Psychonomic Bulletin & Review (2019) 26:103–126
https://doi.org/10.3758/s13423-018-1501-2

THEORETICAL REVIEW

Using experiential optimization to build lexical representations

Brendan T. Johns[1] · Michael N. Jones[2] · D. J. K. Mewhort[3]

Published online: 2 July 2018
© Psychonomic Society, Inc. 2018

Abstract
To account for natural variability in cognitive processing, it is standard practice to optimize a model's parameters by fitting it to behavioral data. Although most language-related theories acknowledge a large role for experience in language processing, variability reflecting that knowledge is usually ignored when evaluating a model's fit to representative data. We fit language-based behavioral data using experiential optimization, a method that optimizes the materials that a model is given while retaining the learning and processing mechanisms of standard practice. Rather than using default materials, experiential optimization selects the optimal linguistic sources to create a memory representation that maximizes task performance. We demonstrate performance on multiple benchmark tasks by optimizing the experience on which a model's representation is based.

Keywords Cognitive modeling · Model optimization · Language processing · Corpus-based modeling · Distributional semantics

Cognitive models specify constructs, such as a decision criterion or encoding probability, that are controlled with free parameters (Shiffrin, 2010). The parameters' values are determined by adjusting them systematically to minimize the discrepancy between the model's output and behavioral data; that is, by fitting the model to data. The idea is to give the model its "best shot" at accounting for the behavior in question, and there are many increasingly sophisticated algorithms with which to fit a model's parameters (e.g., Myung, Cavagnaro, & Pitt, 2017; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

A tacit assumption in cognitive modeling is that behavioral differences across individuals, or tasks, can be explained in terms of processes controlled by the fitted parameters. But an important source of variance comes from differences between individuals' memory, independently of the processes controlled by the fitted parameters. As Hummel and Holyoak (2003) 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" (p. 247). Everyone has had different experiences with the world, leading to variability in people's memorial representations.

The assumption that aspects of the external world are stored internally is almost universal; it appears in theories of memory (Anderson & Schooler, 1991), of perception (Barsalou, 1999; Shepard & Metzler, 1971), and of language processing (Landauer & Dumais, 1997; Tomasello, 2003). In effect, the assumption acknowledges that humans are embedded in a structured environment that constrains behavior because it informs learning.

Early on in cognitive modeling, Estes (1955) urged the field to shift "the burden of explanation from hypothesized processes in the organism to statistical properties of environmental events" (p. 145). Simon (1969) expanded on Estes's call, emphasizing that the "apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves" (p. 53).

Simon (1969) describes a simple parable on the importance of the external environment in understanding behavior: He describes the difficulty of ascribing internal processing mechanisms to an ant's path on a beach. Although the path that the ant takes may seem complicated from a birds-eye view, the complexity is likely a reaction to obstacles in its way. If one were to examine the ant's path without regard for its environment, one might be motivated to ascribe sophisticated internal mechanisms to the ant instead of acknowledging that the ant is a simple organism reacting to a complex environment. If different paths are taken by different ants, it is possible that they

✉ Brendan T. Johns
btjohns@buffalo.edu

[1] Department of Communicative Disorders and Sciences, University at Buffalo, 122 Cary Hall, Buffalo, NY 14214, USA

[2] Indiana University, Bloomington, IN, USA

[3] Queen's University, Kingston, Canada

have different internal process rules; however, it is more plausible that they have identical process rules, but were started at different points on the beach and hence encountered different environmental regularities.

The parable of the ant on the beach applies broadly to theories of cognition (Estes 1975), and to language in particular. For example, theories of word recognition and retrieval (e.g., Goldinger, 1998; Murray & Forster, 2004; Norris, 2006) use a word's frequency of occurrence as a central organizing principle to acknowledge that words that occur more frequently are processed more efficiently (Broadbent, 1967; Brysbaert & New, 2009; Forster & Chambers, 1973). Similarly, distributional models of semantic memory propose that knowledge of what words mean can be inferred by how words are used in language, abstracted across large text corpora (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997). Exposure to different patterns of linguistic information is often used as an explanatory variable to account for behavioral differences in memory (Johns & Jones, 2010; Johns, Jones, & Mewhort, 2012; Mewhort, Shabahang, & Franklin, 2017; Nelson & Shiffrin, 2013), bilingualism (Gollan, Montoya, Cera, & Sandoval, 2008; Johns, Sheppard, Jones, & Taler, 2016; Taler, Johns, Young, Sheppard, & Jones, 2013), in syntactic processing (Johns & Jones, 2015; Reali & Christiansen, 2007; Wells, Christiansen, Race, Acheson, & MacDonald, 2009), language acquisition (Abbot-Smith & Tomasello, 2006; Bannard, Lieven, & Tomasello, 2008; Tomasello, 2003), and in aging (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014; Ramscar, Sun, Hendrix, & Baayen, 2017).

In this article, we explore how to integrate a person's environmental information into a process for maximizing a model's performance on a task. In doing so, we do not wish to underplay the importance of the processing mechanisms addressed by standard techniques. Rather, our focus is on the neglected source of variance; the background knowledge and experience that subjects bring to tasks, knowledge that corresponds to the environment in Simon's (1969) parable of the ant. We present an existence proof demonstrating that it is possible to account for differences in performance on cognitive and linguistic tasks in corpus-based models without changing process parameters, but rather by acknowledging the learning history.

Subjects differ in what they know, and they recruit different memories relevant to different tasks; the differences should prompt a corresponding divergence in behavior. Accumulated linguistic knowledge should have a larger impact on a lexical-decision experiment than on a perceptual-identification task. Hence, including linguistic knowledge when modeling lexical decision makes a good deal of sense.

One way to build linguistic information into a model is to use a representation of word meaning constructed from a standard corpus, such as the TASA corpus (first used by Landauer & Dumais, 1997). TASA is a set of paragraphs from textbooks, sampled from Grades 1 to 12. The TASA corpus has been used as the gold standard in tests of co-occurrence models (e.g., Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997), and it has frequently been integrated into processing models in cognate areas (e.g., Johns & Jones, 2015; Johns et al., 2012; Mewhort et al., 2017).

Although the TASA corpus is likely representative of the linguistic experience that many subjects have experienced, it is not intended to map exactly onto the experiences of specific individuals. Indeed, subjects likely have experienced wildly different linguistic sources, depending on culture, geography, educational system, and so forth. Hence, for any group of subjects, there is a natural variation in their knowledge—variation that should impact their behavior on specific laboratory tasks.

A recent distributional analysis by Johns and Jamieson (2018) illustrates the underlying variability in natural language. In their study, a large sample of fiction was organized by author and genre. There was a small genre effect, where authors who wrote in the same genre had a small increase in the similarity of their writings when compared with authors who wrote in different genres. However, the biggest difference emerged at the individual-author level: Each author had a unique signature of language usage. Given that an individual's exposure to different texts is influenced by a number of factors (including demographics and personal preferences), the author-signature effect suggests that there is a great deal of variance in the natural language environment.

Standard parameter-fitting techniques capitalize on potential variability in cognitive processes. Because different linguistic sources contain different information, variability in a model's behavior should depend on the experience that a model incorporates. Just as optimizing a model's process parameters allows it to have its "best shot" at accounting for particular behavior, a language-based model should provide a like advantage given optimized language experience.

To demonstrate the scope of the issue, we will apply our experiential optimization (EO) approach to several substantive tasks, including lexical semantics, lexical organization, sentence processing, and false memory. The common thread among the examples is representational dependency: Each model incorporates linguistic information into the mental lexicon. Our aim is to use EO within realistic cognitive models to produce benchmark accounts of language-based behaviors.

The use of specialized corpora is not without precedent. For example, Rehder et al. (1998) used a specialized "heart" corpus to analyze the essays of medical students. The specialized corpus was composed of medical texts, as they contain more information about human physiology than TASA (where "heart" is more commonly used in a literary sense). Similarly, Stone, Dennis, and Kwantes (2011) improved performance by

using a search engine with targeted queries to reduce a large Wikipedia corpus into subcorpora. Combined, the past work provides compelling evidence that experiential optimization can benefit cognitive models. Our goal is to isolate important parts of the lexical environment and to provide a general method with which to improve a model's performance.

Although the procedures outlined here provide a solid foundation for data fitting, the primary focus of the research is theoretical, namely, to illustrate the power of variance in language in accounting for linguistic behaviors. Because language is an explanatory variable in a wide variety of theories of cognition (e.g., Gollan et al., 2008; Johns & Jones, 2015; Johns et al., 2012; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Ramscar et al., 2014; Tomasello, 2003), we need to examine its power. The simulations reported in this article provide a powerful look into this problem.

More specifically, we focus on two main aspects of EO: (a) as a new technique for understanding how variance in knowledge controls human behavior, and (b) as a fitting technology for optimizing models based on natural language. As stated previously, the first aspect focuses on optimizing a model's experience with language.

The second application of the technique builds upon much work in the computational cognitive sciences in building cognitively plausible distributional models of natural language acquisition and processing (Bullinaria & Levy, 2007, 2012; Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; for a review, see Jones, Willits, & Dennis, 2015). Distributional models have been highly successful at capturing human behavior and at solving applied problems, such as automated essay grading (see Jones & Dye, 2018). However, if the applied field is to continue to grow, more attention needs to be paid to the materials that are used to train its models. EO provides a promising framework that not only gives a model its best shot at accounting for a set of human behavior but also allows it to perform better in applied situations, by ensuring that the model has the most powerful possible knowledge base for a given task.

The first section of this article will provide an outline of the EO methodology. The following four sections will apply EO to different areas within language and memory processing. The first two topics explored are representational models of lexical semantics and lexical organization. Both explore the power of EO, while not adding additional complexity in the form of a processing model. These two sections contain the foundational simulations demonstrating the varied uses of EO in accounting for behavioral data, including using standard optimization procedures such as cross-validation. The next two topics, sentence processing and false memory, have both a representational and a processing component. The combination of different topics tests the power that optimizing both process and representation provides, and are used to demonstrate the generalizability of EO. The four topics, although not all encompassing, provide a cross-section of the power of the EO procedure across varied tasks, cognitive processes, and data types. Each section of the article has been written to be relatively self-contained, so the reader can skip a topic if it does not align with their interests.

Optimization framework and language sources

To find the optimal linguistic information for a model on a given task, we started by selecting a wide sample of language sources. Specifically, we split a large collection of text into smaller sections, and a searching algorithm iteratively determined which sections maximized the fit of the model under consideration. At the end of the sampling process, materials should be selected that maximizes the model's performance on a task, similar to the way parameter fitting provides a model with its "best shot" to explain data.

Training materials

The texts come from five sources: (a) Wikipedia (Shaoul & Westbury, 2010), (b) Amazon product descriptions (attained from McAuley & Leskovec, 2013), (c) 1,000 fiction books, (d) 1,050 nonfiction books, and (e) 1,500 young-adult books. All were e-books, and the vast majority were written in the past 50 years by popular authors. Table 1 shows the characteristics of the different corpora. The fiction and young adult had relatively shorter sentences than the nonfiction sources did and hence have fewer total words.

The sources—from an online encyclopedia, to books targeted at young adults, to marketing materials for a large range of products—were selected to represent as broad a range of written language as practicable. It is impossible, of course, to span the entire range of possible source information, but these sources represent a substantial range of texts with which to give EO a fair test.

To equate each corpus's contribution, each source was trimmed to six million sentences, for a total of 30 million sentences across all texts (approximately 450 million words). The data-fitting method was designed to determine which texts are the most informative in accounting for a particular experiment, just as statistical methods are used to estimate the optimal free parameters of a model.

The corpora were split into small sections. Although there was some variation in the size of the sections (because some models relied on sentence information whereas others used paragraphs or documents), the standard section size was 50,000 sentences. Sections were split within each of the different corpora (so each fiction section consisted of fiction

Table 1 Descriptions of the different book sets

Collection	# of books	# of sentences per book	Sentence size	# of words
Fiction	1,000	6,518	13.44	80,640,000
Nonfiction	1,050	6,320	17.25	103,500,000
Young adult	1,500	4,417	12.22	73,200,000
Wikipedia			16.04	96,240,000
Amazon			17.29	103,740,000

books), and each section was composed of whole books (for the book corpora). That is, sentences were not randomized, but were kept in their surrounding context. When possible, books written by the same author were assigned to the same language sections. Using 50,000 sentence sections, the individual corpora were split into 600 different sets across the corpora (120 sections for each language source).

Although each section was small relative to the total, it was, nevertheless, a large chunk of language, representing approximately seven-and-a-half fiction novels. Each section was large enough to measure how much linguistic information the section contains, but was small enough to allow the different sections to be combined into an optimal set. Splitting the corpora into sections allowed us to capture a wide range of linguistic backgrounds.

To illustrate the variability of the lexical materials, we used the BEAGLE model of semantics (Jones & Mewhort, 2007, described in the next section) to examine the diversity of the meanings of the words that were acquired from the different corpora sections. To do so, we used the words from the Toronto word pool (Friendly, Franklin, Hoffman, & Rubin, 1982), a standard word set often used in studies of memory. Semantic representations for the words were constructed with BEAGLE for each of the 600 different sections (that is, 600 sets of different BEAGLE vectors were constructed, leading to 600 different representations for the same word). The similarity of the semantic representations for the same words was compared across the sections by computing the vector cosine for each word in the Toronto word pool with its corresponding representation in every other section. The similarity measures estimate the variance of meaning for the same word across the different corpora.

Figure 1 presents the word-to-word similarity distributions for each corpus. As shown in Fig. 1, different sections of language provide different meanings for the same word. Note that no average similarity measure exceeded a value of 0.7, demonstrating that there are large differences in the usage of words across different samples of language. A human trained on one section would have a different mental representation of these word's meaning than a human trained on another.

There was also considerable variability among the corpus types. For instance, the variability of semantic representations in the Amazon product descriptions was quite limited, as all of the different sections were highly similar to each other. That is, words used in product descriptions are used in a very similar manner. By contrast, the distributions for the sections from Wikipedia were variable, documenting that this corpus is a more diverse source of language. Additionally, the similarity of the book collections was also quite variable, suggesting that different authors use words in different ways, an intuitively satisfying result.

To help understand how the language sources lead to different semantic representations, the similarity distributions in Fig. 1 were reduced to their mean. The resulting measures were used to build a multidimensional scaling solution, displayed in Fig. 2.

Figure 2 illustrates intuitively pleasing patterns: fiction books are similar to young-adult fiction books, while Wikipedia is close to nonfiction books. Note also that the corpus types span the semantic space; that is, the diversity of the source materials accounts for a large number of language types. Together, Figs. 1 and 2 demonstrate that there is variability in the lexical knowledge that is contained in the various

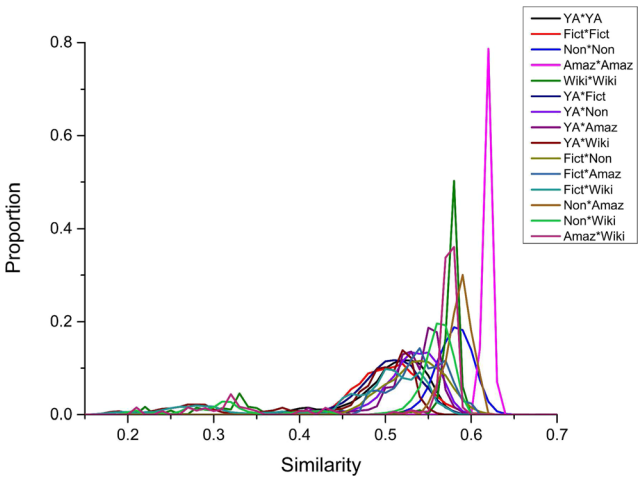

Fig. 1 Semantic similarity distributions for the words from the Toronto word pool for each section of each corpus relative to other sections. These similarity distributions are of the same words (e.g., similarity from *farm–farm*, *dog–dog*, etc.) learned by the model across the different sections. Distributions demonstrate that different sections provide different meanings for the same words, and thus there is natural variability in semantic content across the different sections. (Color figure online)

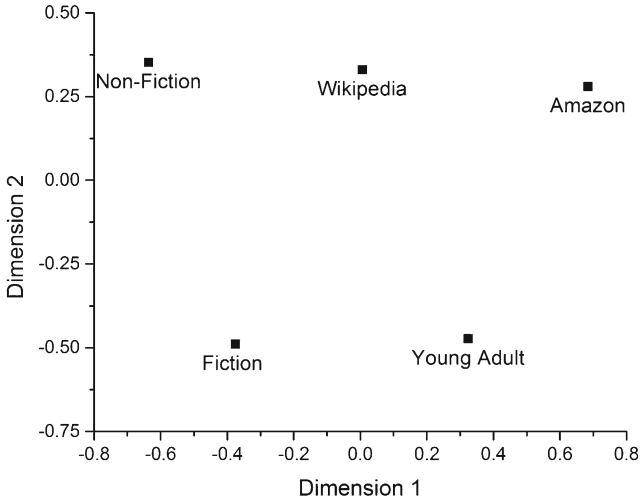

Fig. 2 Two-dimensional scaling solution for the collapsed similarity distributions from Fig. 1

corpora. In short, the corpora provide a solid base with which to explore the power of experiential optimization.

Data fitting

The fitting algorithm's goal is to determine the combination of the text sources that best fits a representation-dependent model to a set of data. To do so, we used a simple hill-climbing algorithm to select the corpora sections iteratively to maximize the model's fit to a behavioral data set. A hill-climbing algorithm is an iterative search algorithm, in which a model is fit by incrementally improving its match to a set of data. Once it is no longer possible to improve the fit, the algorithm terminates.

On the first iteration, the algorithm selects the singular section of language that provides the best fit to the data under question, the beginning of the construction of the optimized corpus. The model is trained on all 600 language sections, and the section that offers the best fit is selected.

The selected language section is added into the optimized model's representation, and the selected section is removed from the search set, meaning that it cannot be used again (i.e., sampling without replacement). Details of how the selected set is used to update a model's representation depends on the model being optimized. Typically, a model would need to be retrained on the complete set of language selected, but there may be computational shortcuts.

The next iteration of the algorithm finds the section that again maximizes the model's performance when combined with the model's current representation, and the new section is removed from the search set and added into the model's representation. The process iterates until a further section does not increase the model's performance. If a section does not cause the model's performance to decrease, however, the algorithm will select it and reiterate. The algorithm terminates only when a new section decreases the model's performance. The stopping rule forces the training materials to increase their resolution continuously, and, eventually, to maximize the set of language materials that provide the best explanation for the data.

The use of a simple search algorithm was intentional: It allows for the language materials themselves to determine the fit to data and ensures that the fit derives from the combination of language materials, not tricks of the data-fitting algorithm. That said, future work should explore the use of more efficient and intelligent search mechanisms (see the General Discussion for more).

An initial illustration of EO using BEAGLE and the TOEFL

To illustrate experiential optimization, we conducted a simulation with the BEAGLE model of semantics. BEAGLE is a random vector accumulation model; it uses sentences to update a word's semantic representation in memory. Each time BEAGLE encounters a word in text, that word's representation is updated with information about the other words in the same sentence. Across a corpus, BEAGLE forms deep representations of word meanings (Jones & Mewhort, 2007).

In BEAGLE, words are coded initially by a static environmental vector (composed of values sampled from a unit normal distribution). Each environmental vector identifies the word and can be thought of as a perceptual (visual/auditory) label for the word. As learning proceeds, BEAGLE builds context and order vectors that store the updated information about the other words in the sentence and about the word's position in the sentence, respectively. For context information, updating is done by summing the environmental vectors of the other words in the sentence (with high-frequency function words removed). Accordingly, the context representation accumulates pure co-occurrence information. Order vectors, by contrast, accumulate rudimentary syntactic information, by recording the word's relative position in the sentence.

There are multiple implementations of BEAGLE (Jones & Mewhort, 2007; Recchia, Sahlgren, Kanerva, & Jones, 2015). Here, we used the sparse implementation described in Recchia et al. (2015). [1] For the simulations in this section, we only used the item vectors (sometimes called context vectors).

A classic test to assess semantic memory is the Test of English as a Foreign Language (TOEFL; first used by Landauer & Dumais, 1997). The TOEFL is a synonym test; subjects are given a target word and must select the word closest in meaning to the target from a set of four foils.

[1] Vectors had a dimensionality of 10,000, and environmental vectors had six nonzero values, similar to `Recchia et al. (2015). Consistent with model architecture, the stop list from `Landauer and Dumais (1997) was used to train context vectors, but not order vectors (`Jones & Mewhort, 2007).

Performance is assessed by how many of the correct synonyms are found.

To demonstrate EO in action, BEAGLE vectors were fit using 120 sections of the nonfiction corpus. To avoid local maxima, we used 10 unique starting points, in a rank order of the best first fitting sections. The best fit across the different starts was displayed in all simulations that follow.

Figure 3 shows the performance on TOEFL for each iteration. On the first iteration, performance ranged from about 10% correct to 35% correct, but EO's real power is shown in subsequent iterations: As the best-fitting sections are integrated into the overall representation, performance increased for all remaining sections. That is, past acquired information scaffolds current information. At the second iteration, for example, performance jumped to between 20% and 45% across all sections. By iterating the process, EO built a representation that performed to a high degree on a difficult task.

Recall that Landauer and Dumais (1997) reported accuracy of about 55% using latent semantic analysis (LSA) on the same TOEFL test, approximately the same value we obtained on the fourth iteration, and EO maximized at about 70% accurate after 17 iterations. That equals 850,000 sentences used in training the model, only slightly larger than the TASA corpus (~750,000 sentences) on which LSA was trained. As Fig. 3 illustrates, a combination of language produced a rapid gain in BEAGLE's power to explain semantic data. Later in the paper, we will illustrate even better performance on the TOEFL by including text in addition to the nonfiction materials used here.

The simulation presented in Fig. 3 demonstrates the difference between EO and standard parameter fitting methods. Typical parameter fitting algorithms, such as a SIMPLEX algorithm, shift the objective fit of a model by exploring the parameter space of that model. Likewise, EO shifts the objective fit of a model by shifting the informational content from which a model derives knowledge. Different tasks may require different knowledge, and EO allows task requirements to determine the best lexical information for the situation. That is, instead of exploring the parameter space to determine optimal model behavior, EO explores the space of possible knowledge that a model could acquire.

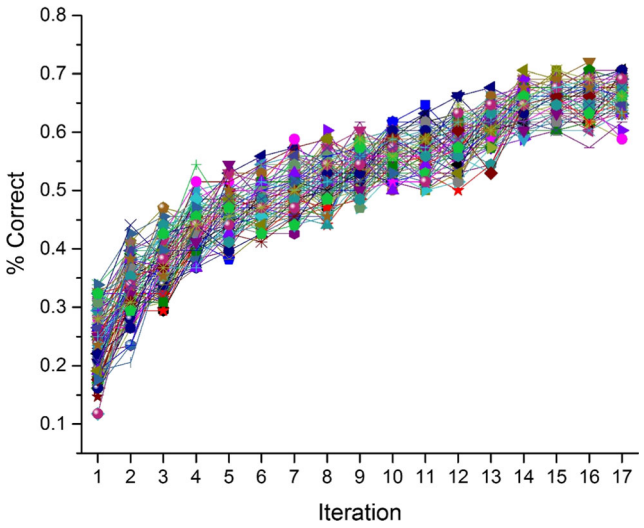

Fig. 3 Example of experiential optimization applied to the BEAGLE model on the TOEFL task. Each line represents a different section from the nonfiction corpus. Across iterations, the process assembles a corpus that produces optimal accuracy on the task. (Color figure online)

For BEAGLE, the integration of new sections is computationally simpler than in other models, as word vectors can be pretrained for each individual section. The pretrained vectors contain the semantic information contained in the respective language sections. The optimized representation can then be updated by summing these word vectors into the optimized representation. However, this is not the case for all models. Different models may need to be retrained with the different sections to determine the optimal linguistic information sources. This will be explored across the different simulation examples contained below.

Although hill-climbing algorithms are often problematic (e.g., they sometimes get stuck in local maxima), they nonetheless provide a simple method to determine how much linguistic information is necessary to optimize the fit to human task performance. To avoid local maxima, 10 unique starting points were made, in a rank order of the best-first fitting sections. Using multiple starting points reduces the risk of the algorithm becoming stuck in a local maximum. The best fit across the different starts was displayed in all simulations that follow.

Discussion

To illustrate EO, text was assembled from a diverse set of sources. To determine the optimal sources needed to explain the data, the texts were split into smaller pieces, and a hill-climbing algorithm was used to find the selections of text that maximally increased the fit of a model to a set of data. The process is a kind of parameter fitting (see Shiffrin et al., 2008), but instead of optimizing the internal process parameters to explain data, we optimized a model's knowledge base (i.e., memory representations).

BEAGLE does not have process parameters; hence, we manipulated only the linguistic material given to the model, with the representational parameters being held constant from previous studies (see Recchia et al., 2015). In our next examples, we will also restrict our analysis to manipulation of linguistic material because of the massive amount of computation that would be required to do both (more is provided on the point in the General Discussion). To demonstrate the power of EO, the optimization procedure was applied to four distinct areas: (a) lexical semantic memory, (b) lexical organization, (c) sentence processing, and (d) false memory. The following four sections will contain these four examples.

Example 1: Lexical semantic memory

Models of semantic memory, beginning with latent semantic analysis (LSA; Landauer & Dumais, 1997), have strongly influenced behavioral studies on the effect of linguistic experience. LSA demonstrated that a simple dimensional reduction mechanism, when combined with sufficient linguistic experience (derived from a large text corpus), could construct a representation of the meaning of words that approximates human semantic similarity data.

Here, we use a model derived from the previously described BEAGLE model (Jones & Mewhort, 2007). The original BEAGLE model used circular convolution to form an *n*-gram representation of the order information. As stated previously, we use a simplified form of the model that uses random permutations of sparse binary vectors to record order information because it reduces the computational expense of the fitting procedure (see Recchia et al., 2015). In the following simulations, we used order, context, and composite (the sum of the order and context vectors) representations.

BEAGLE is a very simple mechanism; it records the word's use across text. No higher level information is used.

Data sources

We used three different tests: (a) synonym tests, (b) semantic similarity ratings, and (c) item-level semantic priming. In the previous section, we introduced the synonym test using the TOEFL.

The second test used semantic similarity ratings in which subjects are given a pair of words and asked to rate their similarity (see Recchia & Jones, 2009). We used three standard similarity-rating tasks, one from Rubenstein and Goodenough (1965), one from Miller and Charles (1991), and one from Finkelstein et al. (2002). Rubenstein and Goodnough used 65 noun pairs, consisting of pairs that are synonyms to pairs that are completely unrelated in meaning. Miller and Charles took 31 pairs from the Rubenstein and Goodenough study to replicate the original study. Finally, Finkelstein et al. used 353 word pairs and included common nouns, proper nouns, verbs, and adjectives.

Finally, we tested semantic priming (e.g., Jones, Kintsch, & Mewhort, 2006; Hare et al., 2009; Lund & Burgess, 1996). In a semantic priming experiment, subjects are asked to perform a simple task, such as lexical decision, and the target word is preceded by a prime word. The prime can be semantically related or not, and priming is measured as the processing speedup observed when the target is preceded by a semantically related item relative to a semantically unrelated word.

Hutchison, Balota, Cortese, and Watson (2008) have shown that models of semantic representation succeed at the mean level across items but fail at the item-level word level. They examined priming in lexical decision for 300 different items and found that semantic variables were not good predictors of the data; forward association strength yielded the best correlation to overall levels of priming at $r = .164$, $p < .01$, while LSA had a nonsignificant correlation of $r = .053$. Clearly, semantic priming data are challenged when examined at an item level; this provides an excellent test for the power of EO.

Data-fitting methodology

We used the same data-fitting method described in the initial illustration of EO. To build a baseline for comparison, 50 resamples of the full corpus (of 30 million sentences) were taken. That is, 50 randomized corpora were assembled by randomly ordering the 600 sections. These randomized corpora will serve as a comparison for the increase that the optimized corpus provides. The average performance increase across each 50,000 section of these corpora was recorded.

Results

Although we used BEAGLE and the TOEFL earlier, to show how optimization works (e.g., see Fig. 3), we used only one language type (nonfiction books). Figure 4 shows accuracy on the TOEFL test when all corpus types are used in fitting this task. Figure 4 shows performance as a function of the number of sentences included in the fit, and the three kinds of information (complete, context, and order). In addition, Fig. 4 shows performance on the control condition in which the sections of text were assembled randomly using the complete (context + order) representation.

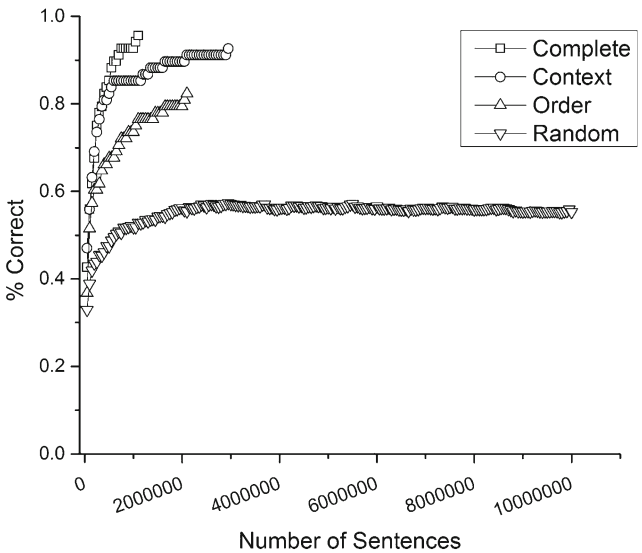

Fig. 4 Results of experiential optimization on the TOEFL task for BEAGLE's three representation types. Random line represents the complete BEAGLE model trained on randomly constructed corpora

For the random corpora, the results were concatenated at 10 million sentences in order to aid visualization. The context representation maximized at 92% accurate at 3 million sentences, while the order representation maximized at 82% accurate at 2.1 million sentences. For the random corpora, the average maximum performance was 57%, consistent with the past results (Jones & Mewhort, 2007; Landauer & Dumais, 1997). The complete model achieved maximized performance on the TOEFL at 97% accurate at only 1.1 million sentences. The complete representation performed at the same level as a native English speaker, an impressive level of performance for a rather simple model.

As Fig. 4 illustrates, by selecting the most informative sections at each iteration, EO formed a combination of language materials that highly matches the task.

Table 2 shows data for the different sets of semantic similarity ratings. The random model is the complete BEAGLE model (context + order) trained on randomly assembled corpora. All three representation types achieved very high levels of performance across all sets. Again, the complete model provided the best fit. For the Rubenstein and Goodenough (1965) data, the rank correlation was $r = .962$, $p < .001$; for the Miller and Charles (1991) data, the rank correlation was $r = .974$, $p < .001$ set. For the larger Finkelstein et al. (2002) data set, the rank correlation was $r = .791$, $p < .001$. In contrast, the average rank correlation for the randomly composed comparison corpora was .582, .592, and .527, respectively. Again the EO accounted for similarity ratings over and above fitting to randomly composed corpora.

Understanding the variability in language is key. As Fig. 1 shows, there is significant variability in the information contained in the different sections. A natural question is what effect this variability has on accounting for lexical behavior.

To answer, we examined the number of sections contained in the search set on the ability of EO to fit to the data from Finkelstein et al. (2002). To do so, we manipulated the number of sections available from 100 to 600 in steps of 100. Each of the five corpora contained 20% of the sections (e.g., at a size of 100 sections, each corpus type contributed 20 sections). Section sizes were kept constant at 50,000 sentences. Sections were randomly sampled, and the resulting fit was the best fit across 25 different resamples of the sections.

Table 2 Correlations between fitted and random representations to semantic similarity norms

Data set	Context	Order	Complete	Random
Rubenstein	0.921	0.943	0.962	0.584
Miller & Charles	0.95	0.835	0.974	0.592
Finkelstein	0.752	0.724	0.791	0.527

Note. Values represent Spearman rank correlations. All correlations are significant at $p < .001$

Figure 5 plots the increase in correlation to the Finkelstein data as a function of number of sections included in the search. There was a constant increase in fit as the number of sections was increased: The added variance provided by additional sections allowed greater flexibility in the representations constructed. In turn, the flexibility allowed BEAGLE to gain a better fit using more finely tuned representations. The benefit of finely tuned representations is a promising outcome because it suggests that the EO's power can only increase.

An additional question about EO concerns how deterministic it is. That is, when EO is run, does it always select the same sections, or is there significant variability in the likelihood of a particular section being chosen. A follow-up question is how the stimuli that are being fit to impacts the sections that are selected.

The issue is important because it tells us about the interaction between the content of the language sections and the behavioral data being fit. If the same sections are always selected, some sections may just provide a better general fit to lexical data. However, if there is variability in the sections selected, the method must be sensitive to both the content of the various language sections and the behavior that is being fit.

To examine these issues, we took the word-pair similarity of Finkelstein et al. (2002) and split the pairs into two parts (a similarity set and a relatedness set), corresponding to the proposals of Agirre et al. (2009). The similarity set contains 203 word pairs, while the relatedness set contains 252 word pairs. BEAGLE was then optimized to the two sets independently.

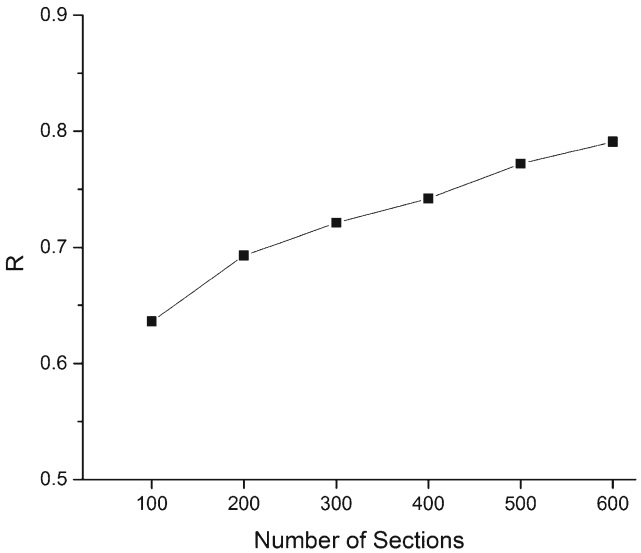

Fig. 5 Simulation examining effects of increasing the number of sections available to the experiential optimization method. Number of sections was manipulated from 100 to 600, with each corpus type occupying 20% of the sections. As the number of sections, and hence variability of the language materials, increases, there is a corresponding increase in the power of the optimization method. This suggests that as the variability of the language materials available to the experiential optimization procedure increases, there is a corresponding increase in the ability to account for variability in lexical behavior

Each set had 30 optimized corpora generated, by initializing the optimization algorithm with different first starting points. The overlap between sections selected was then compared for the corpora generated with the same data set and the corpora generated for different data sets. EO was terminated after 20 iterations, to limit the size of the search set.

When the algorithm was optimized to the same data, on average, 34% of the sections selected were the same across runs of the algorithm. That is, the model is not overly predictable in terms of which language sections are selected during optimization—it depends on the other sections that have already been selected. When the model was optimized to different data sets, the overlap was reduced to 11%. In other words, EO is not very deterministic: Different runs on the same data can yield quite different corpus construction. Additionally, different data force selection of different sections, which makes sense if one considers optimization as using behavioral data as queries to a search algorithm—different queries will return different information. However, it is difficult to know exactly how different the runs of the algorithm are, as the semantic content of two sections could be very similar.

Finally, we also simulated Hutchison et al.'s (2008) item-level semantic priming. Recall that it is difficult for semantic-space models to account for item-level priming results. Figure 6 shows the fitted correlation as a function of the number of sentences for context, order, complete (combined) and random controls (randomly assembled corpora for the complete BEAGLE model). All representation types (except the comparison corpora) provided a good fit to the item-level data in semantic priming. There was not a great deal of difference among the nonrandom representations, with the complete model offering the best fit at $r = .412$, $p < .001$.

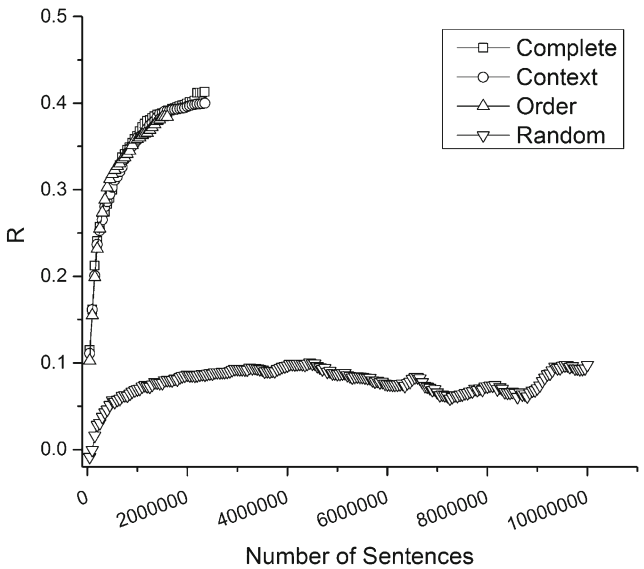

Fig. 6 Results of experiential optimization on item-level priming data from Hutchison et al. (2008) for the three representation types and the randomly assembled corpora for the BEAGLE model of semantics

A skeptic may ask whether the EO can be fit to any set of data. If EO attains high levels of fit to *any* set of data, it may be capitalizing on noise within the different text sources, not necessarily on any systematic connection between natural language and linguistic behaviors. Likewise, if EO were able to attain a high quantitative fit only on sets of real language-based behaviors, the method can find the connections between language and behavior only when the connections are actually contained within natural language.

To evaluate this possibility, BEAGLE was fit to randomized data from the TOEFL, the similarity ratings from Finkelstein et al. (2002), and the semantic priming data from Hutchison et al. (2008). To randomize the TOEFL, a target word was given a set of alternatives from a different target word (e.g., for the target word *grin*, the model would have to learn that it is associated with *mild* and select it from the set of alternatives *mild cold short windy*, instead of associating it with the word *smile* in the set of alternatives *smile exercise rest joke*). For the word–word similarity data, two word pairs were selected and their associates were switched, and their data were randomized (e.g., in the data, the pair *tiger–cat* has an association value of 7.35, and the pair *football–basketball* has an association value of 6.81; in the randomized data, the pairs would be *tiger–basketball* and *football–cat*, with respective association values of 6.81 and 7.35). The same randomization technique was used for the semantic priming data, but with randomized related and unrelated primes. For all three data types, the data were randomized 25 times, EO was applied to each sample, and the average fit across samples was recorded.

The results are displayed in Fig. 7. For all three sets of data, the fit to the intact data far exceeded the fit to randomized data. For example, the fit to the randomized TOEFL task is 26% correct, which is at chance. That is, the model was unable to learn the new associations; EO is unable to acquire associations that are not available in the statistical patterns of the natural language environment. Likewise, for the Finkelstein et al. (2002) data, the correlation was $r = .271$ to the random data, but $r = .791$ for the intact data, a difference of about 55% in the variance explained by the correlation. The fact that the fit to the randomized data was not zero suggests that EO can capitalize on random noise, but the amount of variance accounted for is much greater for the intact data (62.5% vs. 7.34%). The difference was not as large for the semantic priming data, as the intact data had an $r = .412$, and the randomized data had an $r = .213$. This signals that the fit to these data is not as impressive as was first thought. Indeed, the fact that EO is only providing a small advantage to semantic priming data is concerning and likely signals that there is little variance for EO to account for using lexical experience as a guide. Particularly, this smaller advantage for the priming data suggests that the EO method that the structure of the lexical environment is less powerful in accounting for priming data (as

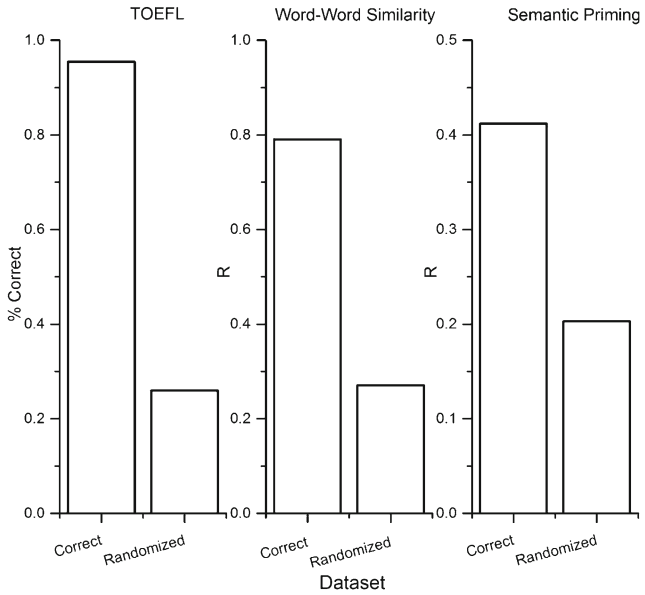

Fig. 7 Performance of BEAGLE optimized to correct and randomized data across three different tasks. Simulation demonstrates that the experiential optimization procedure is only able to optimize to behavioral data when the pattern is contained within the structure of natural language materials. The model is not able to simply fit to any type of data, but is limited to data whose organization is related to the natural language environment

compared with the TOEFL and word similarity data), a result that is consistent with the large individual variance in semantic priming (e.g., Yap, Hutchison, & Tan, 2016). It also points to other approaches, such as processing accounts of semantic priming (e.g., Cree, McRae, & McNorgan, 1999). They may be more use to understanding semantic priming than distributional approaches. The differences in variance accounted for across the tasks suggests that EO could form the basis for a new method with which to assess how much a given task is based in information from the lexical environment versus from internal cognitive mechanisms. This issue will be discussed further in the General Discussion.

In this section, we have only used the BEAGLE model of semantics to examine lexical semantics. As described previously, BEAGLE is an ideal model to use with EO because it is computationally efficient (especially when using the sparse implementation of Recchia et al., 2015, which is the implementation used here), and the use of EO requires significant amounts of computation. The other feature that makes it useful for EO is that it is compositional—vector sets trained on different corpora can be averaged to combine the knowledge gained from the different language sections. This is not possible with dimensionality reduction models, such as latent semantic analysis (LSA; Landauer & Dumais, 1997) or topics (Griffiths et al., 2007). However, one objection to the use of only one model to demonstrate the effectiveness of EO means that the method may be optimizing to some unique aspect of the mathematical framework of BEAGLE, and not necessarily determining the most informative sections of text to simulate a task. To test this possibility, a final simulation was done where a different distributional model—latent semantic analysis—was trained on the optimized corpora from runs of the BEAGLE model. If the LSA model trained on the optimized corpora exceeds performance compared with when the model is trained on randomly selected corpora, it would signal that the optimization method is finding text sections that are generally informative of task performance, and not just optimizing to the mathematical peculiarities of BEAGLE.

To test this possibility, 30 optimized corpora were constructed by running EO with BEAGLE context vectors (as context information aligns best with the operations of LSA) with 30 different starting points, meaning a different corpus is generated with each run. The optimization process was stopped at 20 sections, providing optimized corpora of 1 million sentences each. Additionally, 30 control corpora were generated by randomly selecting 20 sections to correspond to each optimized corpora. LSA was then trained on each of the optimized and randomized corpora, and model performance was assessed by taking the correlation to the Finkelstein et al. (2002) word similarity data. Unlike BEAGLE, LSA uses paragraphs as its unit of analysis. Paragraphs were formed by combining 20 sentences in a moving window across the text sections. This means that each text section is reduced to 2,500 paragraphs.

Figure 8 contains the average correlation of the optimized BEAGLE models, LSA models trained with the optimized corpora from BEAGLE, and LSA models trained with randomized corpora. The BEAGLE model performed the best,

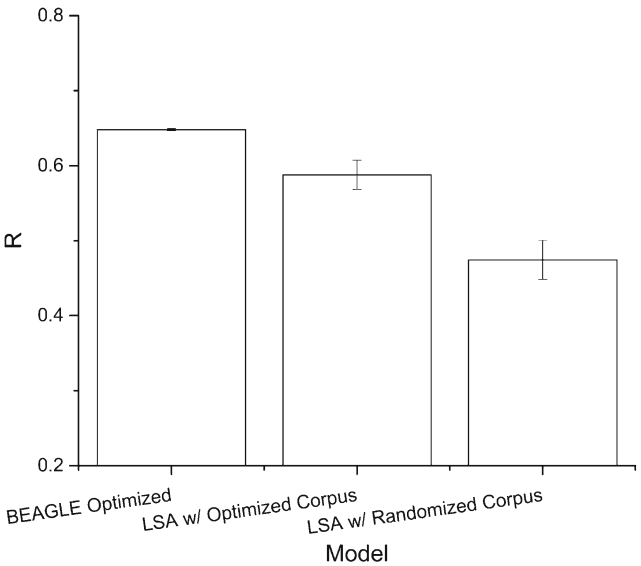

Fig. 8 Simulation using two-step optimization where the latent semantic analysis (LSA) model of Landauer and Dumais (1997) is trained on optimized corpora attained from the BEAGLE model. When trained on optimized corpora, latent semantic analysis shows an advantage compared with randomized corpora that contain equivalent amounts of text

which was expected given it is the model that was being optimized (it is lower than the performance of the model contained in Table 2, as this is average performance across the runs and the searching process was cut off after 20 iterations). However, when LSA is trained on the optimized corpora, there is a significant advantage over the model trained with randomly composed corpora, $F(1, 59) = 12.179$, $p = .001$. Thus, even though BEAGLE and LSA use different mechanism to exploit statistical regularities within natural language, the optimization method is finding sections that generalizes to a different distributional model. This suggests that EO is not only using the mathematical properties of BEAGLE to optimize model performance but is also finding sections of text that contain information linguistic stimuli under question. This simulation also suggests that two-step optimization is possible, where a computationally simple model like BEAGLE can be used to determine informative corpora, which can then be used to train more computationally complex models like LSA. However, more research is needed to determine the most efficient way to perform this type of optimization.

Discussion

Semantic space models have been fundamental in exploring the influence of linguistic structure on human behavior (Jones et al., 2015; Riordan & Jones, 2011). In the present section, we explored EO's power when combined with a simplified form of a popular semantic space. Across three data types, EO proved highly powerful in accounting for multiple data types, producing benchmark performance for every data set analyzed. Similar to the optimization of a model's parameters, by optimizing the knowledge base of a distributional model, a model's performance can be massively increased. However, as the simulation in Fig. 7 shows, EO does not work with randomized data; it requires the data to reflect the structure of natural language.

Example 2: Lexical organization

Research in word recognition has recently focused on the influence of environmental variables on the efficiency of lexical access. Classically, word frequency has predicted the lion's share of variance as a lexical variable: Words that are higher in frequency are processed more efficiently (Broadbent, 1967). As a result, word frequency has become a central component in models of lexical access (e.g., Goldinger, 1998; Murray & Forster, 2004; see Brysbaert, Mandera, & Keuleers, 2018, for a recent review).

The importance of word frequency has spawned a variety of norms used to select stimuli in psycholinguistic studies. Norms derive from different corpora, ranging from the classic Kucera and Francis (1967) counts from the Brown corpus, to the more recent SUBTLEX counts from subtitles of television shows and movies (Brysbaert & New, 2009). Such norms provide an excellent fit to large-scale data, demonstrating their utility for both theoretical and methodological applications.

The exact nature of frequency effects has been questioned (see Jones, Dye, & Johns, 2017, for a review). Adelman, Brown, and Quesada (2006), for example, showed that a measure that builds a word's strength in memory by counting the number of contexts in which it appears (operationalized as the number of document occurrences across a corpus) provides a superior fit to lexical access latency than word frequency does. The measure is commonly known as contextual diversity (CD), and its superiority over word frequency has been demonstrated using several corpora (Adelman & Brown, 2008; Adelman et al., 2006; Brysbaert & New, 2009).

Adelman et al.'s (2006) document-count measure ignores the semantic diversity of the contexts in which the word is found. To examine this possibility more closely, Jones, Johns, and Recchia (2012) used an artificial language-learning experiment that manipulated word frequency and contextual diversity. Specifically, certain words occurred with different sets of words (high semantic diversity), while others occurred repeatedly with the same set (low semantic diversity). Although there was no effect of diversity for low-frequency words, high-frequency words were retrieved more quickly when they had been learned across multiple diverse contexts. That is, processing savings occurred only with a change in context. Based, in part on these results, Jones et al. (2012) proposed a new model that builds a more accurate measure of a word's strength in memory, the semantic distinctiveness memory (SDM) model.

SDM builds a word's strength in memory by weighting each new context by the amount of unique information that the context provides about the meaning of the word. Across various corpora, SDM was able to account for a larger amount of variance in a mega data set of lexical decision and naming times over word frequency and a document count. Additionally, Johns, Gruenenfelder, Pisoni, and Jones (2012) demonstrated that the advantage for a semantic diversity extends to spoken word recognition performance, suggesting that the contextual variability of a word is a general property of linguistic organization. Johns, Sheppard, et al. (2016) have extended this empirical work to examine the use of this information source across aging and bilingualism. Johns, Dye, and Jones (2016) extended Jones et al.'s (2012) artificial-language experiment using natural language materials, and confirmed the importance of semantic diversity across a diverse range of areas, such as in age of acquisition effects (e.g., Hills, 2012; Hills, Maouene, Riordan, & Smith, 2010; see also Hoffman, Ralph, & Rogers, 2013).

The goal of the next section is to determine if experiential fitting can provide a better fit to a large set of previously

published lexical-decision data using SDM. We also provide an additional test on the word frequency, contextual diversity, and semantic diversity accounts of lexical strength.

Data sources

The main source of data was 40,000 lexical decision times from the English Lexicon Project (ELP; Balota et al., 2007). ELP is a standard data set that has been used to differentiate different lexical information sources (Brysbaert & New, 2009; Jones et al., 2012). Subject-level data will also be used from the ELP. Additionally, a set of 2,900 lexical decisions times for young and old adults, attained from Balota, Cortese, and Pilotti (1999), was used to test the sensitivity of the experiential fitting method to different subject groups.

Data-fitting methodology

Because the SDM uses paragraphs, we split each corpus into 3,000 paragraphs/documents. For the Wikipedia corpus, we took a single document in the encyclopedia. For the Amazon product descriptions, one product description was considered a separate document. For the books, there was no simple method to split them into paragraphs, due to differences in formatting. Instead, we used a moving window, with a size of 15 sentences, to assemble paragraph-like units.

Because it is a dynamic model, SDM is typically trained on a whole corpus: previously experienced information determines what should be stored for any new context. However, the SDM is computationally complex, so magnitudes were derived separately for each section. To construct an overall magnitude with experiential optimization and the SDM, at each iteration the method selected the section whose magnitudes have the closest correlation to ELP lexical decision times, and that section's magnitudes were added into an overall magnitude. This process was iterated until the fit was maximized. To compare the performance of the SDM model, the same fitting was done for word frequency (WF) and contextual diversity (CD), a standard comparison (Johns, Dye, et al., 2016; Johns et al., 2012; Johns, Sheppard, et al., 2016; Jones et al., 2012; Jones et al., 2017). WF is the count of the number of times a specific word occurred in a section. CD is the count of the number of times a word occurred within a paragraph within the section, ignoring repetitions within a paragraph (Adelman et al., 2006). For example, for experiential optimization with WF, if the word *molecule* occurs 4 times in the first section selected, and 5 times in the second section, the overall magnitude for *molecule* would be 9. This count would be slightly different for the CD variable as repeats within the same paragraph would be removed, leading to a slightly lower count. All variables were transformed with a natural logarithm before assessing the correlation to the data, a standard transformation (Adelman & Brown, 2008).

In concordance with our previous analyses, 50 randomly composed corpora were constructed to be used as a comparison for the experientially fitted magnitudes. The average correlation constructed with the SDM provides a baseline for the optimized model's performance.

Results

To fit to ELP data, all 40,481 lexical decision times contained in the ELP were used. For each iteration, the best-fitting section to the ELP data is used to update the resulting WF, CD, and SDM count. The results of the experiential fitting method on the *z*-transformed ELP lexical decision-time data are displayed in Fig. 9. Only the results of the SDM are displayed in this figure, because WF, CD, and the SDM produced similar results (explored further below). The SDM result is contrasted with the fit that CD values from the SUBTLEX corpus provides for this data set (Brysbaert & New, 2009), the current most widely used norm set. Figure 9 demonstrates that the use of experiential fitting increases the fit for retrieval latencies, even when compared against a very well-constructed corpus. Additionally, the randomized corpora also achieved a correlation that equaled the SUBTLEX corpus, demonstrating that the source materials used in experiential fitting was of very high quality.

As has been found in past studies (Johns et al., 2012), magnitudes from SDM had the highest correlation to the lexical-decision data, $r = -.708$, $p < .001$, compared with $r = -.702$, $p < .001$, for contextual diversity, and $r = -.701$, $p < .001$, for word frequency, demonstrating that all metrics optimized efficiently. For the sake of comparison, the correlation for CD values from SUBTLEX for the same data is an $r = .666$, $p < .001$. To determine the amount of unique variance attributed to each variable, we used the same linear regression

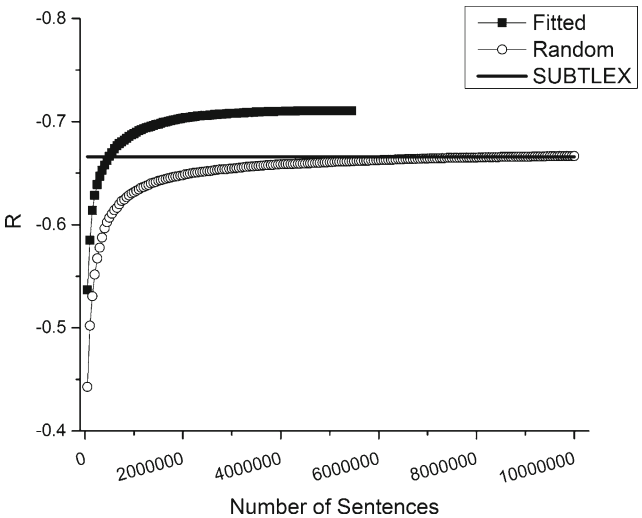

Fig. 9 Experiential optimization applied to the English Lexicon Project database. Performance of the fitted and random corpora are contrasted with the fit of the Brysbaert and New (2009) SUBTLEX corpus

procedure used by Adelman et al. (2006) and Jones et al. (2012) to partial out the unique variance associated with a variable while systematically partialling out the other lexical variables.

Figure 10 shows that the SD variable accounts for the most variance, compared with the WF and CD variables (both of which account for little unique variance). That is, a method that takes into account the semantic diversity of the contexts in which a word appears provides a significantly better fit to retrieval times, a finding coherent with past results (Johns et al., 2012; Jones et al., 2012).

As discussed in the connection with lexical semantic memory (Example 1), a question remains concerning the source of variance that experiential optimization is using in fitting data, as it is possible that it is not accounting for group or individual characteristics, but is instead capitalizing on random noise within the different data sets. As an initial test, 2,900 lexical decision times were attained from Balota et al. (1999) for younger and older adults.

Figure 11 displays the fits to these data with the experiential optimization method for SDM (as well as for randomly assembled comparison corpora), and demonstrates that a high level of fit was produced for both subject groups across corpora sampling, but with a greater fit to younger than to older subjects (a finding also found in Johns, Sheppard, Jones, & Taler 2016, across five different corpora). Additionally, comparing Fig. 11 with Fig. 9, it is obvious that the EO method forms a comparatively larger corpus for the ELP than the Balota et al. (1999) data. The larger corpus reflects the relative size of the data sets (e.g., the ELP contains more than 40,000 words, while Balota et al., 1999, contains 2,900), where the

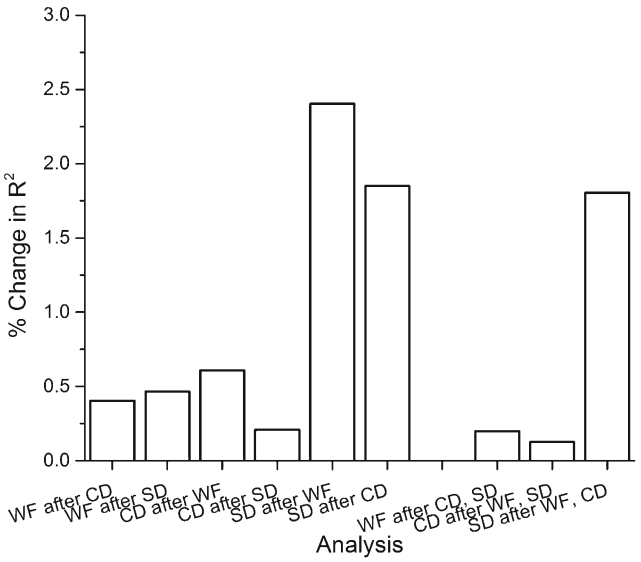

Fig. 10 Results of the regression analysis over the word frequency (WF), contextual diversity (CD), and SDM variables. The values represent the amount of unique variance each of the variables account for. As displayed, the SDM variable accounts for the greatest amount of variance, while reducing the contribution of the other variables

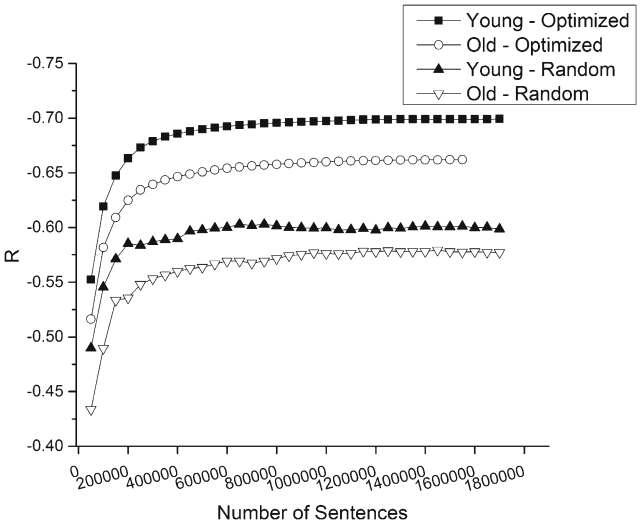

Fig. 11 optimization applied to young and older lexical decision from Balota et al. (1999). Random bars represent performance of the model on randomly assembled corpora

ELP has a proportionally greater amount of variance that needs to be accounted for compared with the Balota et al. (1999) data. This leads to additional sections being integrated into the corpus for the ELP data set. The results of Brysbaert, Stevens, Mandera, and Keuleers (2016) point to the simulation contained in Fig. 11 selecting less lexical information than young adults have likely received. The simulations contained in Figs. 9 and 11 demonstrate why this is—as there are a greater number of data points to account for, there is a corresponding increase in the amount of lexical information necessary to account for the data.

A more compelling analysis examined the composition of the resulting corpora for the two subject groups. To do this, the proportion of the different sections that were selected across optimization was recorded on 20 iterations of the hill-climbing algorithm. The iterations were done by removing the previously selected first section for the current run, so that each run has a unique starting point.

The results of this analysis are shown in Fig. 12. There was no difference in the proportions selected for the nonfiction, Wikipedia, and Amazon sections, but there was a highly significant difference for the young adult sections, $F(1, 39) = 203.51$, $p < .001$, and the fiction sections, $F(1, 39) = 219.45$, $p < .001$. These differences emerge because the young subject group had a higher proportion of young adult sections, while older adults were better described by the fiction sections. Given the composition of the different corpora, this suggests that the retrieval-time data of these different groups are sensitive to the statistics of different linguistic sources that the subjects have experienced: Young adults are better described by simpler examples of language as encoded in young adult books, but older adults are better accounted for by more linguistically diverse fiction and literature books. At least

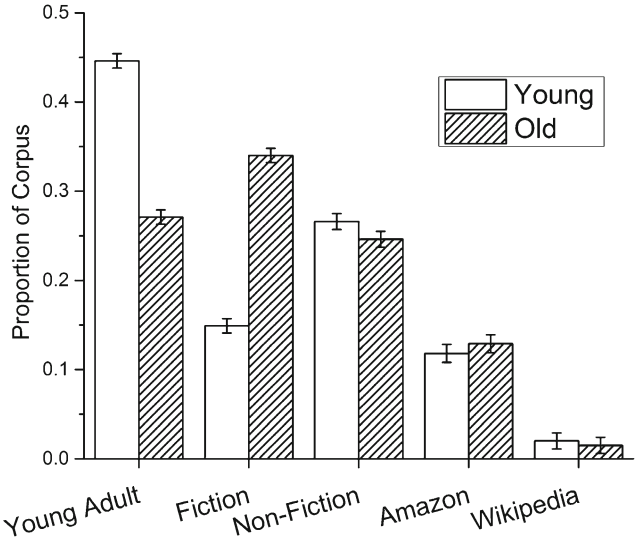

Fig. 12 Proportion of sections selected by the experiential optimization for the young and older subject data from Balota et al. (1999). Error bars are standard error of the mean

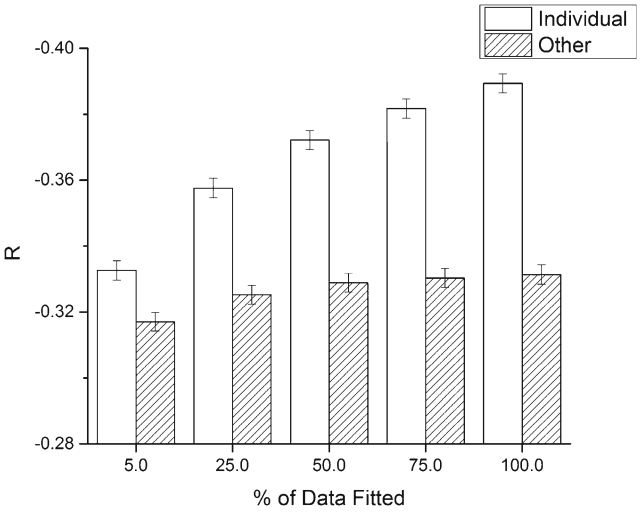

Fig. 13 Cross-validation simulation, using another subject's optimized representation as a comparison. Open bars represent fit to the subject's own full data from the English Lexicon Project when a certain proportion of their data is used as an optimization criterion. Striped bars represent the average fit to other subject's optimized representations

anecdotally, this is consistent with the type of linguistic experiences these subjects presumably had.

Even though the simulation displayed in Fig. 12 provides strong evidence that experiential optimization is fitting a subject's group characteristics, it still does not conclusively demonstrate that the procedure is not simply capitalizing on noise contained within a data set. To demonstrate that the method is indeed capitalizing on individual-level patterns of experience that people have, a simulation was done using a cross-validation procedure on the subject-level data contained in the ELP (Balota et al., 2007). The ELP contains data from 810 subjects, who had on average 1,415 trials with real words. A random percentage of each subject's data was selected, from 5%, 25%, 50%, 75%, and 100% of a subject's data, and the model was fit to each slice of the subject's data. After the EO fitting, the language sections that were selected in the optimization process were used to construct a full SD distribution in order to fit to the subject's complete data set. As a comparison for the resulting fit, the correlation between a subject's data and the average correlation to all other subject's optimized representations was assessed. If the other subject's representations show the same increase in fit, this would suggest that the optimization procedure is simply picking up on general characteristics of lexical-decision data, and not on an individual's pattern of behavioral data.

Figure 13 displays the results of this simulation. The fit of the other subject's representations were considerably smaller than the fits of the representations optimized to an individual's data. There was a small increase in fit of other subject's optimized representation as a function of the amount of data fitted, suggesting that there are some general patterns in the data that the method is picking up on, but that this is much less than the increase that is seen when optimizing to an individual's data. This simulation suggests that the variance that the method is capitalizing on is primarily at the individual level, and not general patterns in lexical-decision data.

However, the simulation displayed in Fig. 13 assesses the fit to the subject's complete data set, and compares the fit to other subject's data to which it was not optimized. An additional question is whether EO allows for an increased fit to the data to which it was not optimized. This is the standard goal of cross-validation procedures, although it is typically used in training models to perform classification and prediction tasks. That is, in a typical cross-validation procedure, a set of data is split into smaller parts and then trained on one set and tested on the unseen set. Performance is then assessed as to whether the performance of the model generalizes to the test data, typically on a classification task. Given that classification and prediction is not the goal of the current article, a comparison to determine relative performance is needed. That is, we need to compare the fit of the optimized representation to both the target subject's data and also to a control subject. This will compare whether there is generalization to an individuals subject's unseen test data.

To do this as controlled as possible, a Monte Carlo simulation was run where, on each sample, two subjects from the English Lexicon Project were selected. The set of words that these two subjects had in common was then found. The use of common words allows for confidence that the sections picked by the searching mechanism is due to differences in the values of the lexical decision data, and not the different words that were included in the search set. To move forward, the subjects had to have at least 100 words in common. This requirement was put into place to ensure that there were enough words to provide variance in the data. The set of common words was

then randomly split into a training and test set by splitting the data in half for both subjects. An optimized representation was then constructed for the training data for each subject, which means that the model was optimized to the same words, but different behavioral patterns. Fits were then assessed across four different tests: (a) fit to training data, (b) fit to same subject test data, (c) fit to other subject training data, and (d) fit to other subject test data. This was done across 5,000 resamples of two subjects using the SDM.

The results of this simulation are contained in Fig. 14. Figure 14 shows that there was a small but consistent advantage for the test set when it was from that subject's data, when compared against the fit of the fitted representation to the other subject's training and test data. This demonstrates there is an advantage for that subject's unseen data, which the optimization method would not have been trained on, suggesting that the method is picking up on the latent structure of a subject's event history—at least enough to provide a small, but measurable, advantage to unseen data.

This simulation has a number of issues. For example, reducing the number of words to which the fit is made necessarily limits the lexical diversity of the language sources that are selected and may cause too narrow of a search space. That is, the words that are included in the data split necessarily impact the searching mechanism, such that other words, especially lower frequency words, may not be included in the resulting representations. Of course, there are ways to modify the searching mechanism to compensate for this, such as by including constraint requirements about baseline frequencies of different words. However, given that this article is meant to introduce the concept of experiential optimization, all derivations of the method are not possible. Additionally, standard behavioral collection techniques are not meant to address these problems, as trial-level data are noisy, and the collection

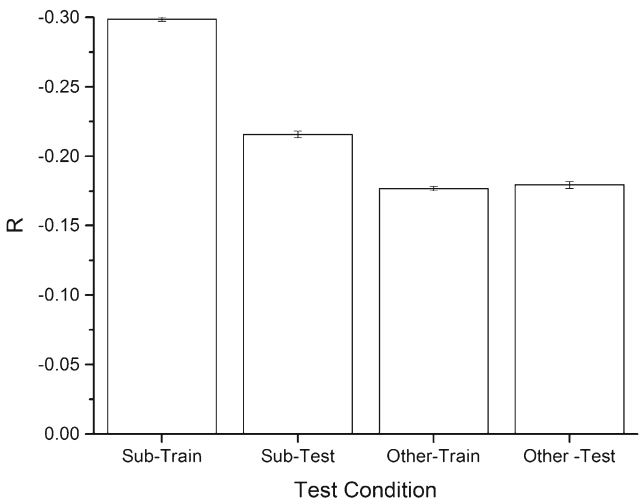

Fig. 14 Monte Carlo cross-validation simulation examining generalization to unseen data

and public release of these data is not standard, except in the case of mega-studies (e.g., Balota et al., 2007; Hutchison et al., 2013). Of course, this is just the start of an important theoretical and empirical issue within the language sciences, which requires much follow-up research.

Discussion

This section has demonstrated that EO can be expanded easily to examine lexical retrieval and subject-level characteristics of data. There is a rich history of using environment variables (i.e., word frequency) to examine word retrieval patterns, with recent research pointing to the importance of contextual and semantic variables in the construction of a word's strength in the mental lexicon (Adelman et al., 2006; Jones et al., 2012). The SDM model, previously shown to provide a superior fit to large-scale lexical decision data than word frequency or a document count, when combined with experiential fitting, provides a better accounting than previously published norms. In a study of young and older adult lexical decision data (Balota et al., 1999), it was determined that the method was sensitive to group characteristics, suggesting that the method is fitting to the experiences that a group of subjects may have had with language, not random noise. Additionally, simulations using cross-validation and Monte Carlo procedures ensured that the EO procedure was capitalizing on subject-level characteristics of the data, not just exploiting random noise within the data and lexical sources.

Example 3: Sentence processing

The next two sections extend EO to processing models, beginning with a model of sentence processing advanced by Johns and Jones (2015). Their model, based in usage-based theories of language (Tomasello, 2003), denies a role for grammatical rules and argues that the utterances one hears and produces form the basis of language processing.

It is known that an exemplar memory system can accomplish some of the fundamental operations of language (Abbot-Smith & Tomasello, 2006). Jamieson and Mewhort (2009a, 2009b, 2010, 2011) and Chubala, Johns, Jamieson, and Mewhort (2016), for example, have shown that such a model can account for several artificial-grammar and implicit-learning results. Johns and Jones (2015) extended the approach; their account explained additional results across sentence processing, grounded cognition, and the cultural evolution of language.

Johns and Jones' (2015) model combines the BEAGLE model of semantics (Jones & Mewhort, 2007) and the MINERVA 2 exemplar memory model (Hintzman, 1986, 1988). The model proposes that the storage and retrieval of linguistic experiences are the fundamental operations of language processing. The theoretical foundation of the model is

consistent with the usage-based view of language (Abbot-Smith & Tomasello, 2006; Tomasello, 2003). The model is fully described in Johns and Jones (2015), along with many supporting simulations, so only a brief description of it will be provided here.

Unlike BEAGLE, the Johns and Jones's (2015) exemplar comprehension model (ECM) stores each sentence encountered as a single exemplar within memory. To construct an exemplar, the ECM uses BEAGLE's encoding scheme to construct a linear ordering of a sentence. Each sentence within a corpus is used to generate an exemplar, which is then stored within memory. Specifically, each word is represented with a randomized environmental vector (equivalent to the classic BEAGLE model), and a sentence exemplar is constructed by forming a composite vector of the environmental vector of each word, with each environmental vector being permuted by its location within a sentence (i.e., the first word in the sentence has a unique permutation, as does the second, etc.). The result is a distributed representation of the linear ordering of the sentence.

The mechanics of Hintzman's MINERVA 2 model are used to retrieve information from the exemplar memory store (Hintzman, 1986, 1988). In MINERVA 2, when a probe is presented, it activates all memory traces in parallel proportional to the similarity between the probe and the trace. The traces are then summed into a composite vector (referred to as an "echo") weighted by their activation values. In a cued-recall task, the echo represents the item that was associated with a probe during study.

In the ECM, the memory probe is the environmental vector for a word, along with the location of that word in a sentence. The echo retrieves a vector from memory that contains the latent expectations for the words that should surround that word in that location. It is called the expectation vector. Expectation vectors are retrieved for each word in a sentence and are summed into an overall sentence representation, called the comprehension vector, which represents the meaning of a sentence in a multidimensional space (similar to the operation of a semantic space model, but for sentences instead of words).

The comprehension vector is used to calculate an expectation value (EV) for each word in a sentence, and the EV is the metric needed used to simulate behavioral data. If a word is expected, its EV should be similar to the comprehension vector, as the previous words processed should have generated a prediction about the upcoming words in a sentence. An EV signals how expected the current word was. A large similarity value promotes increased processing efficiency (and hence a decrease in processing time), because traces in memory that require activation will already be active. The EV for a specific word is calculated by taking the similarity between the comprehension vector and the retrieved expectation vector for that word. The EV is the information source used in the simulations to follow.

The unique aspect of the model, compared with other semantic space models, is that it does not encode a singular representation of a word's meaning. Instead, meaning is distributed across the experiences that the model has had with language. The representation of a word depends on context, so different representations can be retrieved in response to different memory cues (see Jamieson, Avery, Johns, & Jones, 2018; Johns, Jamieson, Crump, Jones, & Mewhort, 2016, for additional capabilities of this approach). As a result, the model can process ambiguity and word sense effects in natural language (see Johns & Jones, 2015) that other types of semantic space models have difficulty explaining.

Because the model computes a formal value of processing ease for each word in a sentence, it is a plausible model to combine with experiential optimization. Additionally, the model is very simple, with only a single parameter (a scaling parameter used to limit the contribution of each individual exemplar in the retrieval process). Hence, the model is completely experience dependent. Experiential optimization allows us to study how much variance in complex linguistic tasks can be attributed to the linguistic experience.

Data sources

Johns and Jones (2015) tested the model using data from several well-controlled psycholinguistic studies. Here, we studied item-level data. Two paradigms were tested: (a) sentence completion norms and (b) ease of processing ratings. The two paradigms examine complementary aspects of the model. Sentence completion norms examine the model's specific expectations, while ease-of-processing ratings examine the global difficulty that people have with different syntactic constructions. Two experiments of each type will be simulated.

In the sentence-completion task, subjects are provided a sentence with the final word missing and are asked to generate the most probable word to fill that position. The first simulation of this type used a sentence set constructed by Rayner and Well (1996). They used Schwanenflugel's (1986) completion norms to assemble a set of 72 sentences, divided into high, medium, and low completion-norm probability values. Rayner and Well (1996) demonstrated that the probability difference manifests in eye-tracking data, with subjects spending a greater amount of time processing low-constraint versus high-constraint words.

Rayner and Well's (1996) data represent a highly controlled and simple data set. We contrasted it with performance on Bloom and Fischler's (1980) much broader sentence-completion norms; their norms consist of 330 sentences and 7,500 response words, with a wide variability in completion probabilities. Given the size of this data set, it provides a

strong test of how well the model is able to account for variance in a complex linguistic task.

The ease-of-processing data come from plausibility/acceptability sentence-rating tasks, in which subjects are asked to rate how sensible those sentences are. The first set of ratings comes from Hare, Tanenhaus, and McRae (2007). They collected plausibility ratings for 144 sentences, which consisted of reduced relatives, unreduced relatives, and passive sentences. Difficulty of the sentences was manipulated by the plausibility of the noun in the sentence.

The second set of ratings, from Rayner, Warren, Juhasz, and Liversedge (2004), consisted of plausibility ratings collected from 90 sentences, with the difficulty of the sentences manipulated by the thematic plausibility. The ratings tested the model on comprehension measures at a global level.

Data-fitting methodology

The fitting methodology was virtually identical to the approach taken in Example 1. Each corpus was split into 120 sections, consisting of 50,000 sentence apiece. A hill-climbing algorithm was used to determine the best combination of sentences that maximized the model's performance. For the sentence completion-norms data set, the EV for each test word (calculated as the cosine similarity between the preceding sentence and each test word) was used to determine the correlation between production probability and the expectations determined by the model. For the ease of processing, the average EV of nonfunction words was used to calculate the correlation between how easy the model expects the sentence to be and the plausibility/acceptability ratings. Johns and Jones's (2015) model has only one processing parameter. The parameter is a scaling parameter and is represented with λ (from MINERVA 2; Hintzman, 1986, 1988). This parameter is used to minimize the contribution of any single exemplar to the retrieval process.

In MINERVA 2, the default parameter setting is 3, but the ECM stores many more exemplars than the classic memory model, and so was set at 11 for all simulations, the same as Johns and Jones (2015). Environmental vectors had a length of 3,000 and were constructed by sampling four nonzero values. Each nonzero entry had an equal probability of being 1 or −1. For comparison purposes, 10 random corpora were formed by sampling 75 sections equally from all five corpora, providing a base rate for any increase in performance associated with the experiential optimization technique.

Results

Figure 15 shows the sentence-completion correlation between the model and the norms as a function of the number of sentences. As shown in the figure, experiential optimization provided a substantially better fit for the model over the randomly constructed corpora. The top panel of Fig. 15 shows the model's fit to Rayner and Well's (1996) highly controlled sentence set. For their data set, the model maximized at an r = .918, $p < .001$, for the fitted model at 800,000 sentences, while the random corpora maximized at an r = .217, $p < .05$, at 2.4 million sentences. The high correlation to the Rayner and Well (1996) data reflects the highly structured nature of the data set, where the sentences were presorted into low, medium, and high bins.

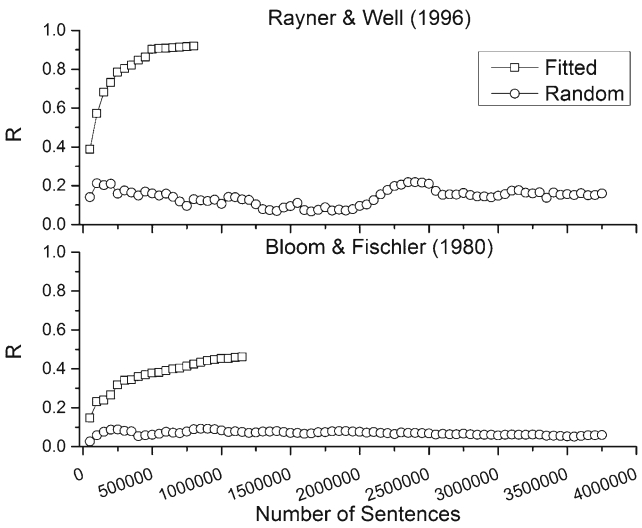

Fig. 15 Results of the experiential optimization and the sentence processing model of Johns and Jones (2015) on two sentence completion norms from Rayner and Well (1996; top panel) and Bloom and Fischler (1980; bottom panel)

The results for Bloom and Fischler's (1980) much larger sentence-completion-norm data set are displayed in the bottom panel of Fig. 15; the fitted model maximized at an r = .462, $p < .001$, at 1.15 million sentences, while the random corpora maximized at r = .093, $p < .001$, at 900,000 sentences. Given the complexity of their data set (which contains more than 7,500 responses), the fitted model accounted for a large amount of variance.

Figure 16 shows the results for the-ease-of-processing simulations. The top panel shows the results for the fitted and random corpora models for Hare et al.'s (2007) data. Again, the model was also quite capable handling their data. The fitted model maximized at an r = .685, $p < .001$, at 950,000 sentences, while the random model maximized at r = .134, *ns,* at 400,000 sentences.

The bottom panel of Fig. 16 shows the ratings from Rayner et al. (2004); the results were similar to the Hare et al. (2007) data set. The fitted model maximized at an r = .634, $p < .001$, at 650,000 sentences, while the random model maximized at r = .151, *ns,* at 1.85 million sentences. The simulations show that the model is very capable of accounting for global ratings, even with different sentence types.

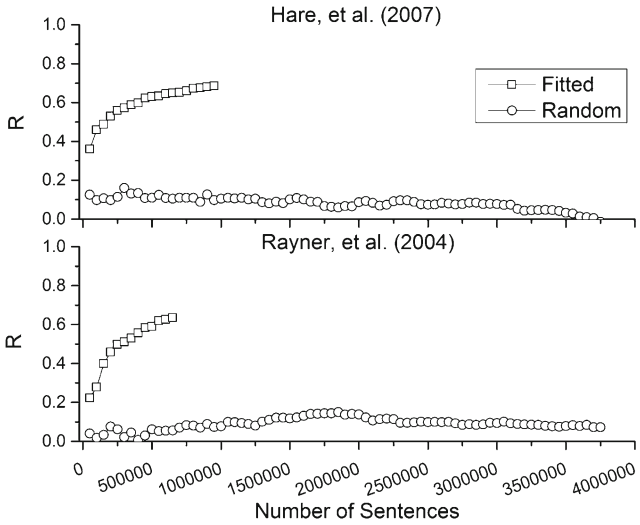

Fig. 16 Results of experiential optimization and the sentence processing model of Johns and Jones (2015) on two plausibility rating data sets from Hare et al. (2007; top panel) and Rayner et al. (2004; bottom panel)

Discussion

The exemplar-model approach to natural language (Johns & Jones, 2015) proposes that much of the variance in linguistic behavior reflects the structure of the individual's experience with language. The model stores sentences constructed using a popular semantic space model (Jones & Mewhort, 2007) and uses a classic exemplar model of memory to retrieve information (Hintzman, 1986, 1988).

For our current purpose, the attractive feature of this model is that its performance is tied entirely to the information that it has experienced—there is no extra or built-in complexity. The model's only parameter, a simple scaling parameter, limits the impact of any single exemplar on the retrieved echo. Thus, the model provides an ideal way to test the power of experiential optimization. Equipping the model with experiential optimization enabled it to provide excellent fits to two different measures of language processing, sentence completion and ease of processing. Compared with performance on randomly assembled corpora, the EO method provides a massive increase in performance. Together with the simulations reported by Johns and Jones (2015), the present work poses an interesting problem for psycholinguistic theory: How much of language usage is based on the people's experience versus a highly structured abstract representation of language? The results in this section provide a promising avenue to analyze this question.

Example 4: False recognition

The previous two sections applied EO to models that have been designed to learn the structure of language. This section applied EO to a processing model of memory, demonstrating the generality of the approach.

False memory is a phenomenon in which people strongly believe to have experienced items to which they had not been exposed. False memory has received a great amount of empirical attention and is known to be sensitive to linguistic structure. The dominant empirical paradigm is the Deese–Rodiger–McDermott paradigm (DRM; Deese, 1959; Roediger & McDermott, 1995).

In the DRM paradigm, subjects are asked to study items (e.g., *nurse*, *hospital*, *medicine*) that are related to a critical item (e.g., *doctor*), but the critical item itself is not presented during study. On subsequent tests of memory, subjects falsely recognize the critical item, and it is recalled at a rate similar to the studied items (for comprehensive reviews, see Brainerd & Reyna, 2005; Gallo, 2006).

Johns, Jones, and Mewhort (2012) proposed a comprehensive theory of both standard and false recognition, called the recognition through semantic synchronization (RSS) model. The RSS is based on the premise that if one wants to explain phenomena based in language (such as the majority of false memory studies, and all DRM experiments), the basis of the model's representation should be routed in material learned from the actual language environment. The representation is important because models of semantics construct very different similarity distributions than are typically assumed by memory models (e.g., they are heavy-tail distributed; Johns & Jones, 2010). Thus, in the RSS model, the word representations that are used are constructed with a semantic space model.

Processing in the RSS model is based on neural synchronization (Singer, 1999), where a probe word is attempted to be put into sync with an episodic memory trace (akin to trying to fit a puzzle piece in a slot). The efficiency of the synchronization process is determined by the amount of semantic information that is contained about a word in an episodic memory trace. Decision operates by accumulating information about whether a word occurred across the synchronization process.

The RSS model accounts for standard results in recognition memory and for a wide variety of false-memory results. The latter include levels of false recognition to different lists (Gallo & Roediger, 2002; Roediger & McDermott, 1995; Stadler, Roediger, & McDermott, 1999), item-level fits to different lists, effects of associative and thematic strength (Cann, McRae, & Katz, 2011; Hutchison & Balota, 2005), and developmental reversals in false recognition (Brainerd, Reyna, & Ceci, 2008; Brainerd, Reyna, & Forrest, 2002). Additionally, the RSS has been modified to account for recollection-based false memories (Johns, Jones, & Mewhort, 2014).

The RSS model can assess the effectiveness of experiential optimization as a general method in a model that has both representation and processing assumptions. For the purposes of the current article, the fact that the model is based on lexical

semantic representation means EO can be applied, and its benefit evaluated. RSS was fit to item-level differences in levels of false recognition to critical items across multiple DRM norms.

Data sources

The current simulation's goal was to maximize the model's fit to item-level differences in false-recognition rates. Different critical items elicit different levels of false recognition. For example, the critical item *window* is falsely recognized 87% of the time, while the critical item *king* is only falsely recognized 27% of the time (Stadler et al., 1999). Because it uses a realistic semantic representation, it can account for item-level variance, one of its appealing aspects.

The model's fit in Johns et al. (2012) to item levels of false recognition was an r = .41, p < .001, a value comparable to the best semantic predictor of false recognition, backward association strength. The model's fit was assessed across two sets of data: the normed lists from Stadler et al. (1999) and Gallo and Roediger (2002). Between these two studies there are 53 critical item lists. Given the success of EO so far, we expect the RSS to surpass the level of performance found by Johns et al. (2012) by a wide margin when augmented with EO.

Data-fitting methodology

Because the RSS uses paragraphs, or documents, to construct its semantic representation, the fitting method split each corpus into 5,000 paragraphs/documents. A larger paragraph size was used than before, because stabilizing the RSS model requires a significantly greater amount of computation than the previous models, and the amount of computation increases linearly with the number of documents integrated into the model's representation. There was a total of 325 sections used in this simulation.

To simplify the analysis, the same parameter set that had been used to model false recognition in Johns et al. (2012) was used here. That is, the internal cognitive parameters were kept constant, but the linguistic experience of the model was manipulated. Although manipulating both processing parameters and the experience that the model is exposed to should provide higher performance, the techniques to accomplish joint fitting have not been developed (because of the massive amount of computation that it would require: this is discussed further in the General Discussion).

To construct a comparison value for the fitted model, the best pure co-occurrence representation was constructed with EO. It is the co-occurrence representation that the model used, without any of the machinery of the RSS being used to make decisions. Additionally, keeping with the previous simulations, the model was trained with 25 random corpora, to form a comparison for the fitted representation.

Recall from the simulation in Fig. 7 that EO was unable to fit to randomly assembled data. A similar randomly assembled manipulation was repeated: the item-level probability of different critical words eliciting false recognition was shuffled randomly across critical words. The data were randomly shuffled 25 times, and the average item-level fit was assessed. The objective was to test whether EO can capitalize on random data.

Results

Figure 17 depicts the results of the simulation with the RSS augmented with EO. The figure also displays the fit of the original RSS, $r = .41$, $p < .001$, and the fit of the best EO optimized representation-only model, $r = 0.482$, $p < .001$. As before, EO could not optimize to the randomly shuffled data: it achieved an average fit of only $r = 0.251$ to the randomized data. The result corroborates the results in Fig. 7.

As Fig. 17 demonstrates, the EO-RSS model provides a substantial increase, even at only 5,000 paragraphs, $r = .59$, $p < .001$, and maximizes at an $r = .756$, $p < .001$ at 45,000 paragraphs. As a comparison, the model trained with randomly composed corpora performed similarly to the RSS trained with TASA, as shown in Fig. 8. By combining realistic process and representation types, better performance can be attained without positing additional complex parameters (cf. Estes, 1975).

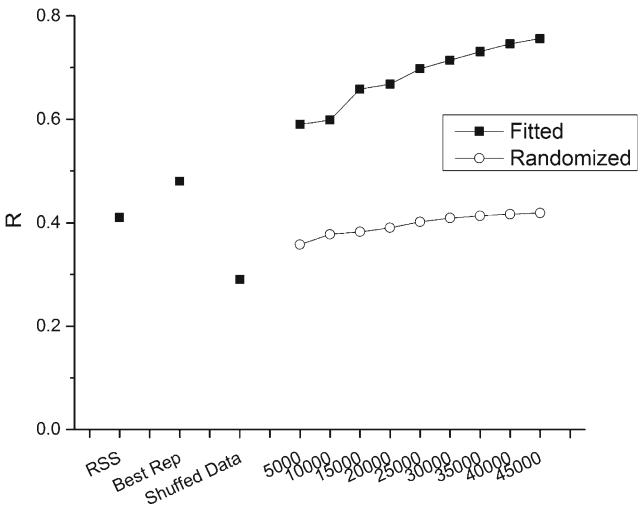

Fig. 17 Experiential optimization applied to the recognition through semantic synchronization (RSS) model of false recognition. The first point is the RSS's performance on the item-level fit to false recognition rates published in Johns et al. (2012), and the second point is the best fit that could be constructed by fitting the representation alone. The closed squares depict the optimized representation and the open circles depict the model trained with randomized corpora. By combining an optimized representation with a powerful processing mechanism, the model was able to account for behavior at a much greater resolution.

The model's fit is particularly impressive given that the best fit to this data is a regression model of seven variables (including behavioral data, such as veridical recall rates) obtained a fit of $r = .69$, $p < .001$ (Roediger, Watson, McDermott, & Gallo, 2001). That is, the EO-augmented cognitive model accounts for more variance in false recognition rates than a regression of all relevant variables (including many similar behavioral variables).

Discussion

The RSS model has proven to be highly successful at accounting for a variety standard and false memory effects (Johns, Jones, & Mewhort, 2012, 2014). The model's advantage is that it is based around a semantic representation learned from a corpus, making it easy to apply EO. The results show that EO is not just a way to optimize models designed to look at lexical semantic tasks; it can also be used to optimize models that are used to explain tasks that use language as stimuli, such as in studies of human memory.

General discussion

The current paper describes a new theoretical approach to optimize cognitive models, experiential optimization, developed to explore the power that variance in language has in accounting for lexical behavior. In EO, the information to which the model is exposed is manipulated to provide the best fit to a set of data. To do so, very large sets of texts spanning multiples subject areas and genres were assembled, including an online encyclopedia, product descriptions, and sets of fiction, non-fiction, and young-adult books. The corpora were split into small sections, and a simple hill-climbing algorithm was used to determine the best combination of these materials for a specific model and set of data. We demonstrated that fitting background knowledge with EO, when combined with experience-based cognitive models, provided benchmark fits to a number of tasks and areas, including semantic memory (Jones & Mewhort, 2007), lexical organization (Jones et al., 2012), sentence processing (Johns & Jones, 2015), and false memory (Johns, Jones, & Mewhort, 2012).

The underlying philosophy of the method is similar to standard parameter-fitting methods commonly applied to process parameters (Shiffrin et al., 2008): we assume that there is natural variability in the parameters that define the cognitive processes underlying behavior and that there is natural variability in the knowledge that different subject groups have experienced. These differences lead to variability in behavior. Likewise, there is natural variability in the experience that people have had with language. By optimizing the experience that groups are assumed to have had, it is possible to dramatically increase the performance of a model with EO.

The fundamental point of this article is that language-based models depend on the content of experience that is given to them. Similar to a model given an incorrect parameter set, a model that has an insufficient or incorrect knowledge base is going to give a poor accounting when tested. That is, even a wonderful model of natural language can be proposed that does not function appropriately because it does not have the correct experience to do so. EO provides a framework under which models of language and memory can be developed that eliminates this possibility. Just as standard parameter fitting techniques point to the need to control the process components of cognition, EO points to the fact that it is equally necessary to control the experiential factors of a cognitive model.

One exciting aspect of EO is that it provides a framework with which to discriminate the contributions of internal cognitive mechanisms and external information sources, one of the classic goals of cognitive science (Anderson & Schooler, 1991; Estes, 1975; Simon, 1969; Tomasello, 2003). If one accepts that language is dictated by a complex interaction of biological and cultural evolution (Chater, Reali, & Christiansen, 2009; Christiansen & Chater, 2008, 2016; Tomasello, 2010), it is necessary to determine how much is derived from evolved mechanisms and how much is explained by the heavily structured environment in which humans are embedded in (and the domain-general learning mechanisms designed to exploit that redundancy). The simulations reported here provide substantial evidence that the content of the information that a model knows is important to its behavior, underscoring the fact that human behavior is sensitive to the knowledge that a person has gained from experience. Further empirical and theoretical work will allow us to examine these issues at a much finer-grain than is reported here (see Nelson & Shffrin, 2013; Wells et al., 2009, for important empirical paradigms used to study these problems). The simulation reported in Fig. 12 (in which EO selected different corpus constructions to account for lexical-decision data in younger and older adults) is a promising first step towards showing that group-level experience can be estimated.

A related application of the method is to differentiate tasks by how dependent task behavior is on lexical experience versus internal cognitive processing parameters. As is demonstrated in Fig. 7, by randomizing a set of data and comparing the fits of intact data to the randomized data, different patterns are found for different tasks. Specifically, semantic priming showed a much smaller advantage when compared to optimizing to the TOEFL test and word-pair similarity data. This suggests that semantic priming may depend more on cognitive processes and individual differences than it is on lexical experience (cf. Yap et al., 2016), especially when compared to synonymy tests or similarity ratings. For future research, the contrast promises a way to differentiate the cognitive and experiential components of different lexical behaviors.

More generally, the present work points to the importance of models capable of extracting information from large text bases, an issue that has been explored in greater detail elsewhere (e.g., Chubala et al., 2016; Hills, Jones, & Todd, 2012; Johns & Jones, 2010; Johns, Jones, & Mewhort, 2012; Johns & Jones, 2015; Johns, Taler, et al., 2017; Mewhort et al., 2017; Taler et al., 2013). Basing a model's performance on large-scale environmental information provides a strong case for the model's plausibility, because it can scale to human levels of experience. The tests in this article show that making a model experience-dependent allows us to examine how much additional power it has as a function of specific experience.

Additionally, EO provides a framework for the general optimization of distributional models. Optimization is especially relevant for applied problems, such as assessing performance of cognitively impaired patients on common neuropsychological tasks (e.g., Johns, et al., 2018). Given a requirement (such as discriminating the performance of cognitively impaired patients from cognitively normal subjects on a semantic task), the model can be optimized to include the most relevant lexical sources, thereby giving the model its best shot at successfully performing the task. EO would ensure that, given sufficient diversity in the lexical sources, a model would fail because of a lack of the correct lexical information but rather because of other aspect of the task's requirements. EO provides a flexible framework within which to embed distributional models; it provides a model with the ability to adapt its knowledge base to any new task that is required of it.

In order to use EO to optimize a model's fit to tasks outside of behavioral data, such as performing classification, will require cross-validation based studies, to ensure that the model is not overfitting the results of one group of people, but can also adapt to another group. However, this changes the underlying philosophy of the technique, as it is no longer being used to estimate the type of information that a group of subjects (or an individual subject) used to accomplish a specific task. Instead it becomes an estimate of the set of language materials that will generalize to the largest number of people as possible. The goal of EO is to optimize and gain additional understanding into the performance of a given model, but this does not preclude its use in machine learning applications.

From an individual learning perspective, the opposite is also important: how much better can an individual learn from materials that are coherent with their past experience? Lexical organization models, such as the SDM (Jones et al., 2012), suggest that learning is dynamic: how much one acquires from a specific experience is dependent on what one has learned previously. Thus, given knowledge about the specific types of experiences an individual has had, it is possible to present materials that should allow for that individual to optimally acquire new lexical information.

As touched on above, an issue with all optimization procedures is the danger of overfitting to data. We have attempted to examine this issue as it pertains to EO in the simulation contained in Figs. 12, 13 and 14. However, there is still considerable research required to examine the connection between lexical experience (e.g. reading the book *To Kill a Mockingbird*) and lexical behavior. Ideally, if a subject had read *To Kill a Mockingbird*, and if lexical experience has a truly measurable impact on lexical behavior, then EO should select this book when optimizing to that subject's behavior. However, most current data collection procedures within cognitive psychology preclude this type of analysis. Specifically, this type of analysis requires large amounts of item-level data, averaged for an individual, a difficult and task-sensitive requirement. It also requires knowing the specific lexical experiences that an individual has had. For experiential accounts of language to continue to develop, it will become increasingly necessary to develop the methodology to understand the impact that individual experiences with language have on the language processing system, with both controlled (e.g., Johns, Dye, et al., 2016) and large-scale experimentation.

A related question to the issue of overfitting is that of model complexity. As Jones, Hills, and Todd (2015) point out, corpus-based models do not conform to standard methods of model complexity. Specifically, classic methods of model complexity (e.g. the Bayesian Information Criterion; Schwarz, 1978) penalize models for having a greater number of parameters. It is not clear how experience-based models of cognition fit into this framework, and whether a model trained on a larger corpus should be considered more complex. As experience-based approaches to cognition continue to develop, there is a need to develop both conceptual and mathematical frameworks to determine the place of this model type in cognitive theory.

One obvious limitation of EO is that it uses a simple hill-climbing algorithm as its optimization method—the selection of hill climbing was purposeful due to its extreme simplicity as a search algorithm. It allowed the overall power of EO to be demonstrated, but there are many published techniques that could be used to increase the power and efficiency of the method. However, unlike many parameter-fitting techniques, experiential optimization is not based on optimizing an underlying function, but, instead, is more akin to optimal sampling. This is due to it being difficult to define the connection between any two selections of language, a limitation that impedes the use of most standard methodologies.

Certain techniques are better suited to experiential optimization than others. For example, population-based methods, such as genetic algorithms (Davis, 1991; Mitchell, 1998), may provide a promising avenue to further this approach, as it does not have to be based on functional approximation, but can be more akin to efficient random search. It is possible that the structure of semantic space could be defined with a great

enough resolution that stochastic optimization algorithms could be applied, but more research needs to be done into defining how linguistic materials combine to form new representations (which is also a model dependent problem).

As Simon (1969) described, in order to provide a complete account of behavior, it is necessary to understand both the internal mechanisms and the environmental information that people use to behave. This is especially important in the study of language, as the vast majority of linguistic theories have focused on the internal mechanisms that are responsible for linguistic behavior, while the influence of environmental information has been downplayed. This was necessary because of a lack of both large amount of texts and computational resources, but neither of these are modern limitations. It is readily possible to examine the impact of linguistic information on human behavior, and by optimizing the linguistic information that a model is exposed to, it allows for a powerful test of a model's ability to account for behavioral data.

Author note Part of this work was presented at the 37th Meeting of the Cognitive Science Society. We would like to thank Rich Shiffrin for feedback during the writing of this manuscript. This research was supported by IES R305A150546 to M.N.J.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275-290.

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

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.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009, May). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27). Stroudsburg, PA: Association for Computational Linguistics.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 339,* 445–459.

Balota, D. A., Cortese, M. J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society* (p. 44). Los Angeles, CA: Psychonomic Society.

Bannard, C., Lieven, E., & Tomasello, M. (2008). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*, 17284–17289.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577–660.

Bloom, P. A., & Fischler, I. S. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition*, *8*, 631–642.

Brainerd, C. J., & Reyna, V. F. (2005). The science of false memory. Oxford, UK: Oxford University Press.

Brainerd, C. J., Reyna, V. F., & Forrest, T. J. (2002). Are young children susceptible to the false-memory illusion? *Child Development*, *73*, 1363–1377.

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin, 134,* 343.

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

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.

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). 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.

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

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*, 890–907.

Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese-Roediger-McDermott paradigm: The roles of gist and associative strength. *The Quarterly Journal of Experimental Psychology, 64,* 1515–1542.

Chater, N., Reali, F., & Christiansen, M. C. (2009). Restrictions on biological evolution in language evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 1015–1020.

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

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., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23,* 371–414.

Davis, L. (Ed.). (1991). Handbook of genetic algorithms. New York, NY: Van Nostrand Reinhold.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

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.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*, 116–131.

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

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation, 14*, 375–399.

Gallo, D. A. (2006). Associative illusions of memory: False memory research in DRM and related tasks. New York, NY: Psychology Press.

Gallo, D. A., & Roediger, H.L. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language*, *47*, 469–497.

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

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*, 787–814.

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

Hare, M., Tanenhaus, M. K., & McRae, K. (2007). Understanding and producing the reduced relative construction: Evidence from ratings, editing and corpora. *Journal of Memory and Language, 56*, 410–435.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition, 111*, 151–167.

Hills, T. (2012). The company that words keep: Comparing the statistical structure of child versus adult-directed language. *Journal of Child Language, 40,* 586–604.

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

Hills, T., Maouene, J., Riordan, B., & Smith, L. (2010). The associative structure of language and contextual diversity in early language acquisition. *Journal of Memory and Language, 63,* 259–273.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

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.

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. (2005). Decoupling semantic and associative information in false memories: Explorations with semantically ambiguous and unambiguous critical words. *Journal of Memory and Language*, *52*, 1–28.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*, 1036–1066.

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., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, *62*, 550–575.

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction time task: Anticipating from experience. *Quarterly Journal of Experimental Psychology*, *62*, 1757–1783.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String-completion and performance for individual items. *Quarterly Journal of Experimental Psychology*, *63*, 1014-1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: A reply to Kinder (2010). *Quarterly Journal of Experimental Psychology*, *64*, 209–216.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of distributional semantics. In C. Kalish, M. Rau, J. Zhu, & T. T. Rogers (Eds.), Proceedings of the 39th Conference of the Cognitive Science Society. Austin TX: Cognitive Science Society.

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

Johns, B. T., Jamieson, R. K., Crump, M. J. C., Jones, M. N., & Mewhort, D. J. K. (2016). The combinatorial power of experience (pp. 1325–1330). In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society.. Austin, TX: Cognitive Science Society

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. (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. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*, 486–518.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2014). A continuous source reinstatement model of true and illusory recollection. In P. Bello, M. Gurarini, M. McShane, & B. Scassellayi (Eds.), Proceedings of the 36th annual Cognitive Science Conference (pp. 248–253). Austin, TX: Cognitive Science Society.

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

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., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., . . . Unverzagt, F. W., & Jones, M. N. (2017). Cognitive modeling as an interface between brain and behavior: Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology, 72*(2), 117–126. doi:https://doi.org/10.1037/cep000013

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

Jones, M. N., & Dye, M. W. (2018). Big data methods for discourse analysis. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), Handbook of discourse processes (2nd ed., pp. 117–124). New York, NY: Routledge.

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 (Vol. 67, p. 43). New York, NY: Academic Press.

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., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

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

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day English*. Providence, Rl: Brown University Press.

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.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text (pp. 165–172). In *Proceedings of the 7th ACM Conference on Recommender Systems* (RecSys). New York, NY: ACM.

Mewhort, D. J. K., Shabahang, K. D., & Franklin. D. R. J. (2017). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review*. doi:https://doi.org/10.3758/s13423-017-1327-3

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L, Bottou, M., Welling, Z., Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Retrieved from https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language & Cognitive Processes*, *6*, 1–28.

Mitchell, M. (1998). An introduction to genetic algorithms. Cambridge, MA: MIT Press.

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

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2017). Model evaluation and selection. In W. H. Batchelder, H. Colonius, E. Dzhafarov & J. I. Myung (Eds.), New handbook of mathematical psychology, Vol. 1: Measurement and methodology (pp. 552–598). Cambridge, UK: Cambridge University Press.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*, 356–394.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*, 5–42.

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological Science*, *28*, 1171–1179.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1290–1301.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*, 504–509.

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. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods*, *41*, 657–663.

Recchia, G. L., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. *Computational Intelligence & Neuroscience*. doi: https://doi.org/10.1155/2015/986574

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*, 337–354.

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.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 803–814.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*, 385–407.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8,* 627–633.

Schwanenflugel, P. J. (1986). Completion norms for final words of sentences using a multiple production measure. *Behavior Research Methods, Instruments, & Computers*, *18*, 363–371.

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

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171,* 701–703.

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, *2*, 736-750.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Simon, H. A. (1969). The sciences of the artificial. Cambridge, MA: MIT Press.

Singer, W. (1999). Neural synchrony: A versatile code for the definition of bindings. *Neuron*, *24*, 49–65.

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create memories. *Memory & Cognition*, *29*, 424–432.

Stone, B., Dennis, S., & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science, 3,* 92–122.

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.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*.

Tomasello, M. (2010). Origins of human communication. Cambridge, MA: MIT Press.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271.

Yap, M. J., Hutchison, K. A., & Tan, L. C. (2016). Individual differences in semantic priming performance: Insights from the Semantic Priming Project. In M. Jones (Ed.), Big data in cognitive science (pp. 203–226). New York, NY: Psychology Press.