Psychological Review
2009, Vol. 116, No. 3, 463–498

0033-295X/09/$12.00 DOI: 10.1037/a0016261

Integrating Experiential and Distributional Data to Learn Semantic Representations

Mark Andrews, Gabriella Vigliocco, and David Vinson
University College London

The authors identify 2 major types of statistical data from which semantic representations can be learned. These are denoted as *experiential data* and *distributional data.* Experiential data are derived by way of experience with the physical world and comprise the sensory-motor data obtained through sense receptors. Distributional data, by contrast, describe the statistical distribution of words across spoken and written language. The authors claim that experiential and distributional data represent distinct data types and that each is a nontrivial source of semantic information. Their theoretical proposal is that human semantic representations are derived from an optimal statistical combination of these 2 data types. Using a Bayesian probabilistic model, they demonstrate how word meanings can be learned by treating experiential and distributional data as a single joint distribution and learning the statistical structure that underlies it. The semantic representations that are learned in this manner are measurably more realistic—as verified by comparison to a set of human-based measures of semantic representation—than those available from either data type individually or from both sources independently. This is not a result of merely using quantitatively more data, but rather it is because experiential and distributional data are qualitatively distinct, yet intercorrelated, types of data. The semantic representations that are learned are based on statistical structures that exist both within and between the experiential and distributional data types.

Keywords: semantic representations, probabilistic models, Bayesian models, computational models, distributional data

In this article, we consider the topic of how human semantic representations may be learned by integrating two distinct types of statistical data. Throughout the article, we use the term *semantic representation* to refer to the mental representation of the meaning of words. This we define informally as the knowledge underlying the ability to make inferences about, for example, which words are synonymous, which are similar, what are the various senses of a given word, what (if anything) are its referents, and so on. The theoretical objective of the article is to consider how this knowledge is acquired. We address this question specifically by asking what the different types of statistical data from which human beings can gain this knowledge are and how these data can be integrated to form semantic representations.

For the purpose of this article, we concentrate on two broad and general types of statistical data. We refer to the first as *experiential data* and the second as *language-based distributional data,* or just simply *distributional data.* Experiential data specify the perceived physical attributes or properties associated with the referents of words. For example, the word *apple* refers to objects in the world whose perceived attributes or properties include being red or green, round, shiny, smooth, crunchy, juicy, sweet, tasty, and so on. As we use the term, experiential data are the entirety of the data obtained directly through sense receptors and from which people gain their knowledge of the world. We also use the term in a more general sense to include affective properties, such as whether something is pleasant, unpleasant, fearsome, and so on, and to include *affordances* (Gibson, 1977, 1979; Norman, 1988), or the properties that tell people how to interface and interact with an object. The second type of data that we consider is distributional data. Distributional data specify how a given word is statistically distributed across different spoken or written texts. We use the term *text* here in a very general sense to refer to any coherent and self-contained piece of written or spoken language. This could include, for example, a newspaper article, a spoken conversation, a letter or e-mail message, an essay, a speech, and so on. If we divide a corpus of spoken and written language into a set of separate texts, the distribution of a given word across this corpus is simply whether and how often it appears within each text.

As we describe them in this article, experiential and distributional data represent data types that are, respectively, extralinguistic versus intralinguistic in their origin. In other words, experiential data are data derived from human perception and interaction with the physical world, while distributional data are derived from the statistical characteristics within a language itself. It is reasonable to assume and, as we explain below, it has been empirically verified that semantic

Mark Andrews, Gabriella Vigliocco, and David Vinson, Cognitive, Perceptual and Brain Sciences, Division of Psychology and Language Sciences, University College London, London, United Kingdom.

This research was supported by European Union (FP6-2004-NEST-PATH) Grant 028714 and U.K. Biotechnology and Biological Sciences Research Council (BBSRC) Grant 31/S18 to Gabriella Vigliocco and by U.K. Economic and Social Research Council (ESRC) Grant RES-620-28-6001 to the Deafness, Cognition and Language Research Centre, University College London. All the codes necessary to simulate the models described in this article are available at http://www.mjandrews.net/code.

Correspondence concerning this article should be addressed to Mark Andrews, Cognitive, Perceptual and Brain Sciences, Division of Psychology and Language Sciences, University College London, Gower Street, London WC1 6BT, United Kingdom. E-mail: m.andrews@ucl.ac.uk

representations can be learned from either one of these sources. For example, part of the meaning of the word *apple* can be learned from the fact that it refers to that set of objects with the properties or attributes like those mentioned above (i.e., being red or green, round, shiny, smooth, crunchy, juicy, sweet, tasty, etc.). Words like *peach, pear,* or *apricot* refer to objects with broadly similar characteristics and can thus be inferred to be similar in meaning. On the other hand, the word *apple* also occurs within phrases like *the leaves of the apple tree, a glass of apple juice, roast pork with apple sauce,* and so on. Words like *cherry, cranberry,* or *pear* can also be found within broadly similar linguistic contexts, and on this basis, irrespective of their possible physical referents, these words can be inferred as being semantically similar.

As we describe below, the contribution of either experiential or distributional data to the learning of human semantic representations has been studied extensively in recent literature within cognitive science. For the most part, throughout this literature, however, the contribution of either one of these data types has been considered independently and to the exclusion of the other. In almost all cases, attention has been focused exclusively upon either experiential data alone or upon distributional data alone. The primary objective of this article is to consider the combined effects of both sources of data.

Our theoretical proposal is that experiential and distributional data both represent major sources of data from which humans can learn semantic representations. In particular, we propose that the statistical patterns describing the structure of experiential data and those describing the structure of distributional data can be jointly used and combined to learn semantic representations. As such, our primary hypothesis, one that we make precise in later sections, is that semantic representations are the product of what we call the statistical combination of experiential and distributional data types. By this, we mean that semantic representations are derived from the optimal statistical combination of experiential and distributional data, rather than either relying primarily upon one source alone or by simply averaging over the separate effects of both sources.

Two Traditions in the Study of Semantic Representations

In the past literature on human semantic representations, the contributions of experiential and distributional data have been approached largely independently and to the exclusion of one another. This has led, in effect, to two broad traditions in the contemporary study of semantic representation. The first tradition, what we call the *experiential tradition,* describes semantic representations in terms of the attributes or properties associated with words. As such, it argues largely in favor of the primacy of extralinguistic data, emphasizing how semantic representations are learned from human experience and interaction with world. By contrast, the second or *distributional tradition* describes semantic representations in terms of the statistical patterns that occur within a language itself. This tradition argues for the primacy of intralinguistic data and emphasizes how semantic representations are learned by way of human experience and use of language. Each one of these traditions can be seen to have distinct historical origins and philosophical antecedents, and these backgrounds have shaped the nature and focus of the two traditions in contemporary cognitive science research.

Experiential Tradition

The experiential tradition is broadly based on a philosophical perspective that characterizes the meaning of words in terms of objects and events in the world. Specifically, this perspective characterizes the meaning of a word as corresponding to the mental representation of, for example, the object in the world to which that word refers, where these representations are ultimately based on the set of perceived physical properties of the object. This general philosophical perspective can be seen to have its origin in early modern empiricism, particularly that of Locke (1632–1704). Locke's perspective on human cognition is described in *An Essay Concerning Human Understanding* (Locke, 1689/1975). For Locke, all knowledge is ultimately based upon *sensible qualities,* or sensory data derived through various sensory modalities. Locke's theory of word meaning is that the meaning of a word is the mental representation of the object to which it refers, or, as Locke put it, "words in their primary or immediate signification stand for nothing but the ideas in the mind of him that uses them" (Locke, 1689/1975, Book III, Chapter II, Part 2). The semantic representation of a word like *apple* is simply the concept or representation of an apple. This representation is, according to Locke, a hierarchical composition of the elementary perceptual attributes of an apple such as its size, shape, color, and so on. For Locke, word meanings are thus represented as patterns over the properties of objects in the world.

The empiricist theoretical perspective exemplified by the work of Locke has been widely adopted throughout the study of human semantic representations within cognitive psychology and cognitive neuroscience. This perspective is clearly evident in the seminal works in this area. For example, in Quillian (1967, 1969) and Collins and Quillian (1969), knowledge is represented in hierarchical terms: General categories are described in terms of subcategories or their constituent objects, while the constituent objects are described in terms of their perceived attributes and properties. Likewise, E. Smith, Shoben, and Rips (1974) explicitly defined the semantic representation of a word in terms of the set of properties associated with its referent. Following Rips, Shoben, and Smith (1973), E. Smith et al. proposed that semantic memory can be described as multidimensional space, the dimensions of which are this set of properties. This work has led naturally to a more statistical interpretation of the problem of semantic representation. According to this view, words correspond to sets of perceived properties or, equivalently, to points in a high-dimensional space. Learning semantic representations corresponds to learning the intrinsic statistical structure of this space. We can find this general statistical perspective in the early models of distributed memory presented by McClelland and Rumelhart (1985). In this work, words are explicitly described as distributions over elementary attributes, and neural network learning models are used to learn the structure of these distributions. In cognitive neuroscience, these same principles and techniques underlie the models of deep dyslexia by Hinton and Shallice (1991) and Plaut and Shallice (1993) and the models of category-specific deficits by Farah and McClelland (1991); Devlin, Gonnerman, Andersen, and Seidenberg (1998); and Tyler, Moss, Durrant-Peatfield, and Levy (2000).

In the study of semantic representation in normal or cognitively unimpaired subjects, we see these principles in the work of McRae, de Sa, and Seidenberg (1997). In this work, a speaker-

generated data set of 190 common nouns, each described in terms of 1,242 elementary features, was used. Using an attractor network to learn the distributions over word-form and semantic-feature representations of these 190 words, it was confirmed that the distances between the representations of these words in the attractor network predicted human judgments on the similarities between words as evidenced by a priming task. In related work, Vigliocco, Vinson, Lewis, and Garrett (2004) collected feature norms for a set of 456 common words, comprising 240 nouns and 216 verbs. Participants described these words in terms of a set of 1,029 elementary features. A self-organizing map was used to learn the low-dimensional structure of this data. As in the work of McRae et al., it was found that words described in terms of this semantic space were predictive of human similarity judgments, as evidenced by semantic priming, picture-word interference (PWI), and error induction. Further work by McClelland and Rogers (2003) and Rogers and McClelland (2005) has shown that knowledge derived from the distribution of attributes associated with words is sufficient to explain the formation of hierarchical categorical knowledge as in Quillian (1967, 1969), the progressive differentiation throughout development of semantic categories into finer subsets (Keil, 1979; Mandler, 2000; Mandler, Bauer, & McDonough, 1991), the early emergence of the so-called basic level of categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), and inferences about the relative importance of certain properties within different categories (S. A. Gelman & Markman, 1986; Macario, 1991), a phenomenon often otherwise explained by way of innate knowledge of conceptual categories (Carey, 1985; Keil, 1979, 1989).

Recently, there has been compelling evidence that the representation of words in the brain is in terms of distributed patterns over the sensory-motor properties of their referents. For example, the premotor and motor cortices have been found to be consistently activated by language referring to body actions (Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni, 2006; Pulvermüller, 1999, 2001; Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005; Tettamanti et al., 2005; Vigliocco et al., 2006), tool actions, tools, or manipulable objects (Chao & Martin, 2000; Gerlach, Law, & Paulson, 2002; Grabowski, Damasio, & Damasio, 1998). Transcranial magnetic stimulation studies have provided converging evidence that lexical and sentential items with motor associations activate motor areas of the cortex (Buccino et al., 2005; Oliveri et al., 2004) and localized motor cortical areas corresponding to the specific effector of an action (Buccino et al., 2005; Pulvermüller et al., 2005). Alongside the motor cortex, mediotemporal activity is repeatedly seen for body and tool actions as well as tool objects (Damasio et al., 2001; Martin, Haxby, Lalonde, Wiggs, & Ungerleider, 1995; Martin, Wiggs, Ungerleider, & Haxby, 1996; Phillips, Noppeney, Humphreys, & Price, 2002; Tettamanti et al., 2005). There is also some evidence that it is active during comprehension of words referring to fruit or an object's form (Phillips et al., 2002; Pulvermüller & Hauk, 2005). With regard to sensory properties, the fusiform gyrus is documented as playing a role in the representation of object form (Chao, Haxby, & Martin, 1999; Vuilleumier, Henson, Driver, & Dolan, 2002), and different areas of the fusiform have been implicated for different categories, namely, lateral fusiform for animals and medial fusiform for tools (Martin & Chao, 2001). In a series of experiments, for example, Martin and colleagues (Beauchamp, Lee, Haxby, & Martin, 2002; Chao et al., 1999; Chao & Martin, 2000; Chao, Weisberg, & Martin, 2002; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999) showed that naming objects referring to different semantic categories activated a broad and largely overlapping region of the ventral and lateral temporal cortex but that the profile of activation differed depending on category. This suggests that object concepts are represented according to object features, rather than according to semantic categories corresponding to specific and anatomically segregated modules. These results support the role of the fusiform in representing the visual attributes of known objects, and more generally, this area of the cortex as involved in higher order visual association, combining features from different modalities (Vigliocco et al., 2006).

Distributional Tradition

If the philosophical background of the experiential tradition can be traced back to at least as early as 17th century empiricism, the corresponding philosophical background of the distributional tradition is of a more recent vintage. Wittgenstein (1953/1997) famously proposed that "for a large class of cases . . . in which we employ the term *meaning,* it can be defined thus: *The meaning of a word is its use in the language*" (Section 43). Wittgenstein was arguing against what he perceived to be a pervasive conception of meaning and language. According to this view, language is a mirror of the world: Words refer to objects in world, while propositions consist of words arranged into a structure that will mirror the interrelationships of their referents. Against this traditional view, Wittgenstein presented his alternative conception of language premised upon the idea that the meaning of word is based on how it used within a language. Rather than pointing to something exterior to the language, a word's meaning is determined by the role it plays within the language itself.

Although the precise implications of Wittgenstein's ideas were (and, indeed, still remain) elusive, Firth (1957) was inspired by Wittgenstein's characterization of language to propose that humans can learn the meaning of a word by examining the various contexts and circumstances of its common usage. Firth suggested that "you shall know a word by the company it keeps" and that human beings learn at least part of the meaning of a word from "its habitual collocation" with other words (Firth, 1957, p. 11). For Firth, words that are found in identical or similar environments can be taken to share at least some of their meanings. In a similar manner, and writing contemporaneously to Firth, Harris (1954) proposed the *distributional hypothesis* whereby word meanings are derived in part from their distribution across different linguistic environments. Harris suggested, for example, that "if (two words) A and B have almost identical environments . . . we say they are synonyms," while "if A and B have some environments in common and some not . . . we say that they have different meanings," with "the amount of meaning difference corresponding roughly to the amount of difference in their environments" (Harris, 1954, p. 157).

Although the idea that word meanings may be learned from the distribution of words across a corpus was first mooted as early as the 1950s, this hypothesis was not seriously considered, and it is likely that its plausibility remained dubious, until computing resources grew to sufficient power. In this respect, one of the first attempts to adequately address this hypothesis was due to Schütze

(1992). In this work, following the common practice in information retrieval (Salton & McGill, 1983), the distribution of a word across a corpus is represented by a vector describing that word's frequency of co-occurrence with every other word. As such, each word can be viewed as a point in high-dimensional space, and matrix factorization is used to find the intrinsic dimensionality of this space. Every point in the original high-dimensional space can be represented in the lower dimensional space, and the distance between these points is taken as a measure of the dissimilarity between the corresponding words.

The work of Schütze (1992) was to strongly influence the work of Lund and Burgess and their hyperspace analog of language (HAL) model(e.g., Burgess & Lund, 1997; Lund & Burgess, 1996; Lund, Burgess, & Atchley, 1995). The HAL model was motivated as an attempt to automatically derive a model of human semantic memory from the statistics in a text. Using a large Usenet corpus, the HAL model describes each word as a high-dimensional co-occurrence vector precisely as in Schütze. The dimensionality of this space is reduced by simply removing all but the 200 highest variance columns. In Lund and Burgess (1996), it was shown that in this lower dimensional space, semantic categories like, for example, *animals, body parts,* and *geographical regions* form clusters that are easily distinguishable from one another. Lund and Burgess also showed that the distances between word pairs in the HAL space correlate positively, with a coefficient of up to $r = .35$, with semantic-priming reaction times, that is, where these word pairs are the prime-target pairs in a lexical decision task.

The latent semantic analysis (LSA) work of Landauer and colleagues (e.g., Landauer & Dumais, 1997; Landauer, Laham, & Foltz, 1998; Landauer, Laham, Rehder, & Schreiner, 1997) was influenced by the information-retrieval work of Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) but is comparable to the HAL model in its modeling objectives. In LSA, words are described as high-dimensional vectors indicating the extent to which they occur in a large set of documents in a corpus. Matrix factorization is used to find the intrinsic structure of this space, and the cosine of the angle between vectors in this reduced space can be taken as a measure of interword similarity. Using this measure, Landauer and Dumais (1997) reported comparable performance between LSA and humans. Using 80 items from a Test of English as a Foreign Language synonym test, the authors reported a performance of 64.4% correct for LSA, compared to 64.5% by nonnative English speakers applying to U.S. colleges. They also showed that in cases where LSA was incorrect, its choices were positively correlated ($r = .44$) with the choices made by the college applicants. In Landauer et al. (1998), it was further reported that LSA performed comparably to children ($r = .5$) and adults ($r = .32$) on a word-sorting task.

More recently, the work of Griffiths and Steyvers (2002, 2003) and Griffiths, Steyvers, and Tenenbaum (2007) provided a probabilistic model of human semantic representation that is based upon the latent Dirichlet allocation (LDA) model of Blei, Ng, and Jordan (2003). The LDA model can be seen as a probabilistic generalization of LSA whereby each text in a corpus is a probabilistic weighting of a set of discourse topics, with each discourse topic corresponding to a probability distribution over words that emphasizes a certain theme. For example, the discourse topic labeled *sport* may place most of its probability mass on words like *game, ball, play, team, competition,* and so on. The aim of learning in these models is to infer the component topics. Each word in the model can then be represented as a distribution over these latent topics, and this can be taken to be its semantic representation. Using these representations, the relationships between words, namely, interword similarities, can also be inferred. For the purposes of comparison between the similarity relationships inferred by the model and those of human judgments, the Nelson word-association norms (Nelson, McEvoy, & Schreiber, 2004) were used as a standard. Griffiths and Steyvers (2003) and Griffiths et al. reported that across a large set of words, the most highly related word according to their model was exactly that of the most highly associated word according to the norms. They also reported superior performance of their model, according to this measure, in comparison to the original LSA model of Landauer and Dumais (1997).

The Necessity of Combining Data Types

From what we have reviewed so far, it is evident that on the basis of either experiential data alone or distributional data alone, intuitively correct semantic relationships may be derived and that, moreover, these have psychological validity as evidenced by comparisons with human-based measures of semantic similarity. However, in all of the studies that we have reviewed, the focus of attention has been upon one of these data types alone, independent and to the exclusion of the other. In what follows, we show that there are obvious problems with any perspective that advocates the importance of one data type to the total exclusion of the other.

An obvious criticism of experiential data is that they are largely limited to the so-called concrete terms, that is, those words that have tangible physical referents or instantiations. Concrete terms constitute only a small subset of the commonly used words in the language. Most other terms, even if we exclude the function words whose role is primarily syntactic, do not have obvious physical instantiations. These terms include not just the canonical examples of abstract terms such as *truth, art, justice,* and so on but also more mundane terms such as *government, finance, crime,* and so on. While these latter terms can be, either directly or indirectly, related to objects or events in the world, the meaning of these terms is in no way exhausted by these referents. It is arguable that their meaning is fully appreciated only in the context of a much richer body of knowledge about, for example, how people, economies, and societies work and that this body of knowledge can be only be acquired through language and the representational medium it affords. Knowledge acquired by way of the physical or experiential attributes of the referents of these terms, albeit substantial and important, is nevertheless inadequate to fully account for their semantic representations.

In contrast, a fundamental criticism of distributional data is that they are disconnected or disembodied from the physical world. In other words, distributional data describe the relationship of words only to one another but not to the physical world or anything else beyond language itself. This fact alone is taken by some as an a priori argument against the plausibility of distributional models as accounts of human semantic representation. For example, Glenberg and Robertson (2000) argued that because distributional approaches propose that "the meaning of an abstract symbol (a word) can arise from the conjunction of relations to other undefined abstract symbols" (p. 381), this is grounds for their rejection

as plausible models of semantic representation. For Glenberg and Robertson, to know "the meaning of an abstract symbol such as . . . an English word, the symbol has to be grounded in something other than more abstract symbols" (Glenberg & Robertson, 2000, p. 382). Even if this perspective is not found to be convincing, the disembodied nature of distributional data does present real challenges for distributional models of semantic representation. For example, as a consequence of their disembodied character, distributional models cannot account for any of the previously described neuroscientific evidence showing that words are represented in the brain according the sensory-motor characteristics of their referents. More generally, distributional models cannot explain how any knowledge acquired by way of distributional data can be related back to the world. Although two words may have similar distributional patterns and from that it may be inferred that they are semantically related, what they refer to in the world still cannot be known, nor can any world knowledge, otherwise acquired, be integrated with knowledge derived from distributional patterns. For example, on the basis of distributional patterns, it may be inferred that the terms *dog* and *cat* are related, but nonetheless, it is not clear that these words also refer to familiar domestic animals and pets, nor could any knowledge of these domestic creatures, acquired through interaction with them in the world, be integrated with this distributional-based knowledge. For knowledge acquired from language to be pragmatically useful, it must ultimately relate back to the world. It is challenging to explain how this can be the case if words are known only through their relationship with other words.

It is not, however, necessary to choose between experiential and distributional data as if they were mutually exclusive. Both types of data are available to humans when learning the meaning of words. Words are encountered simultaneously within two rich contexts: the physical world itself and the discourse of human language. As such, it is reasonable to assume that both data types are used concurrently to learn word meanings. For example, one can imagine the following scenario: On the basis of perceptual experience, a child learns that the term *dog* refers to creatures that make barking noises, have four legs and waggy tails, and so on. In addition, through general experience with language, the child learns that the term *dog* co-occurs with terms like *pet, animal,* and so on. In this learning scenario, it is as if there is a dual, or parallel, corpus of data. On the one hand, there is the stream of words that is the language itself, and on the other, there is the set of perceived properties associated with (at least a subset of) the words. Knowledge that the word *dog* refers to those creatures that have four legs and tails, that bark, and so on can be integrated with the knowledge that *dog* co-occurs with *pet, animal,* and so on. The two sources of information could then be combined to provide a richer understanding of the semantics of the word *dog* than could be learned by either source alone.

In this article, we describe how experiential and distributional data can be combined to learn semantic representations. The manner in which we model this follows the same general rationale as that of the statistical models reviewed so far. While, in those models, the objective was to infer the statistical structure underlying either experiential data alone or distributional data alone, in the models we use here, we aim to infer the statistical structure underlying the joint distribution of both data types.

To our knowledge, the joint role of experiential and distributional data in the learning of semantic representations has yet to be thoroughly investigated. There are, however, some notable studies that relate, or are precursors, to this work. The well-known dual-coding theory of Paivio (e.g., Paivio, 1971, 1986/1990) was one of the first theories to propose a distinction between information acquired by way of sensory processes and information acquired through language. Recently, studies such as, for example, Yu and Smith (2007) and L. B. Smith and Yu (2008) have shown how statistical regularities between the co-occurrences of words and the co-occurrences of their referents can facilitate the learning of word–object mappings, and others such as, for example, Roy and Pentland (2002) have shown how simultaneous statistics from visual and auditory senses can facilitate the discovery of words and their referents. Similarly, Louwerse (2008) investigated the parallels between information encoded in linguistic structures and that of more sensory-motor, or embodied, information. More relevant still is the work of Howell and Becker (2001); Howell, Becker, and Jankowicz (2001); and Howell, Jankowicz, and Becker (2005), who have shown that the acquisition of a language by a simple recurrent network is improved when words are augmented with sensory-motor representations. From this, it is argued that grounding words by way of sensory-motor representations can provide a form of semantic bootstrapping for the learning of a grammar. Despite the importance of these studies, however, none have specifically addressed the distinct nature of the semantic information provided by experiential versus language-based distributional data or how both information types could be integrated to form semantic representations. This, we believe, is the primary contribution of the work we describe in this article.

The Consequences of Combining Data Types

Any theory of semantic representation that is based on combining experiential and distributional data will clearly redress some of the fundamental limitations inherent in theories based on using only one source alone. However, the importance of combining data is not simply that two major sources of data are being used. Rather, due to the fact that these two sources are correlated with one another, learning from both sources jointly permits more knowledge to be acquired from the available data and allows knowledge acquired from language to be related to knowledge about the world. This has important consequences for any theory of how knowledge—semantic knowledge in particular but any knowledge more generally—is acquired and used.

By learning semantic representations from the joint distribution of experiential and distributional data, more semantic knowledge is gained from the available data than is possible using one source exclusively or using both independently. This is a consequence of the elementary statistical fact that all the information in a joint probability distribution cannot be known by reference to its marginal distributions. We can see this in Figure 1, where the joint distribution $P(x, y)$ varies across the subfigures, but both marginal distributions, $P(x)$ and $P(y)$, remain unchanged. By a direct analogy, all the information from which semantic knowledge can be attained is given by the joint distribution over both experiential and distributional data. Obviously, considering only one of these sources will lead to information loss. Equally important, though perhaps less intuitively obvious, is that treating the two data sets

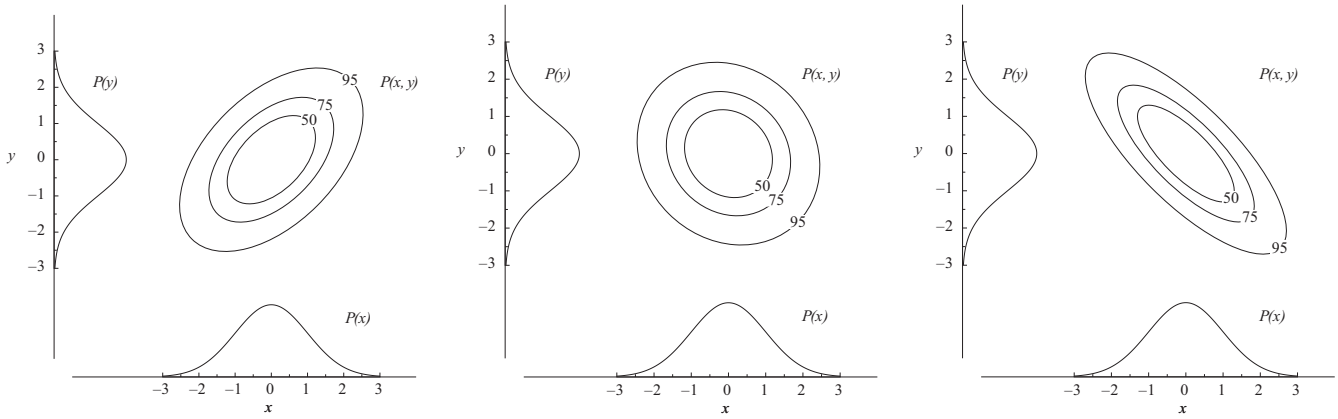

Figure 1. An example illustrating how the information in any joint distribution cannot, in general, be known from the marginal distributions. In this example, the joint distribution P(x, y) varies considerably across the subfigures, but the marginals P(x) and P(y) remain unchanged.

independently will also lead to a loss of information. Only by treating the data as a single joint data set can all the available information be utilized. The structure of the data will be obscured whenever they are treated independently or in the mutually exclusive manner.

The importance of this fact is that when experiential and distributional data are treated jointly, new structures in the data will become apparent. These new structures are specifically those that bridge between the structures in the individual data sets. For example, returning to the example described earlier, a child may learn on the basis of experiential data that *dog* and *cat* are related, as they share salient properties with one another. On the other hand, he or she may also learn on the basis of distributional data that *dog* and *pet* or *animal* are related, as they are distributed similarly across language. By combining the knowledge gleaned from both sources, the child may infer that *cat* is also related to *pet* or *animal.* By so doing, the child is able to generalize to a greater extent than would be possible by using either data set individually or by averaging between them.

The fact that combining experiential and distributional data allows more knowledge to be acquired from the available data and generalization to be improved furthers the ability to explain how semantic knowledge, or any knowledge at all, can be acquired from the limited and finite amounts of data available to human beings. This problem has often been referred to as *Plato's problem* (e.g., Chomsky, 1986). Distributional models, in particular the LSA model of Landauer and Dumais (1997), have been explicitly proffered as solutions to this foundational problem. From our theoretical perspective, distributional accounts offer a partial solution to this problem. A more complete solution should take into account not only that there are two major sources of statistical data available to human beings but also that these two sources are intercorrelated. The statistical patterns necessary for successful inference and generalization exist both within each source and also between them. By using all of these patterns, more knowledge can be acquired than is possible otherwise.

Outline of the Article

The primary hypothesis put forward in this article is that semantic representations are the product of the what we call the statistical combination of experiential and distributional data. By this, we mean that semantic representations are learned from the statistical structure underlying the joint distribution of experiential and distributional data. This contrasts with alternative possibilities such as learning from either one data source alone or by using both data types independently and averaging between them. In this article, we address this hypothesis by providing probabilistic models that learn semantic representations either from experiential data alone, from distributional data alone, or from both in combination. We then compare the semantic representations that are learned by each of these models, both with one another and with a set of measures of human semantic representations. With the inclusion of some necessary control conditions, these comparisons allow us to assess the roles of both experiential and distributional data, and their integration, in learning semantic representations.

The article proceeds as follows: In the section entitled Models, we introduce and motivate the general modeling framework that we employ throughout the remainder of the work and provide a description of the data sets we use for training. The section entitled Model Analysis and Evaluation provides a detailed qualitative and quantitative analysis of the semantic representations learned by the models and how they compare to human-based measures of semantic similarity. In the General Discussion, we describe the general significance of these findings, particularly how semantic representations learned by combining experiential and distributional data can lead to improved semantic knowledge acquisition and knowledge-based inferences. Throughout the article, we attempt to keep technical details to a minimum. In the Appendixes, however, we supply sufficient detail to describe the training, inference, and analysis of our models.

Models

The problem of learning semantic representations from statistical data—whether experiential, distributional, or both combined—is an example of a general problem of statistical learning. Its solution involves describing the statistical structure underlying the data in terms of a set of basic or elementary statistical patterns and then representing the semantics of words in terms of these patterns. For example, experiential data describe words as distri-

butions over sensory-motor features. Distributional data describe words as distributions over texts. When using both experiential and distributional data concurrently, words are joint distributions over sensory-motor features and texts. In each case, learning semantic representations involves discovery of a repertoire of elementary statistical patterns underlying these data sets and then representing words as distributions over these patterns. While, in each case, the data are distinct, the learning problem itself is a general one and affords one general solution.

To understand the models that we use, it is helpful to consider a simple example where the data are sets of independent arrays each describing a frequency distribution over a finite number of discrete elements. A graphical illustration of such data is provided in Figure 2. On the left of this figure are shown $J = 15$ rows of data vectors, each one being a frequency distribution over $V = 9$ elements. The shade of grayscale represents relative frequency, with lighter shades indicating higher frequency. In other words, each vector represents the frequency of occurrence of each of a finite number of elements. From visual inspection, certain patterns revealing the nonrandom structure in these data begin to become apparent. In a sense, a statistical model is a means to formally execute this pattern-discovery process that the human visual system does naturally. In this respect, one particularly general modeling approach is to assume that each data vector is generated by sampling from a composition of elementary probability distributions. The learning problem then reduces to using statistical inference to infer what these elementary distributions are likely to be. In the case of Figure 2, each frequency distribution shown on the left is in fact a sample drawn from compositions of $K = 8$ elementary probability distributions. These distributions are depicted on the right, where each row is a probability vector and the lighter the shade of grayscale, the higher the probability that it denotes.

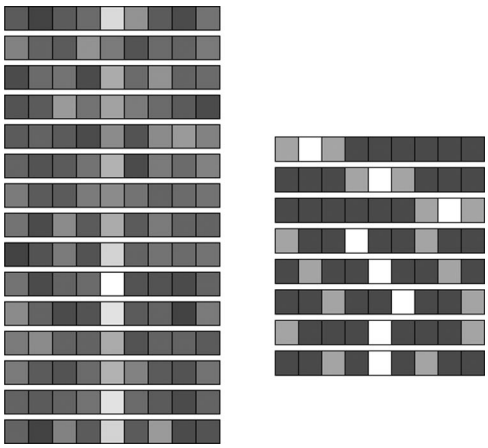

Figure 2. An illustrative example of the statistical structure underlying a set of data. The arrays shown on the left are observed data, each being a frequency distribution over $V = 9$ elements. The shade of grayscale indicates relative frequency, with lighter shades indicating higher frequency. Each of these data arrays was generated by sampling from a specific weighted composition of $K = 8$ elementary probability distributions. These are shown on the right, that is, each row on the right is a probability vector where the shade of grayscale indicates the probability it assigns to each of the $V = 9$ elements; the lighter the shade of the grayscale, the higher the probability it denotes. The elementary probability distributions, as well as the weighted composition specific to each data array (not shown), describe all the statistical structure inherent in the observed data. The general task of unsupervised statistical learning is to infer these directly from the data.

The problem of learning semantic representations from either experiential data alone or distributional data alone can be readily understood in relation to the illustration in Figure 2. In the case of experiential data, for example, we can view each row of the left subfigure as denoting an object or event in the world that is the referent of some word. Each column denotes a particular sensory-motor feature. As such, the data vector signifies the observed frequency of each feature for each object or event. The statistical structure of this data set is described by compositions of the elementary distributions depicted on the right. Each of these distributions, by placing most of its probability mass upon a subset of the features, denotes a *feature cluster* that specifies a set of intercorrelated features found in the data. For example, and as we show below, these clusters might represent human body parts, typical properties of fruit and vegetables, typical properties of household furniture, and so on. Each object or event in the world, or the word that refers to it, can then be represented in terms of how much each one of these feature clusters typifies it. This we can take to be the word's semantic representation as derived from experiential data.

For the case of learning from distributional data, we can view the rows on the left of Figure 2 as denoting an individual linguistic context, for example, a text, while the columns represent word types. As such, each data vector specifies the observed frequency of any given word in any given text. The statistical structure of the texts is described by the compositions of elementary distributions shown on the right. In this example, each of these distributions now specifies a probability distribution over word types. By analogy with the feature clusters in the previous example, in this example, these probability distributions can be seen as corresponding to a cluster of correlated words that have the character of discourse topics, that is, interrelated words that denote a specific topic such as sport, finance, politics, and so on. Each text can then be represented in terms of the extent to which each discourse topic typifies it. By the symmetry of the problem (i.e., just as texts are distributions over words, words can be seen as distributions over texts), however, we can also represent each individual word in terms of how typical it is for each discourse topic. We can then take this to be the word's semantic representation as derived from distributional data.

In the case of learning semantic representations from both experiential and distributional data in combination, the illustrative example to which we have been referring must be extended to accommodate the case of joint frequency distributions. This is shown in Figure 3. In the left subfigure are displayed $J = 6$ two-dimensional arrays, each comprising $F = 5 \times V = 9$ elements. We can view each two-dimensional array as denoting a particular linguistic context. In this example, it is intuitive to think of each linguistic context as being, for example, some spoken discourse that is referring to objects and events in its immediate environment. Each row of the arrays is a word type, and each column is a sensory-motor feature. As such, we can see that each linguistic context is a joint frequency distribution over words and features. Each element in this array indicates the frequency of

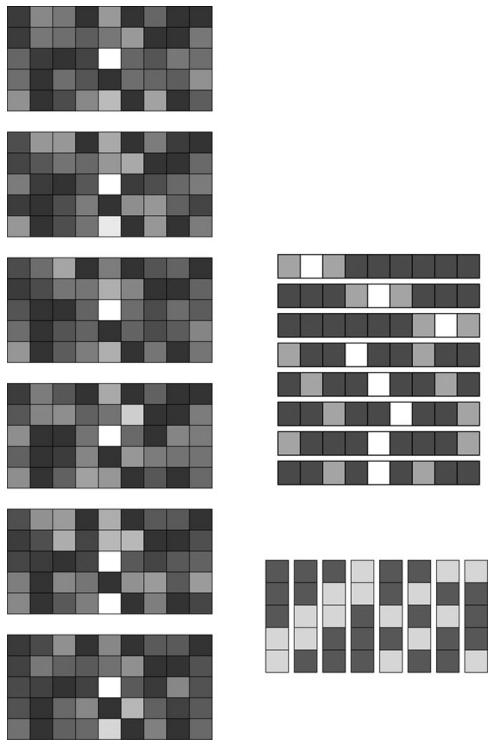

Figure 3. An illustrative example of the statistical structure underlying a combined data set. Each two-dimensional array on the left is a frequency distribution over $F = 5 \times V = 9$ elements. Each of these arrays are generated by sampling from a specific weighted composition of a set of $K = 8$ elementary two-dimensional probability distributions. In this example, these $K = 8$ distributions are formed by the product of a pair of coupled distributions, one over $V = 9$ elements, the other over $F = 5$ elements. These are shown on the left, where each of the $K = 8$ rows in the upper part of the subfigure is coupled to one of the $K = 8$ rows below. As with the previous example, these coupled distributions and the weighted composition specific to each array describe all the statistical structure inherent in the observed arrays. The task of learning in a combined model is thus to infer these from the available data.

observing a particular word and particular feature jointly within the same linguistic context. For example, if the spoken discourse is about fruit, we could expect words like *apples, grapes, lemon,* and so on to be observed simultaneously with sensory-motor features like *juice, seed, sweet,* and so on.

Just as in the previous examples, in this example, each data array is generated by sampling from a specific weighted composition of $K = 8$ elementary probability distributions. These probability distributions are two-dimensional, and for this example, each is formed by the product of two coupled distributions, one over $V = 9$ elements and the other over $F = 5$ elements. The $K = 8$ coupled distributions are shown in the right subfigure, where each of the $K = 8$ rows on top are to be seen as coupled to one of the $K = 8$ columns on the bottom. In this example, we can view these as couplings of discourse topics and feature clusters, where the feature clusters are those features typical of the words signified by the discourse topic. For example, a discourse topic about food and drink might be coupled with a feature cluster about the sensory-motor features of eating and drinking, or a discourse topic about education might be coupled with a feature cluster denoting sensory-motor features relates to teaching and learning. As in the cases of the previous two learning scenarios, we can represent words in terms of these coupled distributions. These representations simultaneously express how typical any given word is of the discourse topic and how well the coupled feature cluster typifies its sensory-motor features. This can then be viewed as the word's semantic representation derived from the combination of experiential and distributional data.

The foregoing descriptions can be put more formally. For the cases of learning from experiential data alone or from distributional data alone, we can view the jth data vector (e.g., the jth row of the left subfigure of Figure 2) as having been sampled from a probability distribution

$$P(y|j) = \sum_{k=1}^{K} P(y|x = k)P(x = k|j), \quad (1)$$

where $P(y|x = k)$ is the kth elementary distribution (e.g., the kth row depicted in the right subfigure of Figure 2) and $P(x = k|j)$ describes the weight assigned to this elementary distribution in the case of the jth data vector. On the other hand, for the case of learning from both experiential and distributional data in combination, we can view the jth data array (e.g., the jth array of the left subfigure of Figure 3) as sampled from the joint distribution

$$P(y, z|j) = \sum_{k=1}^{K} P(y|x = k)P(z|x = k)P(x = k|j), \quad (2)$$

where the couple $P(y|x = k) \times P(z|x = k)$ is the kth elementary distribution couple (e.g., the first row of the upper array coupled to the first column of the lower array depicted to the right of Figure 3) and $P(x = k|j)$ gives the weight assigned to this couple in the jth data array.

In the language of probabilistic models, this description states that each data vector or array is drawn from a probabilistic mixture model. While they share their component distributions, each vector or array is derived from a unique mixture of these components. By assuming that each of these mixing proportions is itself drawn from a common distribution over mixing proportions, the entire data set, in any of the three learning scenarios, is modeled as a hierarchical mixture model based on the LDA model. As mentioned, this type of probabilistic model was introduced by Blei et al. (2003) and applied to the particular case of learning semantic representations by Griffiths and Steyvers (2002, 2003) and Griffiths et al. (2007), where it was referred to as the *topics model.* The LDA models have two set of parameters: those that describe the elementary distributions and those that describe the common distribution over mixing proportions. These parameters are obviously unknown and must be inferred from the data. This process is referred to as model fitting (or learning, or training). The technical details of the models we use, as well as their method of model fitting, are described in Appendix A.

Training Data

For the purpose of model training, we required two separate data sets, one providing experiential data and the other distributional-

based data. Following common practice (see examples in the introduction), we operationalized the experiential data as speaker-generated feature norms and the distributional data as texts from a large natural-language corpus. These data sets were used separately to train the experiential-only and distributional-only models and were used concurrently to train the combined model. In what follows, we describe these data sets and how they were used as model training data.

In the general procedure of collecting speaker-generated feature norms, human volunteers provide features that they believe belong to the referents of words in a list of words. After suitable preprocessing, this results in a set of unique features, with each word being described by whether and how often any given feature is listed as one its properties. The norms used in this study were collected previously for the purpose of another study and are described in detail in Vinson and Vigliocco (2002) and Vigliocco et al. (2004). In summary, for each one of 456 words (240 nouns and 216 verbs), undergraduate student volunteers provided lists of their features. Each word was concrete in the sense that it referred to objects, events, or actions that have tangible physical referents, for example, words like *bird, chair, run, say,* and so on. The set of words was divided into subsets, and separate groups of 20 participants were assigned to each group. After excluding attributes that were produced less than six times overall across the entire set of 456 words, a set of 1,029 unique features was obtained, such that each word could be described in terms of the relative frequency of occurrence of each of these unique features. For the purposes of analysis, these 1,029 features could be classified as belonging to one of five main types: visual (those attributes dependent upon the visual modality, e.g., *red*), other perceptual (features involving sensory modalities other than vision, e.g., *sweet, loud*), functional (referring to the purpose for which an item is used or the goal of an action, e.g., *cuts*), motoric (referring to the way in which an object is used or describing its motion, e.g., *handheld*), and other (all other features not meeting the above descriptions). Examples of the high-probability features for certain words are provided in Table 1.

While speaker-generated attribute norms are widely used as a proxy for experiential data, there are obvious limitations to data obtained in this manner. Experiential data, as we have defined them, are the totality of data describing human perception and interaction with the physical world. This includes not only the data derived by way of the main senses but also motor-based affordances and affective data. Naturally, this is vast array of data that will be only very crudely approximated by an array of the relative frequencies of 1,029 attributes. More problematic still is the fact the features produced by human subjects are sometimes of a dubious nature, and it is not always clear that they identify data that are derived from perception and interaction with the physical world. It may be argued that, in some cases, the features listed represent so-called encyclopedic knowledge that is derived by way of a formal or informal education. Notwithstanding its inherent limitations, efforts were made to ensure that the features obtained from this method referred, in a broad sense, to sensory, motor, or functional characteristics. Participants were instructed to list only tangible properties of the objects or actions in question and not to engage in free association or to provide dictionary or encyclopedic definitions. Given that the majority of features provided could be classified as either perceptual, motor-related, or functional indicates that, in many cases, the features are of a satisfactory nature.

Given the current state of information technology, resources to obtain almost arbitrarily large quantities of distributional data are readily available. For our purposes, a large standardized text corpus like the British National Corpus (BNC) is sufficient. The BNC is a 100-million word corpus of contemporary written and spoken English in the British Isles. It is standardized in the sense that it is restricted to late 20th century British English, rather than being a mélange of American and British English or of English from different centuries. It is general in the sense that it comprises texts from a wide cross-section of subjects and materials and deliberately subsamples from single-theme or single-author works to minimize idiosyncratic texts. Although the BNC comprises both spoken and written language, the written component constitutes the majority of the corpus, namely, 90%, and consequently, we restricted our attention to this component. According to its publishers, the texts that make up the written component include "extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays" (Burnard, 2009, paragraph 2).

The BNC is annotated both with syntactic information, for example, part-of-speech tags, and also with the structural properties of a text, such as sectioning and subsectioning information. The latter type of annotation facilitates the processing of the corpus. To extract texts from the BNC, we extracted contiguous blocks that were labeled as high-level sections, roughly corresponding to individual articles, letters, or chapters. These sections varied in size from tens to thousands of words, and from these, we

Table 1
Seven Randomly Chosen Concrete Words With Examples of Some of Their High-Probability Features According to the Data Set Collected by Vigliocco, Vinson, Lewis, and Garrett (2004)

Bear	Book	Grape	Pig	Pull	Punch	Strawberry
big	read	sweet	pink	move	hurt	red
fur	story	green	farm	force	fist	taste
teeth	word	red	smell	object	hit	sweet
brown	information	purple	dirt	bring	anger	seed
black	cover	round	mud	action	hand	grow
wood	paper	vine	4-legs	something	action	small
claw	humans	juice	fat	tug	force	garden
aggressive	write	small	pet	close	humans	food
scare	enjoy	wine	tail	grasp	intentional	eat

chose only those texts that were approximately 150–250 words in length. This length is typical of, for example, a short newspaper article. We also restricted our attention to those texts where at least 10% of their words were words for which we also had collected feature data. Following these criteria, we sampled 7,776 individual texts to be used for training. Of all the word types that occurred within this subset of texts, we excluded both stop-words and those words that occurred less than five times overall. The stop-words, for example, *the, of, and, is, to, you,* and so on, were taken from a standard database list of stop-words and were all high-frequency words whose occurrence across all texts was essentially uniform. This restriction resulted in a total of 7,586 unique words. Although this corpus is the primary one used in our analysis, as is elaborated below, we found it useful to compare results obtained from this corpus with those of a larger subset of the BNC comprising these same texts plus an additional 36,282 texts (each ranging from 150 to 300 words in length), with a total of 20,672 unique word types.

The feature norms and corpus of 7,776 texts comprise our two principal data sets, representing experiential and distributional data, respectively. The feature norms clearly are in the form of frequency data vectors. Likewise, each text can be described simply in terms of whether and how often each word, from the vocabulary of 7,586 words, occurs within it. This characterization of text data is usually referred to as the *bag-of-words assumption,* whereby the sequential order of words within the text is taken to be irrelevant. The combined data set is, as its name implies, a combined form of the above data sets: Each text is a frequency distribution over words, while the words are frequency distributions over features. As only a subset of word types have features associated with them, the feature distributions for all other words are left undefined. The combined training set was prepared by using the text corpus in its original form and, every time a word for which features were defined was encountered, sampling a feature from its feature distribution. This resulted in a parallel corpus of text and features. From this parallel training corpus, the constituent features-only and text-only training corpora were then easily derived. The features-only training set was the set of words for which features existed, along with the features to which they were paired in the combined training sets. This ensured that the relative frequency of occurrence of any of these words was identical in the features-only and combined training sets. The text-only training set was simply the combined training set with the features stripped away.

Justification of Our Modeling Approach

Our general perspective on the use of models in cognitive science is pragmatic: Models are tools that allow us to describe and understand phenomena. In this sense, one does not need a strong sense of a model being right or wrong but rather of it simply being more or less useful. There are unlimited ways to model a given phenomenon, and each should be judged in terms of the understanding or insight it affords.

The nature of the problems we consider—the data being sets of independent arrays describing frequency distributions over discrete elements and the learning objective being the discovery of the elementary patterns underlying these data—obviously constrain the types of models that are appropriate. The only popular choice of models for such problems is LSA (e.g., Landauer & Dumais, 1997). By contrast, as we have just described, the models we use are based on the LDA model. However, LDA and LSA need not be viewed in opposition but can be seen rather as members of a family or continuum of models. LDA can be seen as a probabilistic version of LSA, with there being a direct lineage from the original LSA model to the probabilistic LSA model of Hofmann (1999a, 1999b) to the LDA model of Blei et al. (2003) to the instantiation and generalizations discussed in Griffiths et al. (2007).

Some of the primary advantages of the models we use and of using probabilistic generative models generally relate to learning and analysis. The Bayesian approach to learning is to infer a probability distribution over a model's parameters that is conditioned upon observed data. This so-called posterior distribution measures how probable it is that any possible parameter setting could have been the cause of the observed data. This approach is in contrast to the so-called point estimate approaches to learning, whereby a single parameter value, such as the maximum-likelihood value, is chosen. The advantage of using a distribution over a point estimate of the parameters is that all the evidence about the nature of the model's parameters is utilized. If the likelihood function is multimodal or if the modal points and centers of greatest mass are not coincident, choosing a single point estimate of the parameters will result in a limited or spurious summary of the evidence provided by the data. This is a particularly common problem with complex models, due to their inherently vast parameters spaces.

In models of human cognition, the internal representations in the model, for example, the hidden units of a neural network and their connections to other units, are often said to correspond to the knowledge inherent in the model and are of particular importance to explain the behavior of the model. It is not uncommon, however, for the interpretation of these internal representations to be challenging, and consequently, the behavior of the model is often inscrutable. In the models that we present here, all variables and relationships between variables are represented in explicitly probabilistic terms. This facilitates our understanding of the significance of each variable in the system and provides us with the means to precisely describe the inferences or predictions that can be made by the model. Later, we make use of these features to explicitly define the nature of the semantic knowledge in each model and to explain exactly how this knowledge can be used to draw inferences or predictions about semantic relationships.

More generally, the probabilistic generative model approach is useful as it offers a flexible and systematic framework for designing and analyzing models. In particular, this means that generalization of, for example, the topics model of Griffiths et al. (2007) to one that can learn semantic representations from both texts and feature vectors in combination is a straightforward process. This also facilitates intermodel comparison. For example, the models that we use here can all be seen as either identical or otherwise straightforward generalizations of one another, with the primary difference between them being the nature of the data that they use for learning. This gives us a unified framework to explore the learning of semantic representations from data, regardless of source.

Model Analysis and Evaluation

The objective of the statistical models we use is to discover latent statistical patterns in the training data sets. For the case of the model trained with experiential data alone, each of these patterns will correspond to feature clusters, or clusters of corre-

lated features. For the model trained on distributional data alone, the statistical patterns will resemble discourse topics, or clusters of correlated words that together might identify a coherent topic or subject matter in a text. Models trained using both experiential and distributional data concurrently will discover statistical patterns that characterize the correlations both within and between their two constituent data sets. In other words, these models will effectively align feature clusters with discourse topics to form a coupled pair of patterns such that the features in the feature cluster are correlated with the words in its corresponding discourse topic.

The latent patterns in each model represent its semantic knowledge. Any given word can be described in terms of these patterns, and we can view these descriptions as the word's semantic representation. By comparing the distributions of different words, we can measure the extent to which their semantic representations are similar or different. In what follows, we provide examples of the statistical patterns inferred by each model and then provide the interword semantic similarities derived from these patterns.

Latent Distributions

In the experiential model, each latent variable corresponds to a probability distribution over a set of sensory-motor features. In Table 2, we display the 20 features with the highest probability in a randomly chosen set of seven distributions (chosen from a total set of 120). As can be seen, each of these distributions forms a cluster of correlated features that signifies a basic statistical pattern in the training data. In each example shown, we have applied an intuitive label to approximately describe what it signifies.

These feature clusters form a repertoire of elementary statistical patterns from which the objects and events in our feature-based training set are composed. More precisely, each object or event (e.g., the referent of *apple, kill, table,* etc.) can be viewed as a frequency distribution over sensory-motor features that is sampled from a specific weighted composition of these elementary feature clusters. The weight assigned to each feature cluster specifies the extent to which the item is composed of that feature cluster or how typical that feature cluster is of it. For example, we might expect that the feature clusters we have labeled *fruit* or *cooking* would be typical of apples, while *fixing* or *locomotion* would be much less so. In this case, the word *apple* would correspond to a composition of these clusters that weights the former pair more highly than the latter. As we take these feature clusters to be the model's semantic knowledge, the specific weight assigned to them by any given word corresponds to that word's semantic representation.

In the distributional model, each latent variable specifies a probability distribution over the vocabulary of words (i.e., all word types in the corpus). We display the 20 most probable words from seven (from a total of 250) randomly chosen distributions in Table 3. From this table, we see that these distributions signify discourse topics,[1] and we have applied intuitive labels to approximately describe them. By a precise analogy with the case of the experiential model, each text in the training corpus is a frequency distribution over words that is derived from a specific composition of these discourse topics. This composition indicates the extent to which each text is characterized by the discourse topics. For example, a text about finance is likely to heavily weight the topics *economics* and *markets,* but to weight *soccer* less. By an application of Bayes's theorem, we can also represent each word as a distribution over the set of topics. For example, the word *travel* may be represented highly by the topics labeled *trains* or *aircraft,* but less so by, say, *prison.* As with the case of the experiential model, any given word's distribution over these topics is its semantic representation.

The combined model couples feature clusters with discourse topics. The feature clusters, as before, identify sets of intercorrelated features in the world, while their corresponding discourse topics identify sets of intercorrelated words within texts. However, just as the feature clusters and discourse topics are correlated within themselves, they are also correlated with one another: The features in the feature cluster are the characteristic properties of those words that most characterize the discourse topic to which it is coupled. In Table 4, we provide six randomly chosen (out of a total of 350) coupled distributions. For each couple, we have displayed the 20 most likely features of the feature cluster and the 20 most likely words of the discourse topic. As can be seen, each coupling has a coherent meaning, identifying features and words that share meaning with themselves and with one another. For example, in the case of the couple we have labeled *food,* we see sensory-motor features describing eating, drinking, and other actions related to food. This feature cluster is coupled to a discourse topic with words describing food and diet. In the case of the couple labeled *cars,* we see properties of cars and the act of driving coupled to a topic with the words *car, road, drive,* and so on. As before, we can view the coupled distributions in the combined model as its semantic knowledge, and the distribution of each word over these couples can be viewed as its semantic representation.

Interword Similarity and Neighborhood Structure

In each model, the semantic representation of a word is a probability distribution over the model's latent variables. As such, each word can be represented as a point on a probability simplex.[2] The similarity between the semantic representations of any pair of words can then be measured by the distance between points in this space. An appropriate metric to use for the case of a probability simplex space is the symmetrized Kullback-Leibler divergence (see Appendix A for description).

Using this method, for each model, we can calculate the nearest neighbors of any given word. For illustrative purposes, rather than simply presenting sets of near neighbors, it is more informative to present what we term *neighborhood cliques.* The neighborhood clique of word w_i is formed by finding its K-nearest neighbors. For each pair w_j and w_k in this set of K neighbors, we assess whether w_j and w_k are also neighbors of one another.[3] These relationships can be represented as undirected graphs, that is, w_i and its neigh-

[1] Griffiths et al. (2007) used the term *topics* rather than discourse topics to describe latent distributions of this type. We prefer to use the latter term so as to clarify that we are referring specifically to statistical patterns in texts.

[2] A probability K simplex is the subset of $\mathbb{R}^{K+1}$ defined by the set $\{x = (x_0, x_1 \ldots x_k \ldots x_K)' \in \mathbb{R}^{K+1} : \sum_{k=0}^{K} x_k = 1 \text{ and for } 0 \leq k \leq K, x_k \geq 0\}$. In other words, a probability K simplex is the subset of $(K + 1)$-dimensional space where all points define a discrete probability distribution.

[3] In other words, if w_j is in the set of K-nearest neighbors of w_k and if w_k is in the set of K-nearest neighbors of w_j, then we say that w_j and w_k are themselves neighbors.

Table 2
Examples of Randomly Chosen Latent Distributions in the Experiential Model

Fruit	Animals	Speaking	Cars	Cooking	Fixing	Locomotion
juice	fur	speak	wheel	mix	construct	leg
yellow	4-legs	word	transport	rotate	build	fast
red	tail	voice	passenger	spoon	new	exercise
round	pet	talk	gas	turn	fix	feet
grow	big	mouth	automobile	utensil	work	slow
sweet	small	language	drive	dance	create	body
sour	bark	sound	metal	hand	building	intentional
green	black	express	seat	bowl	material	go
taste	hair	converse	door	repeat	house	upright
seed	farm	noise	window	join	break	walk
small	wild	secret	sport	combine	form	sport
peel	domestic	explain	fast	stretch	physical	arm
citrus	ear	say	engine	awkward	action	race
good	white	comfort	move	cook	hole	foot
skin	ride	verbal	destination	bake	hand	speed
eat	zoo	understand	large	stir	finish	shoe
orange	wool	friend	oil	kitchen	carpenter	sweat
pit	friend	gossip	expensive	container	wood	destination
soft	meow	command	plastic	liquid	heavy	long
flesh	large	share	low	repeat	machine	work

Note. In this model, each latent distribution is a probability distribution over a set of 1,029 features. In each of the examples displayed, we provide its most highly probable features. As is evident, each distribution identifies a coherent cluster of features. For each example, we provide a descriptive label.

bors are the vertices of a graph, with an edge between w_i and each of its neighbors and an edge between w_j and w_k if w_j and w_k are themselves semantic neighbors.

Neighborhood-clique graphs provide a useful visualization tool to understand the differences in the semantic representations of words across different models. For example, we can use these graphs to gain an understanding of the possible differences between the semantic information contained in experiential versus distributional data. In Figure 4, we compare the neighborhood cliques of a set of words in the experiential and distributional models. Due to the order-of-magnitude difference in the size of the vocabulary in the experiential versus the distributional model, we

Table 3
Examples of Randomly Chosen Latent Distributions in the Distributional Model

Soccer	Prison	Economics	Drink	Markets	Trains	Aircraft
league	prison	rate	pub	market	railway	air
cup	years	cent	guinness	stock	train	aircraft
season	sentence	inflation	beer	exchange	station	flying
team	jail	recession	drink	demand	steam	flight
game	home	recovery	bar	share	rail	plane
match	prisoner	economy	drinking	group	locomotive	airport
division	serving	cut	alcohol	news	class	pilot
win	office	fall	bottle	trading	run	fly
club	life	economic	whisky	following	engine	jet
games	appeal	year	spirits	index	track	crash
final	case	rise	brewery	yesterday	lines	near
play	justice	confidence	wine	close	running	aviation
home	escape	industry	pint	fall	valley	base
won	cell	billion	brewing	early	passenger	airline
football	given	yesterday	drunk	added	service	flew
coach	guilty	growth	real	stocks	built	ground
second	punishment	spending	ale	value	platform	crew
victory	judge	high	duty	better	freight	squadron
saturday	term	increase	lager	fell	great	helicopter
round	crime	sales	cider	london	british	force

Note. In this model, each distribution is a probability distribution over a set of 7,586 word types and can be viewed as being equivalent to a discourse topic. For each example distribution displayed, we provide its 20 most probable words. In each case, we provide intuitive labels for the discourse topics that the distributions represent.

Table 4
Examples of Randomly Drawn Latent Distributions From the Combined Model

Food	Education	Cars	Body	Cooking	Loans
		Highly probably features			
mouth	teach	drive	body	food	need
liquid	learn	wheel	hand	cook	give
consume	instruct	engine	joint	kitchen	money
food	guide	gas	move	pot	purchase
swallow	school	move	arm	heat	own
ingest	talk	passenger	humans	hot	trade
enjoy	idea	steer	connect	eat	return
hunger	show	window	muscle	stir	borrow
taste	help	fast	bone	oven	goods
thirst	know	adult	bend	cut	swap
stomach	express	transport	point	stove	bank
action	experience	seat	fingernail	knife	buy
nutrition	task	metal	part	mix	store
water	effort	door	limb	bake	service
		Highly probable words			
food	course	car	arms	add	bank
eat	students	road	arm	cook	exchange
drink	english	drive	fingers	oil	loan
eating	language	driving	side	minutes	loans
wine	education	cars	hands	chopped	lend
drinking	college	driver	shoulder	heat	mortgage
drinks	university	drove	body	serve	borrow
alcohol	teaching	van	shoulders	large	terms
ate	student	vehicle	knee	butter	exchanged
meal	taught	front	wrist	salt	banks
lunch	courses	vehicles	elbow	mix	interest
weight	teach	engine	leg	pan	deal
diet	study	speed	finger	stir	borrowed
sugar	higher	drivers	chest	tbsp	purchase
wine	learn	motor	slowly	sauce	finance

Note. In this model, each distribution is a coupling of a distribution over features and a distribution over words. The upper half of the table shows highly probable features, while the lower half of the table shows highly probable words. As can be seen, in each example, both the features and the words are intercorrelated with themselves and with one another. The latent distributions can be seen as alignments of a feature cluster and a discourse topic that are both consistent with a single general meaning.

use graphs of different sizes to represent their semantic neighborhoods. Specifically, for the experiential model, the degree of the central vertex (i.e., the neighborhood size of the word being illustrated) is 10. For the other two models, we chose the degree of central vertices to be 20.

From these graphs, clear and systematic differences in the semantic representations in these two models are evident. The experiential model appears to emphasize more grounded, sensorymotor senses of words, while the distributional model emphasizes more abstract, encyclopedic senses. For example, for *demand,* we see that its sense according to the experiential model is the physical act of making a demand, with its neighbors being words like *plead, request,* and *ask.* By contrast, in the distributional model, *demand* has a distinctly economic sense, with neighbors like *profits, markets,* and *investment.* The same phenomenon is evident for *eat* and *sing,* whose neighborhood cliques are also displayed in Figure 4. According to the experiential model, the senses of both *eat* and *sing* emphasize human actions of singing and eating, along with some of their sensory and motor properties. According to the distributional model, the senses are more encyclopedic. The sense of *eat* relates to diet and nutrition, and the sense of *sing* relates to the music industry.

In Figures 5, 6, and 7, we provide a three-way comparison of the experiential, distributional, and combined models. From these graphs, there is again an evident distinction between the representations in the experiential and distributional models. More importantly, we also see that the combined model integrates the information inherent in these two data types. The resulting representations appear more coherent or refined than that which is derived from either source individually. For example, in the case of *drink,* experiential data alone emphasize the physical act of drinking. Distributional data alone present *drink* particularly in the sense of an alcoholic beverage. On the basis of both data in combination, however, *drink* has a clearer meaning as an act of consumption, related to eating, and related especially to alcoholic beverages. For the case of *kick,* experiential data alone lead to a sense of the term as a quasi-violent physical act. Distributional data alone lead to a distinct sense emphasizing a relationship to sports, particularly rugby and soccer. On the basis of both data types in combination, however, the sense of *kick* is that of the physical act of kicking in a sports-related context, with neighbors referring to the acts of stepping, bouncing, throwing, and so on. A similar pattern is seen in the cases of *kill.* Experiential data emphasize an act of violence. Distributional data emphasize an act related to crime and criminal investigation. In the combined model, *kill* is more clearly the sense of loss of life under violent or criminal circumstances. Again, this sense is both an integration and a refinement of the senses derived from its constituent data types.

By comparing these neighborhood cliques, it is apparent that the senses in the combined model are not simply composites, or sums, of the senses in the experiential and distributional models. This assertion can be tested more formally by evaluating the neighborhood cliques as represented by what we term the *independent model.* The independent model is, in a sense, not a new model but rather a composite of the experiential and distributional models. In both the experiential and distributional models, a word is represented as a probability distribution over latent variables. In the independent model, a word is represented simply as the product of these two probability distributions, that is, the cross-product of the distribution in the experiential model and that in the distributional model.[4] As such, the independent model represents each word as simply the composite of the representations obtained in the individual experiential and distributional models. Furthermore, the distance between any pair of words in the independent model is precisely the sum of their distances in the experiential and the distributional models. The neighborhood cliques formed from such composites are shown in Figures 8, 9, and 10. From these examples, we can see that composite representations are not interesting syntheses of the senses derived from the constituent models. They are simply the result of an average of the constituent senses, emphasizing their overlapping characteristics, if they exist. For

[4] Thus, if $P(x|w)$ is the latent distribution of w in one model and $P(y|w)$ is its distribution in the other, then the distribution in the product space is simply $P(x, y|w) = P(x|w)P(y|w)$.

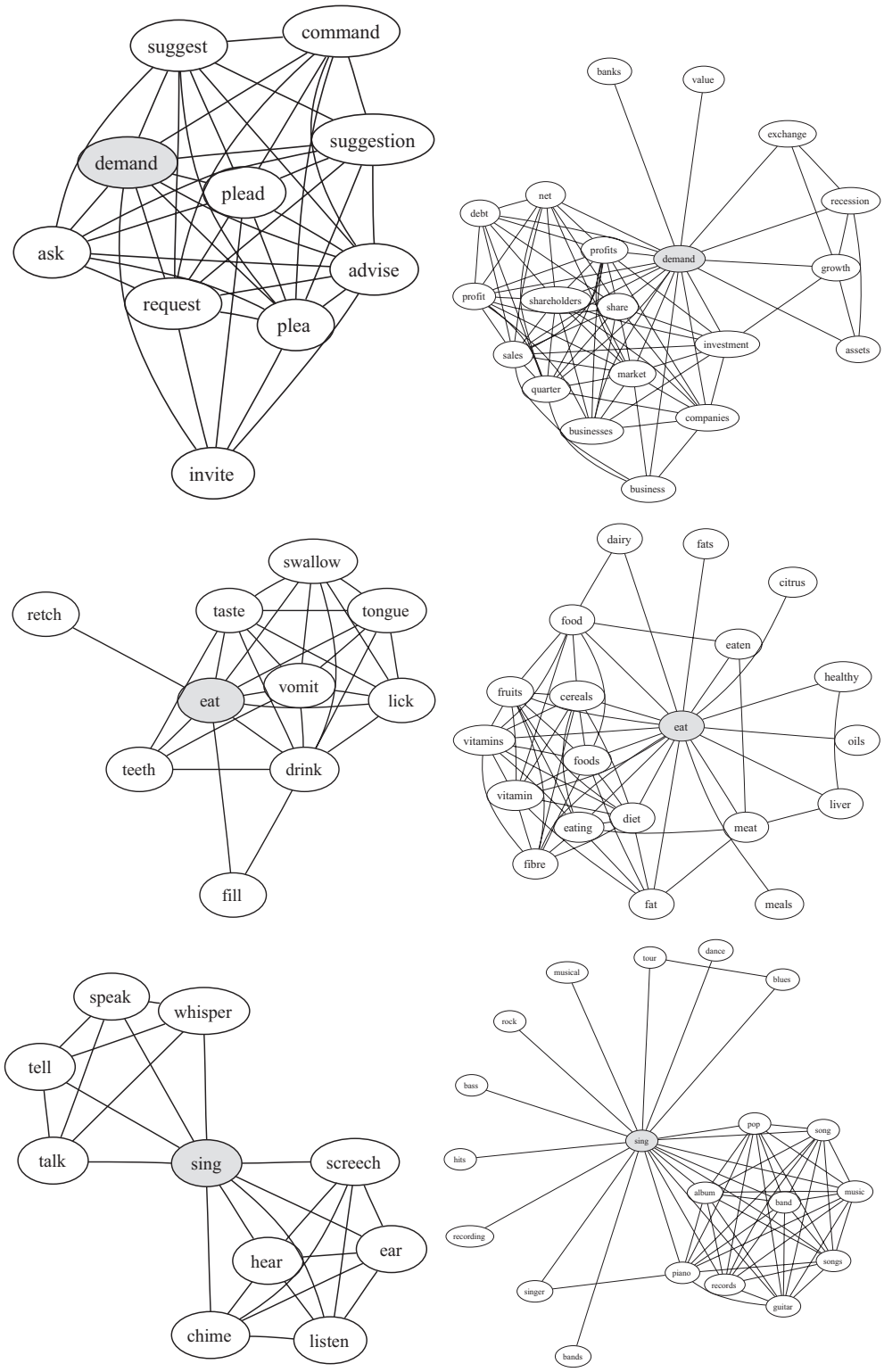

Figure 4. The neighborhood cliques of *demand, eat,* and *sing* in the experiential (shown on the left) and distributional models. These graphs illustrate how semantic representations derived from experiential data are systematically distinct from those derived from distributional data. The experiential model emphasizes more grounded, sensory-motor senses of words, while the distributional model emphasizes more abstract, encyclopedic senses.

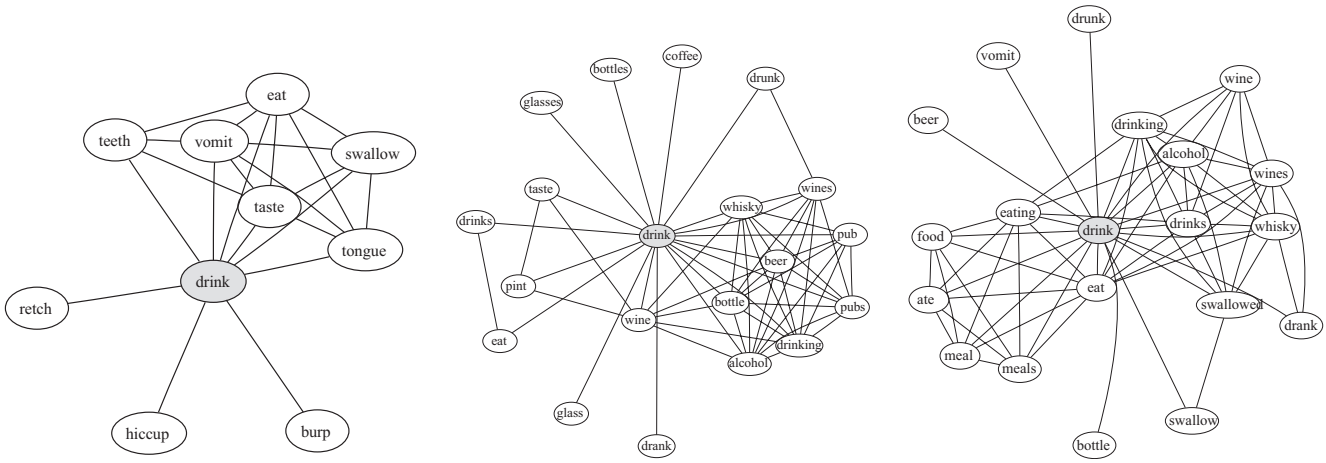

Figure 5. Neighborhood cliques of *drink* according to (from left to right) the experiential, distributional, and combined models.

example, in the case of *drink,* due to the salience of the sense derived from the distributional model and the relative lack of overlap between it and that of the experiential model, the independent model closely resembles the distributional model alone, without any extra emphasis on the sense yielded by the experiential model. Likewise for the case of *kick.* In the case of *kill,* the sense obtained from the independent model largely resembles that of the distributional model, while also emphasizing words that are presumably salient in both it and the experiential model. While these composite senses are not necessarily invalid, reference to the independent model serves to illustrate the importance of what we have termed statistical combination and the extent to which it goes beyond a mere summation of the information derived from two separate sources.

The reason that the representations learned from statistical combinations are richer than those derived from using either one source alone or using both data types independently is due to the fact that when experiential and distributional data are combined, vital statistical structures between the data types will now be revealed. Specifically, these structures are those that bridge between the structures within each individual data set. These bridging structures are readily observable in the combined model. We have provided three case examples in Figure 11. In the upper row, we see how the neighborhood of *explosion* changes from when it is trained on distributional data alone (left subfigure) to when it is trained with both experiential and distributional data in combination (right subfigure). We can see that on the basis of distributional data, *explosion* is related to *bomb.* On the basis of experiential data, *bomb* is related to *destroy* and *bombardment.* Through this chain, *explosion* is now related to *destroy* and *bombardment.* We see the same phenomenon with *fruit.* On the basis of distributional data, *fruit* relates to words like *apple, strawberry,* and so on. On the basis of experiential data, these words are closely related to other fruit terms like *grape, lime,* and *plum.* Through this and similar chains of relationships, the neighborhood of *fruit* changes to include these and other fruits as neighbors. Likewise for the case of *weapon.* On the basis of distributional data, *weapon* is related to *dagger* and *knife.* On the basis of experiential data, *dagger* and

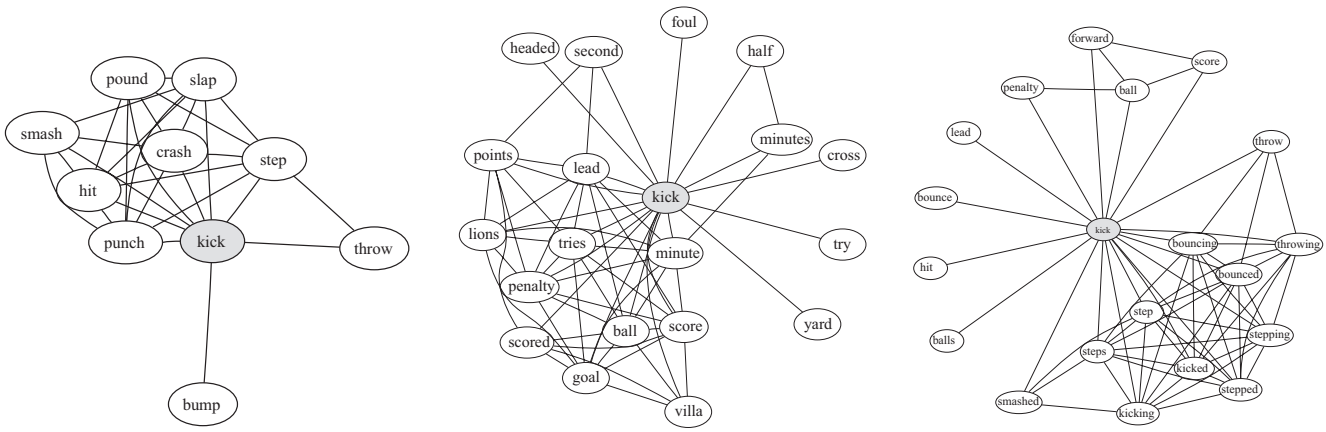

Figure 6. Neighborhood cliques of *kick* according to (from left to right) the experiential, distributional, and combined models.

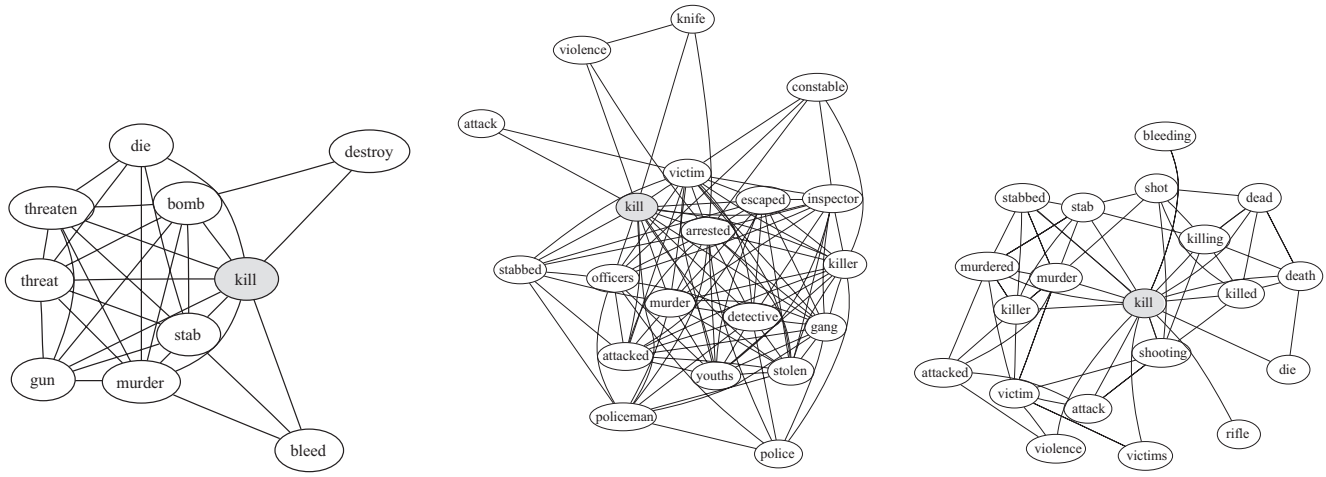

Figure 7. Neighborhood cliques of *kill* according to (from left to right) the experiential, distributional, and combined models.

knife both are related to *kill,* which in turn is related to *gun.* Together, these chains lead to a reshaping of the semantic neighbor of *weapon* in the combined model.

The combined model, as just described, was trained from a joint data set that is obviously quantitatively larger than either of its two component data sets. It is important, therefore, to consider the extent to which the semantic representations obtained by this model are merely the result of a larger training set. To this end, we have trained the distributional model using the original corpus plus an additional, much larger corpus. This augmented training set is approximately 10 times larger than the original set, and in terms of quantity of raw data points, this new training set is clearly far in excess of our combined training set. Neighborhood cliques for *drink, kick,* and *kill* are shown in Figures 12, 13, and 14. As can be seen, despite the far greater quantity of training data, the semantic representations in this new distributional model bear strong resemblances to those in the original model. In the cases of *drink* and *kick,* the neighborhood cliques in the original and augmented models are largely indistinguishable from one another. For the case of *kill,* the sense in the new model keeps the crime-related sense of the original but restricts its scope to refer primarily to gun crimes. From these examples, it is clear that the results evident in the combined model are not simply the result of using a quantitatively larger training set. Using a text-based corpus that is an order of magnitude greater than our original one does not, in itself, lead to remarkably distinct semantic representations.

Quantitative Comparison of Neighborhood Structures

The evidence provided by the neighborhood-clique graphs is corroborated by quantitatively comparing neighborhoods across

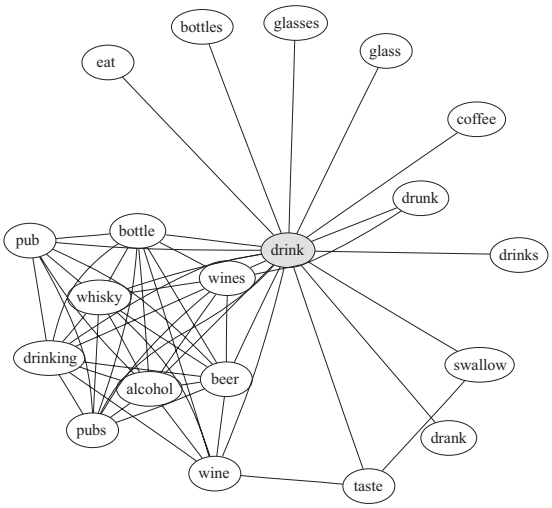

Figure 8. Neighborhood cliques of *drink* according to the independent model.

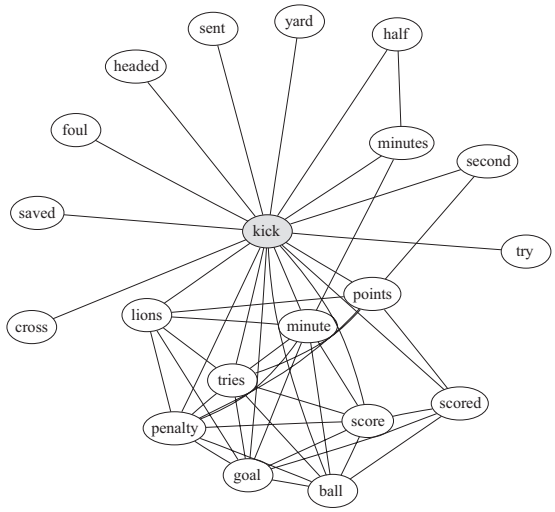

Figure 9. Neighborhood cliques of *kick* according to the independent model.

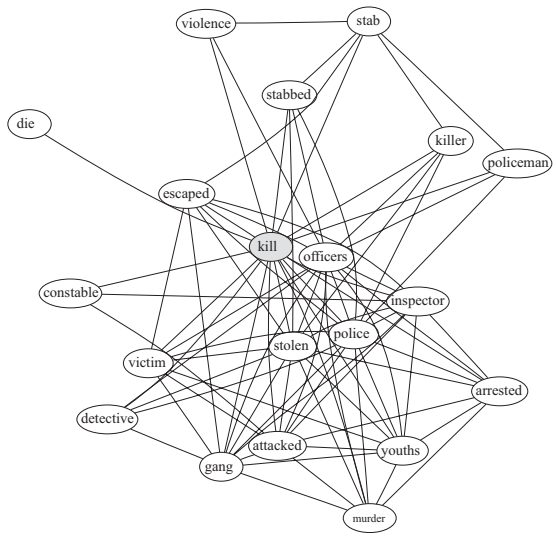

Figure 10. Neighborhood cliques of *kill* according to the independent model.

the models under consideration. For example, in each model, we can calculate the distance between each word and every other word. This leads to a neighborhood-distance vector. We can then measure the correlation between the neighbor-distance vectors for a given word across different models. More precisely, if $\mathbf{d}_i^0$ is the neighborhood-distance vector corresponding to w_i in model M_0 and $\mathbf{d}_i^1$ is the distance vector for w_i in model M_1, then we can calculate the correlation between these vectors, or $c_i^{01} = \text{corr}(\mathbf{d}_i^0, \mathbf{d}_j^1)$. Averaging over all words shared by M_0 and M_1, we can then measure the overall correlation between these two models. As there are five models that we need to compare with one another, that is, the experiential, distributional, combined, independent, and augmented distributional models, this leads to 10 pairs of overall correlations. To perform these comparisons, we restrict our analysis to the subset of words shared by all five models. This subset is essentially all the words in the experiential model, with the exception of a small number of high-frequency stop-words.

For each correlation we perform, we calculate the posterior distribution of the Pearson product–moment correlation coefficient. From this, we can calculate the $1 - \alpha$ high posterior density (HPD) regions. Although full details are provided in Appendix B, in brief, these are the regions of the posterior that contain $1 - \alpha$ of the probability mass and so denote the regions that have $1 - \alpha$ certainty of containing the true value of the correlation coefficient. In Figure 15, we present the $1 - \alpha$ HPD regions, with $\alpha \in \{.05, .5\}$, and posterior medians of the correlation coefficient for each of the 10 comparisons.

From these correlations, we see further evidence in support of our main claims. For example, with respect to our claim that the senses of words derived from experiential and distributional data differ markedly from one another, we see that of all the pairwise model comparisons, the correlation between the experiential and distributional models is the lowest, implying that the neighborhood structures of these two models are the least similar to one another. The claim that the combined model is the result of neither a mere averaging of two data sets nor a quantitatively larger data set is also corroborated by the pairwise correlations. We see that the neighborhoods derived from the distributional and experiential models are correlated with the independent model but are less so with the combined model. This corroborates the impression given by the neighborhood cliques that the representations derived from the independent model are relatively similar to those of the experiential and distributional models, while the representations in combined model tend to be more distinct. We also see from these correlations that the independent and combined models are not highly correlated with one another. This corroborates the claim that the combined model is capturing more information than merely the average of the experiential and distributional models. Finally, we see from these correlations that the original distributional model and its augmented counterpart are also relatively highly correlated. This corroborates the impression that training the distributional model with the addition of a larger text corpus does not lead to qualitatively different semantic representations.

Comparisons With Behavioral Data

If, as we are claiming, combining experiential and distributional data leads to richer semantic representations, we should expect that the interword semantic similarities in the combined model will more closely resemble those obtained from behavioral measures of semantic representations. Likewise, if the representations obtained from the combined model are richer than those obtained by simply averaging over two data sets or by using a quantitatively larger data set, we should expect that the combined model will more closely resemble human-based data than either the independent model or the larger augmented distributional model.

For the purpose of this analysis, we used six data sets providing measures of semantic similarity: lexical substitution errors, two sets of word-association norms, two sets of lexical priming data, and PWI data, all of which are described in more detail below. These data sets were chosen as they represent a range of different behavioral measures in language production and comprehension, all of which reflect semantic similarity among words at a fine-grained level.

Lexical substitution errors. Lexical substitution errors—spontaneous errors in speech in which one word mistakenly replaces the intended word—are among the most frequent type of slips of the tongue (Bock, 1991). For the majority of these, the error exhibits some meaning resemblance to the intended word, such as saying *finger* instead of *hand* or *car* instead of *bike* (Garrett, 1992). Such errors are generally assumed to reflect erroneous selection of a lexico-semantic representation corresponding to the concept the speaker wants to express (Butterworth, 1989; Dell, 1986; Garrett, 1984, 1992, 1993; Levelt, Roelofs, & Meyer, 1999), and as such, the more highly related a word's meaning is to the target's, the more likely that word is to be produced as an error (Vigliocco et al., 2004). Such data therefore represent semantic similarity effects that arise naturally through automatic language-production processes. Here, we used a data set consisting of a collection of spontaneously occurring lexical substitution errors collected over a period of several years by Harley, who recorded every error made by adult native English speakers (or produced himself) in intervals decided in advance (for discussion and anal-

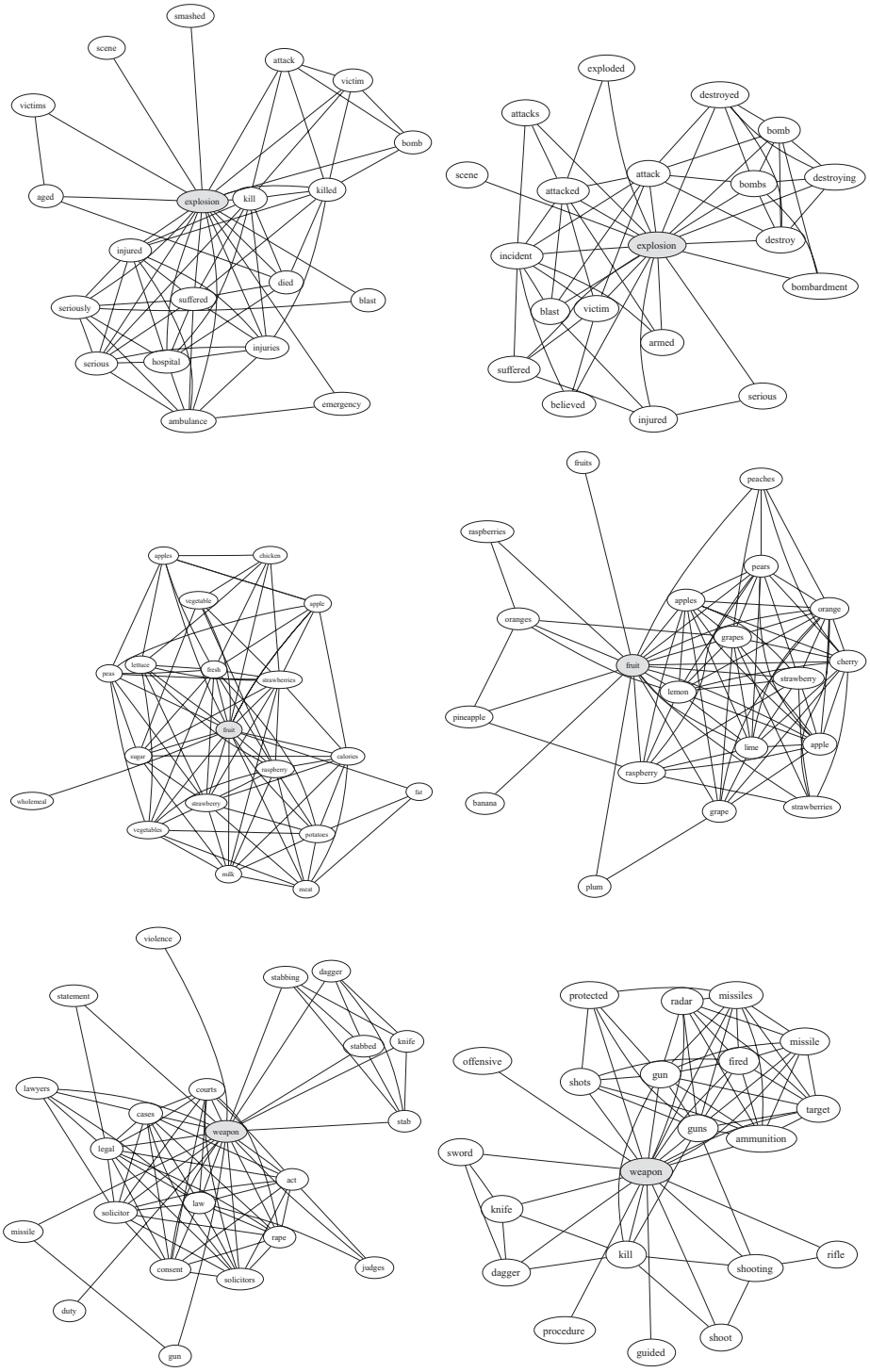

Figure 11. Neighborhood cliques of *explosion, fruit,* and *weapon* in the distributional model (left subfigure) and the combined model (right subfigure). These graphs illustrate the bridging structures that occur when data types are combined. For example, unlike in the distributional model, in the combined model, *explosion* becomes related to *destroy* and *bombardment.* This is because, on the basis of distributional data, *explosion* is related to *bomb,* while, on the basis of experiential data, *bomb* is related to *destroy* and *bombardment.* When experiential and distributional data are combined, these chains allow new relationships to become apparent.

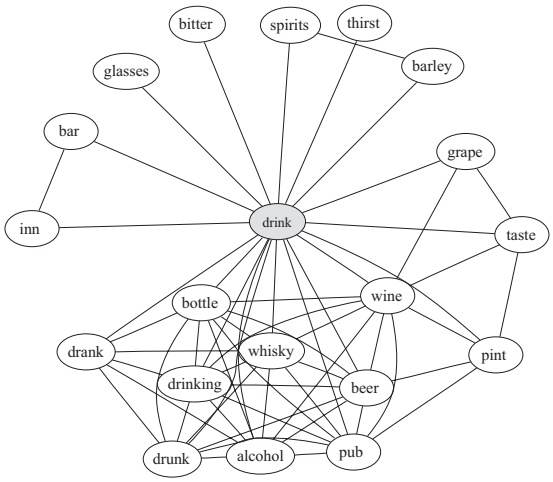

Figure 12. Neighborhood cliques of *drink* according to the augmented distributional model.

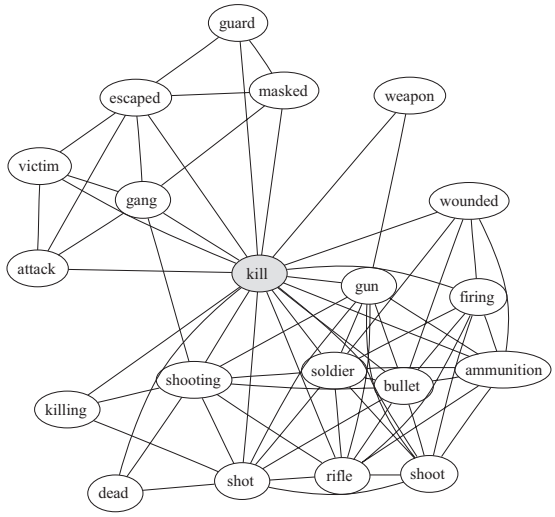

Figure 14. Neighborhood cliques of *kill* according to the augmented distributional model.

yses based on these data, see Harley, 1984; Harley & MacAndrew, 1993, 2001). Of this data set, a set of 420 observed errors were of a semantic, as opposed to phonological, nature. Of these 420, 61 target–errors pairs involved words that occurred in all our models and so could be used in the analysis.

Substitution error data take the form of sets of ordered word pairs, for example, (w_i, w_j), that indicate that w_j was observed as an erroneous substitution for w_i. For each word pair, we can measure the distance between w_i and w_j according to each model. A model with a low average distance for the entire set of substitutions implies that the model regards the members of the word pairs as being, on average, semantically similar. As substitution errors are taken to be indicative of semantic similarities, the higher the average similarity according to a given model, the better that model resembles human patterns of semantic similarities. To compare average distances across models, it is necessary to convert the distances into percentile scores.[5] For example, if the distance between w_i and w_j is assigned a percentile score of .75, this means that w_j is closer to w_i than .75 of other words. The higher the percentile score of the pair of words, the higher similarity according to the model. In other words, the higher the percentile score, the closer, or less distant, is the pair of words.

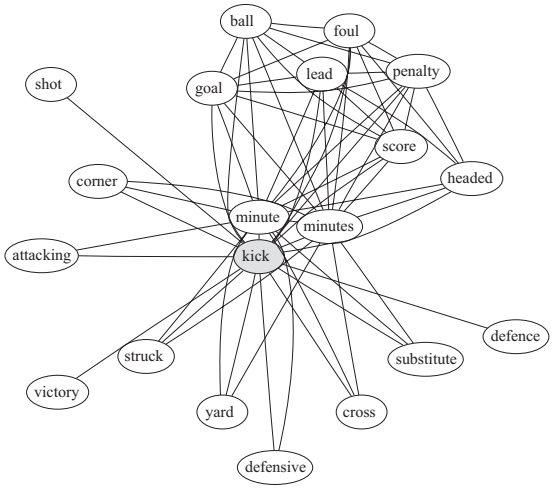

Figure 13. Neighborhood cliques of *kick* according to the augmented distributional model.

To perform statistical analyses, we can treat the percentile scores from each model as samples from normally distributed random variables. Using a Bayesian analysis of variance (ANOVA; see Appendix B for details), we can then determine the HPD regions for the means of these variables, and examine the credible intervals for the differences between the means of each pair of models.

In the upper subfigure of Figure 16, we provide the $1 - \alpha$ HPD regions for $\alpha \in \{.05, .5\}$ and posterior medians for the mean percentile scores of the substitution errors for each of our five models. In the lower subfigure, we provide the same intervals for the differences between the means of the combined model and each of the remaining four models. As the figure illustrates, the mean percentile score of the errors according to the combined model is estimated to be in an interval centered at approximately .97, which is noticeably higher than those of the other models. Overall, the HPD regions of the five models are reliably distinct. The probability that the true means of the five models are identical is $< .02$. Likewise, from the lower subfigure, the HPD regions for the differences between the mean of the combined model and those of the other four models are all clearly greater than zero. The closest model to the combined model is the distributional model. The probable difference between it and the combined model is approximately 7 percentile points.

Association norms. We next turn to sets of association norms: sets of words generated when participants are asked to produce the

[5] We use percentile scores here and elsewhere as the absolute values of distances may vary across models due to factors such as model size, complexity, or the number of items in the model's vocabulary. The relevant quantities are thus provided by relative, rather than absolute, scores and, in particular, by the percentile.

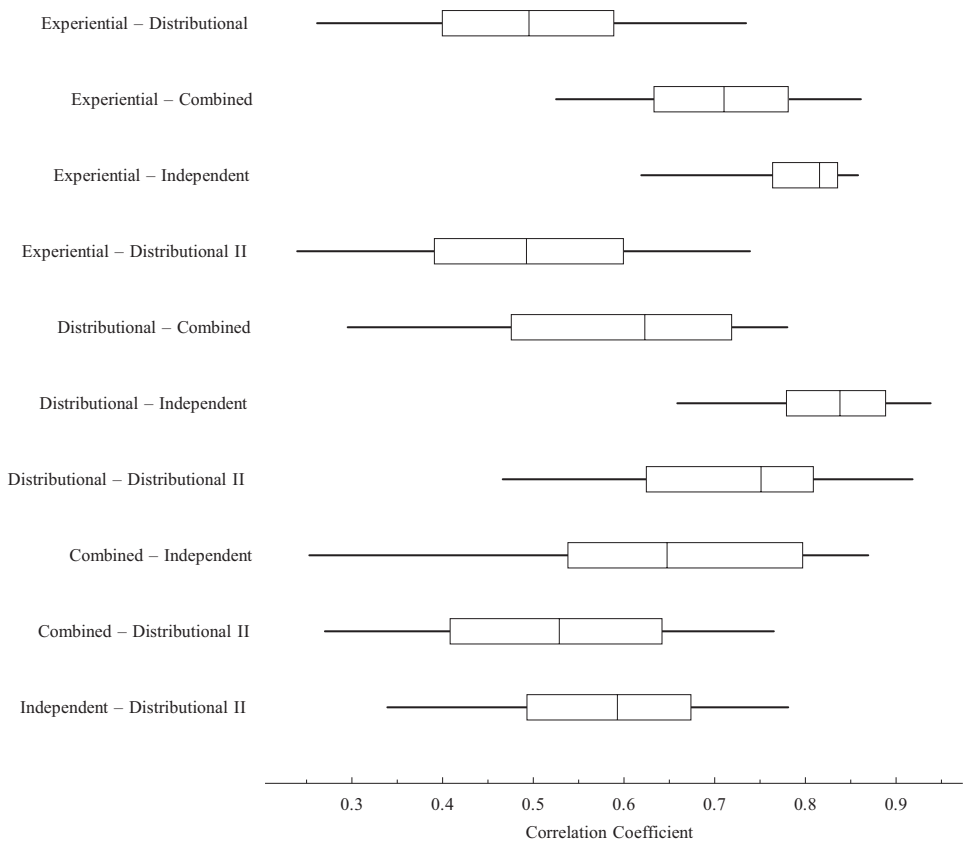

Figure 15. The regions of high posterior density of the pairwise correlations between the lexical neighborhoods in of the five models, that is, experiential, distributional, combined, independent, and augmented distributional models. We label the augmented distributional model here as Distributional II.

first word that comes to mind given a target word. Instructions for such tasks typically request participants to produce words that are "meaningfully related or strongly associated" (Nelson et al., 2004, p. 403) but do not provide any further constraints upon their responses. As such, the range of responses typically includes a very wide range of relations differing across domains of meaning. Inspection of these productions, however, reveals that nearly all of them are related in meaning to the target in one way or another. Such data sets thus serve as a useful counterpart to lexical substitution errors, in that associates are produced via conscious decision processes rather than automatically but nonetheless exhibit strong meaning relations to the target words. An added advantage is that there exist two very large sets of association norms including thousands of words and collected from numerous participants: one using speakers of British English—the Edinburgh Associative Thesaurus (EAT; http://www.eat.rl.ac.uk; see Kiss, Armstrong, Milroy, & Piper, 1973)—and one using speakers of American English—the Nelson association norms (Nelson et al., 2004). From the latter data set, a total of 746 unique word pairs could be used, while from the former a total of 545 word pairs could be used.

Association norms have a similar data character as the substitution norms in that they can also be conceived as sets of ordered word pairs. For example, the pair (w_i, w_j) indicates that w_j has been recorded as an associate of w_i. Association norms also provide an integer n_{ij} that gives the number of times that w_j has been recorded as an associate of w_i. This is equivalent to having the pair (w_i, w_j) occur exactly n_{ij} times in the set. We can then follow the same procedure as described for the case of substitution errors and, for each model, calculate the semantic similarity percentile scores for the set of association norms. The higher the model's average percentile score, the closer its semantic similarities resemble those given by the association norms. We may also analyze these data using the Bayesian ANOVA model as we did for the lexical substitution errors.

In Figures 17 and 18, we present the HPD regions from these analyses for the case of the EAT and the Nelson association norms, respectively. The two cases of the association norms strongly resemble those of the lexical substitution errors presented above. In these cases, because of the large quantity of data points, all the HPD regions are very narrow. The probability that the true means of the five models are identical is negligible, that is, $<10^{-10}$ in both cases. The combined model is estimated to have the highest mean, described by a credible interval centered at approximately .95 in the case of the EAT norms and .96 in the case of the Nelson norms. In both cases, the nearest model is the independent model, being approximately .02 percentile points below the combined model.

Semantic priming in word recognition. In addition to the above data sets including sets of word pairs that are related in meaning, we also investigated chronometric measures of semantic similarity, where degree of semantic similarity is reflected in temporal effects on word recognition or production. We begin with

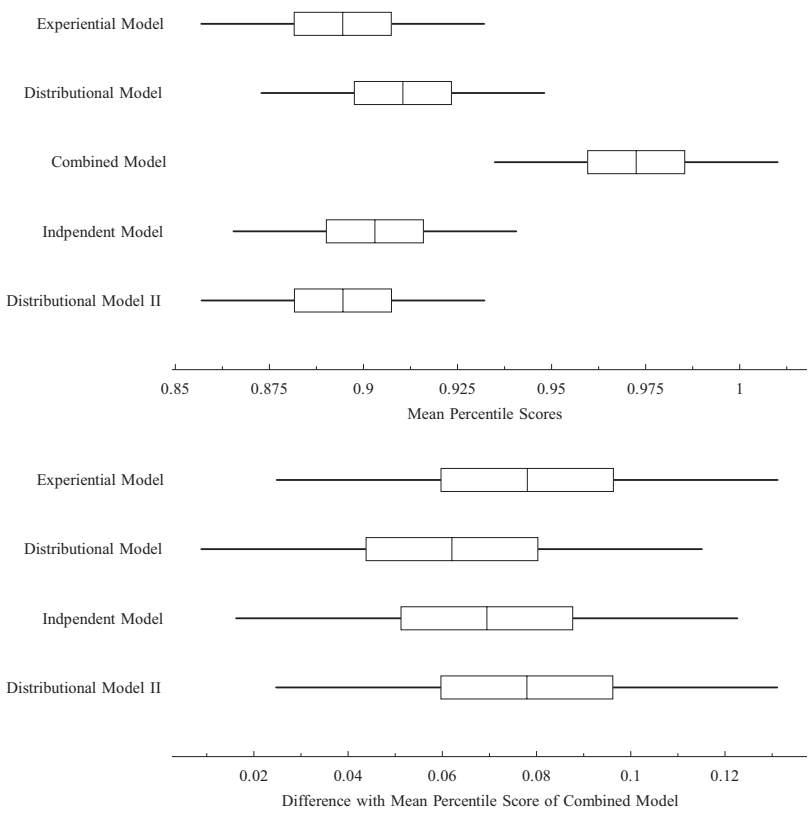

Figure 16. The high posterior density regions of the mean percentile score for each model with respect to the lexical substitution errors (upper subfigure) and the high posterior density regions of the difference between the means of the combined model and the means of each other model (lower subfigure). We label the augmented distributional model here as Distributional II.

semantic priming in lexical decision, the robust finding that speakers typically respond faster to a target word when it is preceded by a semantically related word than when it is preceded by an unrelated word (Meyer & Schvaneveldt, 1971). This finding of semantic priming in lexical decision tasks (word–nonword) has been largely investigated because it has been considered to directly reflect the organization of semantic memory (e.g., Anderson, 1983; Collins & Loftus, 1975; Cree, McRae, & McNorgan, 1999; McRae & Boisvert, 1998; McRae et al., 1997). Importantly, the degree of semantic relatedness between target and prime has been shown to modulate this priming effect (Vigliocco et al., 2004). Here, we tested the ability of the different models to predict the degree of priming effects observed in two studies: one employing lexical decision, Vigliocco et al. (2004), Experiment 5, and one employing semantic decision (a task that might be more sensitive to semantic relations), McRae et al. (1997), Experiment 2A. In the Vigliocco et al. data set, each of 32 experimental target words was paired with four prime words, varying in semantic similarity to the target (very close, close, medium, or far, according to Vigliocco et al.'s, 2004, measures). The McRae et al. data set instead derived from a set of 90 pairs of concepts varying in degree of similarity. In both cases, greater semantic similarity between pairs of words resulted in greater facilitatory effects in the decision tasks.

To assess the ability of each model to predict these results, we performed a correlational analysis of their correspondence with the interword similarities in each model. For priming data, the data consist of a set of word pairs (w_i, w_j). With each pair is associated a reaction time r_{ij} that gives the speed of response to w_i when preceded by w_j. A second reaction time r_i' provides a base-rate reaction time for w_i. The quantity $-(r_{ij} - r_i')$ gives the so-called priming effect of w_j on w_i. The higher this priming effect, that is, the faster the reaction time to w_i when preceded by w_j, the higher the implied semantic similarity between w_i and w_j. To compare these data with the models, we can perform a (Bayesian) correlational analysis (as described above for the case of the comparison between neighborhood structures across models) between the priming effects and the models' set of semantic distances. If the semantic similarities of the model resemble those revealed by the priming data, there should be a negative correlation between distances in the model and priming effects, that is, closer distances should have higher priming effects, and greater distances should have lower priming effects.

In Figures 19 and 20, we provide the $1 - \alpha$ HPD regions for $\alpha \in \{.05, .5\}$ and posterior medians for the correlation coefficients between each model's distances and the priming effects in the Vigliocco et al. (2004) data and McRae et al. (1997) data, respectively. Distances in the combined model are most highly anticorrelated with the observed effects. The medians of the HPD regions for the correlation coefficients of the combined model are $\rho = -.49$ in the case of the Vigliocco et al. priming data and $\rho = -.505$ for the McRae et al. priming data. For the Vigliocco et al.

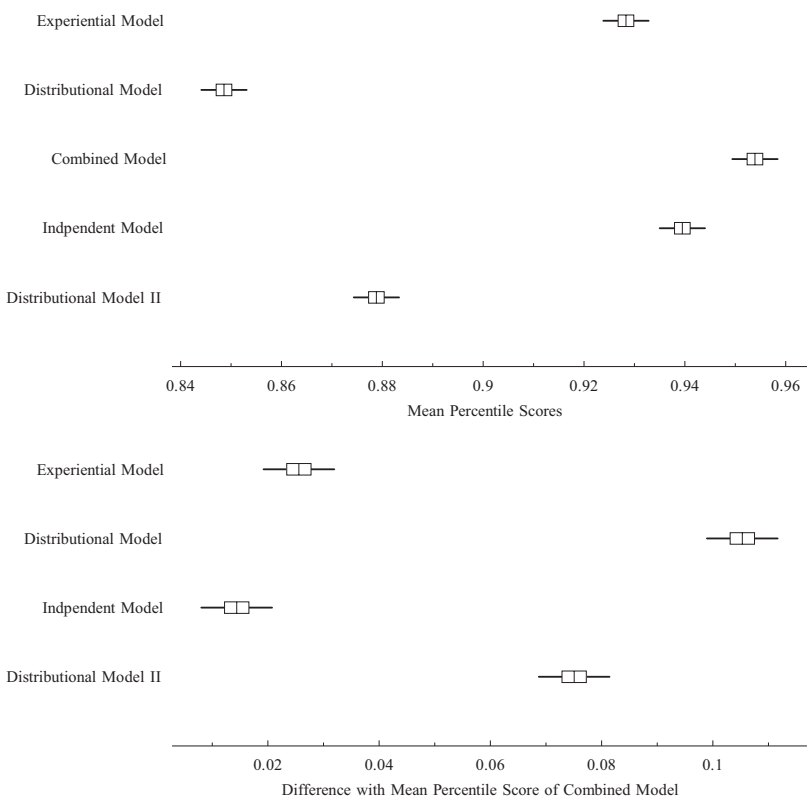

Figure 17. The high posterior density regions of the mean rank percentile score for each model with respect to the Edinburgh Associative Thesaurus association norms (upper subfigure) and the high posterior density regions of the difference between the means of the combined model and the means of each other model (lower subfigure). We label the augmented distributional model here as Distributional II.

priming data, the model with the next highest negative correlation is the experiential model (the median for credible interval is $\rho = -.44$), followed by the independent model (median of interval is $\rho = -.38$). Likewise, in the McRae et al. data, after the combined model, the next highest correlation is from the experiential model (the median for credible interval is $\rho = -.35$), followed closely by the independent model (median of interval is $\rho = -.34$).

Semantic interference in word production. Finally, we return to word production, to a chronometric task similar in character to lexical substitution errors. Lexical substitution errors are cases in which the conceptually driven lexical retrieval process fails in favor of another word semantically similar to the target. However, semantically related words also exert effects during accurate production, as shown in PWI experiments (e.g., Lupker, 1979; Rosinski, 1977). PWI experiments are based on a variant of the Stroop task (Stroop, 1935) in which a distractor word is presented immediately before a target picture to be named. In these experiments, speakers are slower to name the picture when the word is semantically related to the target than when the word is unrelated (Glaser & Düngelhoff, 1984; Schriefers, Meyer, & Levelt, 1990). Crucially, these effects have also been shown to be subject to gradation (Vigliocco et al., 2004) and, as such, provide a useful counterpart to the semantic priming effects described above. Here, we used data from Vigliocco et al.'s (2004) Experiment 3, where participants were asked to name each of 24 pictures that were paired with distractor words ranging in semantic similarity to the target word (very close, close, medium, or far). Data consisted of trimmed correct naming latencies averaged across 36 participants for each distractor–target pair.

PWI data are thus similar in their character to priming data. The data are also word pairs, for example, (w_i, w_j), with corresponding naming latencies r_{ij}. In this case, r_{ij} gives the speed to name a picture depicting the referent of w_i when preceded by w_j, while r_i' gives the base-rate reaction to picture naming of w_i. The quantity $(r_{ij} - r_i')$ gives the so-called interference effect, or the extent to which the presence of w_j inhibits the picture naming of w_i. A greater interference effect implies a greater semantic relationship between w_i and w_j. We can perform a correlation analysis with these interference effects and the semantic distances in each model. The more anticorrelated these are, the more the semantic similarities of the model resemble those revealed by interference norms.

In Figure 21, we provide the analysis for the PWI data. Again, distances in the combined model are those that are most highly anticorrelated with the observed effects. The median of the HPD regions for the correlation coefficients of the combined model is $\rho = -.21$ for PWI data. After the combined model, the next highest correlations are from the independent and the feature models (the medians of the HPD regions of both are $\rho = -.19$).

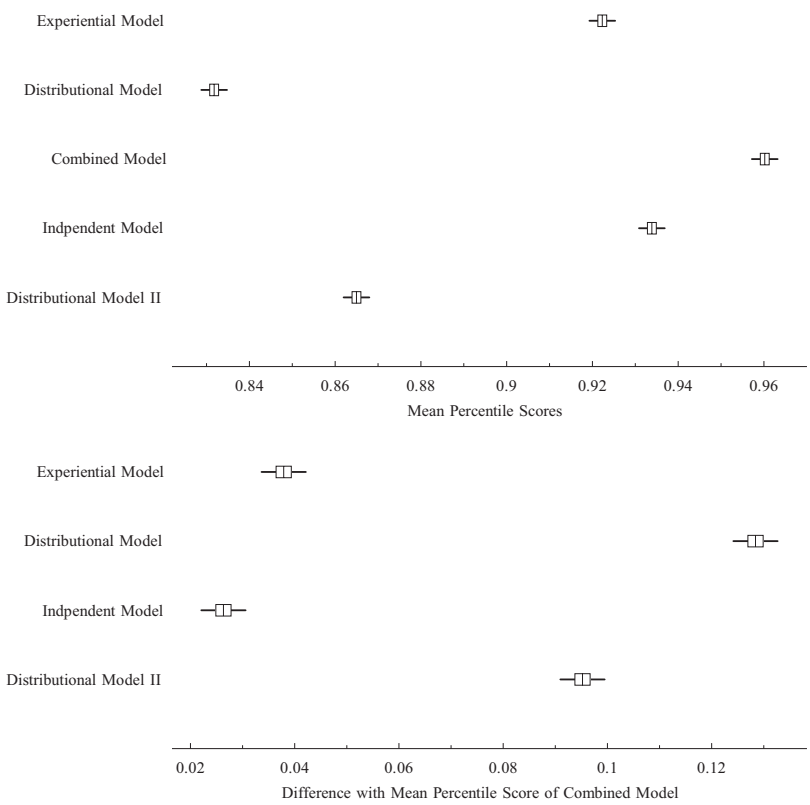

Figure 18. The high posterior density regions of the mean rank percentile score for each model with respect to the Nelson, McEvoy, and Schreiber (2004) association norms (upper subfigure) and the high posterior density regions of the difference between the means of the combined model and the means of each other model (lower subfigure). We label the augmented distributional model here as Distributional II.

Across this set of results from six disparate data sets, a common pattern emerges. In each case, the combined model provides a closer correspondence to the human data than do the other models. In most cases, either the independent model or the experiential model provides the second-closest correspondence. This is then followed by one or the other of the distributional models.

Discussion of Model Analysis and Evaluation

The primary claim of this article is that human semantic representations are derived from an optimal statistical combination of experiential and distributional data. The evidence that we have accumulated in support of these claims is as follows. We have

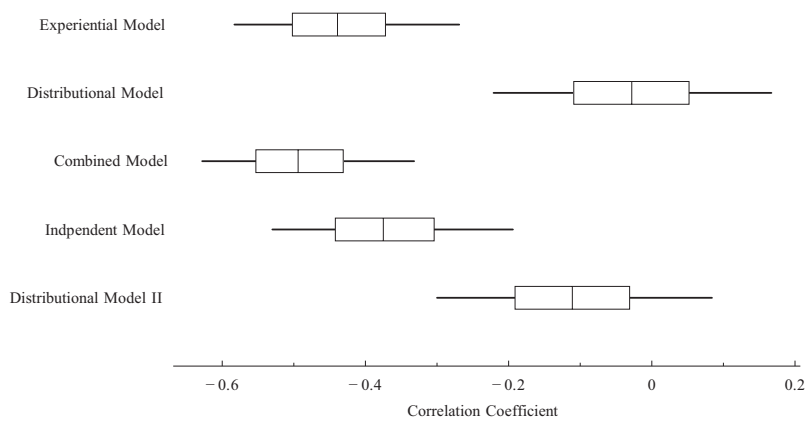

Figure 19. The high posterior density regions over the correlation coefficient between lexical decision reaction times and neighbor closeness in each model. Reaction time data are from Vigliocco, Vinson, Lewis, and Garrett (2004). We label the augmented distributional model here as Distributional II.

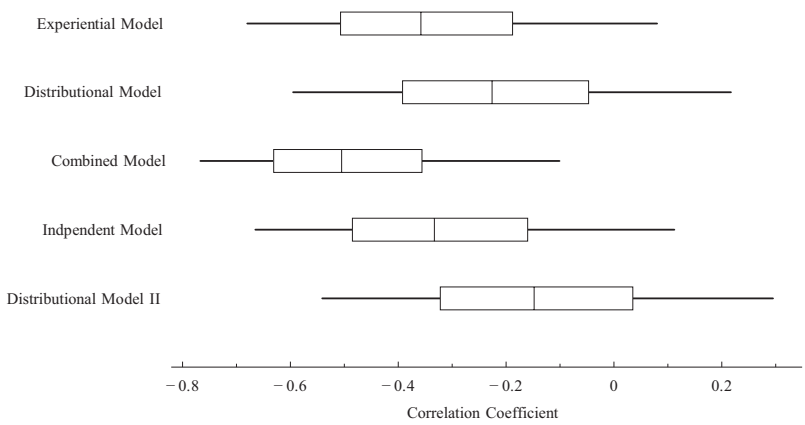

Figure 20. The high posterior density regions over the correlation coefficient between lexical decision reaction times and neighbor closeness in each model. Reaction time data are from McRae, de Sa, and Seidenberg (1997). We label the augmented distributional model here as Distributional II.

established that experiential and distributional data are separate and distinct data types and that both are nontrivial sources of semantic information. By examining the neighborhood cliques of words as they are represented in models trained using either experiential data alone or distributional data alone, it is evident that experiential and distributional data provide different types of information about the meaning of words. By reference to Figure 4 (but also Figures 5, 6, and 7), for example, it appears that experiential data provide semantic information related to the sensory, motor, and generally physical aspects of a given word. By contrast, distributional data provide semantic information that is more abstract or encyclopedic. In general, the neighborhood cliques in these examples suggest that the senses of words derived from experiential and distributional data differ markedly from one another and do so in the systematic manner we have described.

These findings show that both experiential and distributional data are valid, yet distinct, sources of semantic information. They imply that it is necessary to consider both data types to achieve a complete account of semantic representations and how they are learned from statistical data. They also imply that the common practice of modeling semantic representations by exclusive consideration of one data type or the other will necessarily lead to only a partial description of semantic representations. Neither modeling semantic representations by way of their physical and sensory features nor modeling semantic representation using distributional models, such as LSA or the models described in Griffiths et al. (2007), can likely lead to a complete or unbiased account of semantic representations. The former will lack reference to the more encyclopedic aspects of the word's meaning, while the latter will lack adequate details of the more physical, grounded, or embodied aspects of the word's meaning.

These conclusions are clearly premised upon the particular operational definitions of experiential and distributional data that we have adopted throughout this article. In particular, following the common practice found in the literature described in the introduction, we have defined distributional data as the frequency distribution of word across texts. This obviously neglects important information available in the statistics of language. Most dramatically, it neglects all information derived from the synatactic and sequential structure in text, which will be of a more fine-grained nature than that available from treating texts as so-called bags of words. Likewise, we have defined experiential data, also following common practices, in terms of frequency distributions over features obtained from speaker-generated feature norms. Clearly, these data are a drastic oversimplification of the sensory-motor data to which human beings are exposed.

Important as these reservations are, the differences between experiential and distributional data do not appear to be artefacts of the operational definitions we have used. As mentioned, the characteristics of representations derived from experiential data appear to be based on sensory, motor, and generally physical aspects of a given word. By contrast, those derived from distributional data are more abstract or encyclopedic. In particular and as hoped, the unique information derived from experiential appears to be based primarily on the perceived physical characteristics of words. It is difficult to see how this exact information could also be derived from nonexperiential data, such as within relations within language itself. Indeed, explicit attempts to extract information of precisely this kind from language by using text patterns such as "the * of the *C* [is | was]" and so on, where *C* is a given concept, have shown that this is a formidable challenge (see, e.g., Poesio & Almuhareb, 2004).

Assuming the validity of our operational definitions and having established the distinction between experiential and distributional data, we can address the hypothesis that semantic representations are derived from their statistical combination. From Figures 5, 6, and 7, it is apparent that the combined model integrates the information inherent in its two constituent data types to generate new semantic representations that are more coherent and refined than that which is derived from either source individually. In addition, by reference to Figures 8, 9, and 10, we see that, in contrast to the case of the combined model, representations in the independent model are not interesting syntheses of the senses derived from the constituent models. They are simply the result of an average of the constituent senses, emphasizing their overlapping characteristics, if they exist. This serves to illustrate the importance of what we have termed statistical combination and the extent to which it goes beyond a mere summation of the informa-

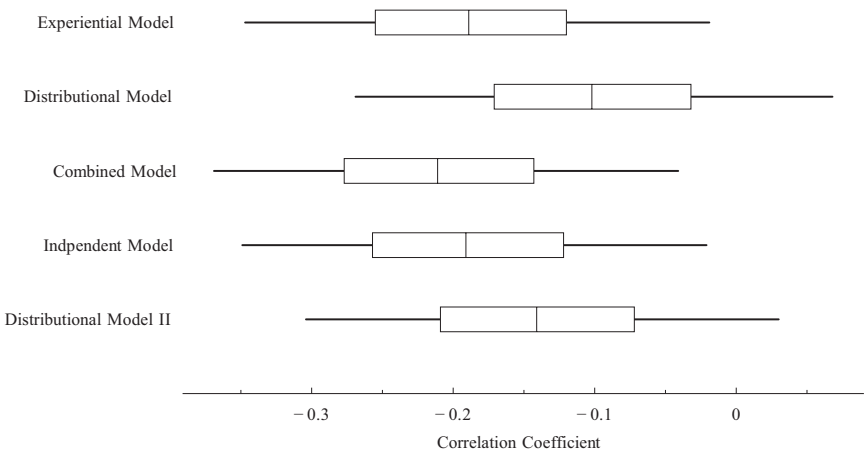

Figure 21. The high posterior density regions over the correlation coefficient between picture-word naming latencies and neighbor closeness in each model. Latencies are from Vigliocco, Vinson, Lewis, and Garrett (2004). We label the augmented distributional model here as Distributional II.

tion derived from two separate sources. Statistical combination treats two data sets as a joint whole and will lead to the discovery of patterns that characterize the correlations both within and between the constituents. By contrast, summing the senses derived from each data set merely leads to an average of both, emphasizing their overlapping characteristics but not leading to a new sense that is their synthesis. The difference between statistical combination and mere averaging is further illustrated by the graphs in Figure 11. These graphs show that when the two data types are combined, the relationship between the structures within each individual data type will now be revealed, and what we have termed bridging structures can be inferred between them. Thus, greater amounts of semantic information are now apparent than would be possible using either one source alone or both data types independently.

We have addressed whether the semantic representations obtained by the combined model are merely the result of a larger training set. We have trained the distributional model, using an augmented training set and shown in Figures 12, 13, and 14, that the representations learned are not remarkably different from those learned in original training corpus. From these examples, we would argue that it is not simply the combination of any two data sets of sufficient size that leads to the kind of richer or more refined semantic representation found in the combined model. What is special about the combined model is that is draws upon two distinct data types that nonetheless are intercorrelated. In other words, the experiential and distributional data types each provide distinct senses of a given word, with the former providing more grounded information and the latter providing more abstract and encyclopedic information. As these senses are not incompatible but rather form a continuum of senses, joining the two data sets and learning from both concurrently will lead to a richer and more comprehensive meaning. By contrast, another text corpus, even if it supplies a far greater quantity of data, will not be qualitatively distinct from the original in the same manner. Both text corpora will be restricted to encyclopedic information, and their combination will likely not lead to qualitatively new and interesting new senses.

These claims are borne out further by reference to the pairwise correlations between the neighborhoods in the models under consideration. The HPD regions for these correlation coefficients are shown in Figure 15. We have seen that of all the pairwise model comparisons, the correlation between the experiential and distributional models is the lowest, implying that the neighborhood structures of these two models are the least similar to one another. In addition, the neighborhoods derived from the distributional and experiential models are more highly correlated with the independent model than the combined model. This shows that the representations derived from the independent model are relatively similar to those of the experiential and distributional models, while the representations in the combined model tend to be more distinct. We have also observed that the original distributional model and its augmented counterpart are relatively highly correlated, showing that training the distributional model with a larger text corpus does not lead to qualitatively different semantic representations.

Finally, in the comparisons with the behavioral data, there is a common pattern to the results. In each case, the combined model provides a closer correspondence to the human data than the other models. In most cases, either the independent model or the experiential model provides the second-closest correspondence. This is then followed by one or other of the distributional models.[6] These results corroborate the evidence provided thus far and in themselves provide compelling evidence that human semantic representations are the product of a statistical combination of experi-

[6] The closer correspondence to the human data by the experiential model, rather than the distributional model, is likely due to the fact that the experiential model is particularly suited to the subset of words used for the analysis. As mentioned, by necessity, these words had to be those shared by all models and hence are all concrete words referring to mundane objects and events. It is reasonable to speculate that experiential model works best as a model of this small and constrained set of words. By contrast, the distributional model and, by extension, the independent model and the augmented distributional model will have much broader coverage, but be less likely to excel at this particular subset of words.

ential and distributional data, rather than either the product of one source alone or simply the sum of both.

General Discussion

Semantic representation, as we have used the term throughout this article, is knowledge about the meaning of words. It is the knowledge that allows a language user to infer, among other things, what words are similar or identical in meaning, what are the semantic or ontological categories to which a word belongs, and what (if anything) are the referents of a word. The general aim of this article has been to consider how this knowledge is acquired.

A standard empiricist approach to this problem holds that semantic knowledge, like all knowledge, is acquired from experience. As Locke (1689/1975) put it, "How comes [the mind] to be furnished? . . . To this I answer, in one word, from *experience*" (Book 2, chapter I). Traditionally, the type of experience with which empiricism has most concerned itself has been sensory experience, or the mass of data acquired by the senses. In the 20th century, however, attention has also focused on natural language itself as an important source of experience through which human beings learn about words. As Firth (1957) succinctly put it, "[we] know a word by the company it keeps" (p. 11). We have denoted these two major data types as experiential and distributional data, respectively.

At least three possible theoretical perspectives on the role of these sources can be adopted. The first is to choose one source and treat it as the primary or principal source of semantic knowledge. This, in fact, has been the prevailing practice in cognitive science in recent years. Despite compelling evidence in support of the utility of either, in the study of semantic representation, attention has generally been focused exclusively upon either experiential data or distributional data. There are, however, obvious limitations to either point of view. To exclude experiential data is to exclude data, grounded in the physical world, that provide direct knowledge of the referents of words. By contrast, to exclude distributional data is to exclude information about all words beyond those that are strictly concrete and that have tangible physical referents. Each source provides information that the other cannot. Treating them in a mutually exclusive manner is to discard otherwise valuable sources of semantic information.

An alternative perspective is to learn from both data sources independently and to thus arrive at semantic knowledge by an appropriate weighting of two essentially separate bodies of information. While perhaps more intuitively correct, this perspective, by treating the sources independently, will ignore any correlations between them. The information within both the experiential and distributional data may be utilized, yet the information between the two will be lost. A third perspective, therefore, the one we advocate here, is to treat experiential and distributional data as representing essentially one large joint data set. According to this perspective, semantic knowledge is gained by simultaneously learning from both the statistical structure within each source and the correlations between them. We have termed this process statistical combination to emphasize that is based on the learning of the structure of a joint statistical distribution and to distinguish it from the former case of merely summing or averaging over two independent data sets.

The evidence we have reported provides support for this general theoretical perspective. We have shown that experiential and distributional data represent distinct sources of semantic information. Each source is individually necessary, but not sufficient, to provide a complete account of semantic representations. We have shown that learning from both sources independently is likewise insufficient to provide plausible representations. On the other hand, as experiential and distributional data are distinct but nonetheless intercorrelated, their statistical combination not only captures the unique structures within each data type but also captures the bridging structures that exist between them. The resulting semantic representations are thus richer and more refined than those derived from either source individually or both independently.

Revisiting Plato's Problem

Semantic knowledge, just like all knowledge, is ultimately derived from a finite and inherently noisy body of data. The general question of how we are able to obtain any knowledge at all from finite and impoverished data has been referred to as Plato's Problem by, for example, Chomsky (1986). In the specific context of semantic knowledge, this is the problem of how human beings are able to proceed from the limited data that they experience to their knowledge of, for example, the similarity relationships between words, the categories into which words cluster, the different senses of words and how they change with context, and so on.

One general solution to Plato's problem is to use statistical inference to learn the underlying structure of the data to which human beings are exposed. On the basis of this structure, we can extrapolate to an essentially unlimited number of new inferences and predictions. For semantic knowledge in particular, this entails learning the structure underlying the statistical data that are related to words and representing words in terms of this structure. From this, we can then infer the extent to which any pair of words is related, the categories or groups to which words belong, and so on. Landauer and Dumais (1997) provided a specific implementation of this solution in proposing that semantic knowledge is acquired by learning the structure underlying the distributional statistics in language, representing words in terms of this learned structure, and making semantic generalizations accordingly.

While we agree with the general principle, we regard the work by Landauer and Dumais (1997) as providing a partial solution to Plato's problem. A more complete solution must take into account two important additional facts. First, there exist two major bodies of statistical data associated with words. In addition to the distributional statistics of language, there is the experiential data derived from human perception and interaction with the world. Second, these two bodies of data are intercorrelated. As such, the entirety of the data that human beings experience is given by the joint data set of experiential and distributional data. A more complete implementation of the solution to the problem is thus to learn the structure underlying this joint data set, representing words in terms of this structure, and making inferences and predictions on this basis.

The objective of this article has been to describe and justify this approach. We have shown that both experiential and distributional data provide distinct sources of semantic information. The structure underlying either one alone can provide a basis for reasonable, although evidently limited, generalization. Generalization is not greatly improved when the structures underlying both data sets

independently are used. However, when the joint structure of both data sets is learned, the resulting generalizations are clearly improved. This can be verified by comparing the interword similarities inferred by models based on each of these different data scenarios to human-based measures of semantic similarity. We can also understand why this improvement in generalization occurs by reference to, for example, Figure 11. There, we see that as a result of learning from both data sets jointly, not only are the structures within each individual data set learned but structures that bridge between them are also learned. These bridging structures create chains of inference that allow for generalizations. For example, *fruit* may be related to *apple* according to one data source, and *apple* may be related to *grape, lime,* and *plum* according to another. This leads *fruit* itself to be eventually related to *grape, lime,* and *plum.* Just as it is an elementary fact that marginal distributions cannot reveal all the information in a joint distribution, so it is that these generalizations are not possible when only one data type is used or when both data types are used independently.

Reconciling Language and the World

As described previously, the two broad theoretical positions in the study of semantic representations that we have identified, that is, the experiential tradition and the distributional tradition, both have inherent limitations, and there are points of serious conflict between them. By treating experiential and distributional data as a single joint source of data, however, we can overcome the limitations in each perspective and resolve the important sources of conflict between them.

The inadequacy of the experiential theoretical position is apparent when we consider the treatment of abstract words. Although there have been suggestions about how to ground abstract words in sensory-motor data (e.g., Barsalou & Wiemer-Hastings, 2005; Glenberg, Sato, & Cattaneo, 2008), it nonetheless remains a formidable challenge to describe how these words are learned from experiential data. Indeed, this problem equally applies to any words that do not have clear and obvious referents or to those whose meanings are not exhausted by these referents. For example, while it is apparent how mundane concrete words like *car, cat,* and *dog* could be learned from experiential data, the same cannot so easily be said about the words *confidence, fact,* and *punishment,* three nouns chosen randomly from the BNC. Combining the data sets avoids the problem of nonconcrete words simply because it incorporates distributional data, and distributional data are available for all words, irrespective of how concrete or tangible they are. As such, the primary source of data for nonconcrete words can be said to be data in language itself. More generally, this perspective argues strongly in favor of different modes of acquisition (see, e.g., Wauters, Tellings, Van Bon, & Van Haaften, 2003) for different words or different classes of words. According to this view, the data from which human beings learn any given word exist on a continuum from the purely experiential to the purely linguistic. The nonconcrete words can be said to be acquired primarily from language, while the more mundane concrete words can be said to be acquired primarily from experience with the world.

By contrast to the theoretical difficulties with the experiential tradition, the theoretical inadequacy of the distributional tradition stems from the fact that any knowledge acquired solely from language is disembodied from the world. While much about the relationship between words and other words can be discerned from distributional data, the same cannot be said about the relationship between words and the world. For example, although one may discern that the word *business* is related to the words *company, market, management,* and so on, one cannot know how any of these words relates back to the objects and events in the world. If the knowledge humans acquire from language is to be pragmatically useful, it must ultimately relate back to the world. In this sense, we can draw an analogy between the knowledge acquired from language and that acquired from natural science. If science is to be of any pragmatic value, no matter how abstract, theoretical, or mathematical it is, it must ultimately relate back to the concrete, physical, or human world. This general issue can be described as the problem of how language hooks onto the world.

Combining data sets allows words to ultimately hook onto the world through chains of inference. By combining experiential data, which directly hook words onto the world, with distributional data, which link words to other words, a model based on the joint data set can permit inferences to be made from all words to the world. We can get an intuition about this by considering the following, albeit contrived, scenario: On the basis of extralinguistic data, humans can learn that the word *dog* refers to those creatures in the world that bark, have waggy tails, have shaggy hair, chase cats, and so on. On the other hand, on the basis of intralinguistic data, humans know that a word like *canine* often appears in similar text and discourse contexts to *dog*. From this, even though one may never have been told what *canine* refers to, one might infer that *canine* is (in some sense) related to creatures in the world that bark, have waggy tails, have shaggy hair, chase cats, and so on.

We can explain this more formally by reference to probabilistic inference in the combined model. Given a word w_i, we can infer the probability of any feature y using

$$P(y|w_i) = \sum_{k=1}^{K} P(y|x_k)P(x_k|w_i), \quad (3)$$

where $P(x_k|w_i)$ is the semantic representation (i.e., the distribution over latent variables) of w_i. If w_i is a concrete word from our training set, the features inferred will correspond (subject to regularization) to the features with which it was paired in the training set. The more interesting case, however, is for words for which features have not been observed or do not exist. In these cases, in intuitive terms, the model infers features for w_i on the basis of how w_i is distributed across texts in a manner similar to other words for which features are known.

In Table 5, we provide the 10 most probable inferred features for a set of seven randomly chosen concrete words. Although they are all concrete words, as defined by the MRC psycholinguistic database (Coltheart, 1981), none of these words appeared in the feature-based training set. As far as the model is concerned, it has directly observed nothing about their referents. As is evident from this table, the model's inferences are intuitively valid. For example, the word *accident* is inferred to have properties related to death, carnage, and driving, and the word *army* is inferred to have properties relating to war and fighting. In all these cases, these inferred properties are due to

Table 5
Inferred or Predicted Features for Concrete Words

Accident	Army	Bowl	Cigar	Harbor	Alcohol	Rat
blood	attack	soup	smoke	sail	drink	fur
flesh	war	spoon	scent	boat	thirst	wild
life	destroy	stove	odor	ship	ingest	hop
kill	anger	pan	disgust	water	consume	cute
death	kill	oven	emit	float	liquid	long
drive	military	heat	breath	lake	swallow	whisker
wheel	oppose	carrot	air	ocean	enjoy	tail
passenger	explode	bake	inhale	cargo	glass	pet
vehicle	deadly	silverware	exhale	transport	mouth	squeak
window	pain	cook	gross	steer	stomach	teeth

how these words are distributed across texts like other words for which we have observed features. A word like *rat,* for example, is distributed quite similarly to the words *mouse* and, to a lesser extent, *cat,* for which features have been observed. On this basis, it is possible to infer that *rat* will share properties of *mouse* and *cat.* Indeed, notice that properties like *cute* and *pet* are ascribed to rats. Strictly speaking, of course, these are errors. However, it merely reflects that the model is inferring features on the basis of how words are statistically distributed. On the basis of this information, such inferences are completely reasonable.

The phenomenon illustrated by these examples demonstrates that knowledge acquired from within a language can lead to plausible predictions about the real-world referents of words. From this, it is apparent that even though distributional data per se are disembodied from the world, knowledge about how words are distributed coupled with knowledge about the features associated with a subset of words can together allow sensible predictions about how words relate to the world.

An interesting implication of this phenomenon relates to cognitive development. If intralinguistic data can provide information about the referents of words, it is plausible that the acquisition of general world knowledge may be bootstrapped from knowledge acquired through language. As we have seen, in the standard empiricist account of the learning of world knowledge, the referents of words are experienced as sets of features. What a word refers to and the various taxonomic and ontological categories into which these referents are organized are learned from patterns discovered in these data. However, according to the phenomenon discussed above, even if the referent of a word is not directly experienced, what it refers to in the world and, in particular, what its properties are can be inferred by using its distribution across language. Likewise, if the referents of a word have been experienced only infrequently, knowledge of its properties and extensions may be supplemented from knowledge acquired from language. A consequence of this is that knowledge of the world may be acquired faster, or with less direct experience with the world, by bootstrapping from data within language itself. This phenomenon is related to, and complements, the phenomenon of syntactic bootstrapping (e.g., Landau & Gleitman, 1985). The primary distinction between the two phenomena is that while in syntactic bootstrapping, fine-grained syntactic information is used, in distribution-based bootstrapping, more coarse-grained linguistic data are used to facilitate conceptual learning.

More generally, the above phenomenon addresses the objection raised by, for example, Glenberg and Robertson (2000) that states that intralinguistic data, particularly distributional statistics, are not capable of providing sufficient information from which the meanings of words can be learned. As mentioned previously, Glenberg and Robertson pointed out that although the relationships between words may be revealed by intralinguistic information, the relationship between words and the world cannot be known. In response to these concerns, we would argue that both the referents of words and, more generally, how language relates to the world can be derived from intralinguistic data. This is not direct but rather through a process of inference. Certain words, namely, the concrete words, refer to objects and events in the world. As words that are about similar things are distributed similarly across language, the distribution of any given word and how it resembles the distributional patterns of concrete words can lead to plausible inference about how this word relates to the world.

Importantly, this phenomenon is equally applicable to either concrete or nonconcrete words. Nonconcrete words—words either having no physical referents or whose meaning is not entirely constrained by these referents—can be hooked onto world through this chain of inferences. A consequence of this is that these words can be always be grounded, albeit indirectly, to sensory experience. In Table 6, we provide examples of predicted features for a set of seven randomly chosen abstract (as defined by the MRC database) words. Using the process described above, an abstract word's sensory-motor features are inferred on the basis of its distributional similarity to other words for which features have been observed. Thus, for example, although the word *feeling* may not have any observed physical instantiation or referent, it is statistically distributed with other words that do. Through this chain, we can infer that it is characterized by features such as *heart, touch, rough, love,* and so on. By way of the same process, an abstract word like *death* is characterized by features like *sad, sick, black,* and so on.

Treating data from the world and data from language as a single joint data source simultaneously resolves the twin problems of learning nonconcrete words and hooking language onto the world. The former represents a formidable challenge for the experiential tradition, while the latter is a frequent point of criticism of the distributional tradition. As each data source can provide information that the other lacks, joining the two is necessary for a complete account of the learning and representation of word meanings.

Table 6
Inferred or Predicted Features for Abstract Words

Obsession	Feeling	Sensation	Chaos	Luxury	Fashion	Death
women	heart	heart	demolish	comfort	women	sad
scream	touch	touch	attack	beautiful	grace	sudden
crave	rough	rough	explode	texture	beautiful	sick
desire	love	physical	war	cloth	style	harsh
relieve	sense	thirst	collide	gift	wear	black
love	texture	sense	impact	decorate	casual	cold
discomfort	hot	stomach	pain	gold	leather	dark
burn	finger	ingest	destroy	desire	cloth	bad
explode	cold	healthy	deadly	enjoy	cotton	violent
attack	hand	love	force	art	party	blood

Summary and Conclusion

The significance of this work is as follows. We have shown that there are two distinct types of statistical data from which human beings can learn the meanings of words. These are the statistics of the physical world, known through sensory experience, and the distributional statistics of words across language. Both of these experiential and distributional statistics provide substantial semantic information—experiential data provide grounded, sensory-motor-based information, and distributional data provide more abstract, encyclopedic information—and neither is derivative of, or dependent upon, the other. While these basic claims are in many ways obvious, the distinction between experiential and distributional data has not been properly addressed or explored in the literature up to this point.

We have shown that word meanings can be learned by using experiential and distributional data in combination, rather than by simply using one source exclusively. Also, this is a distinctly different process than simply using a weighted sum of the two data types.

Rather, it is a result of treating the two data sets as a single joint distribution and learning the statistical structure that underlies it. We have implemented this as a probabilistic model that is a straightforward generalization of models that have been used to learn semantic representations from distributional data alone (e.g., Griffiths et al., 2007). This gives us a single framework to explore the learning of semantic representations from data, regardless of their source, and greatly facilitates analysis and comparison of representations derived from different data types.

Semantic representations that are learned from both experiential and distributional data in combination are measurably more realistic than those available from either source alone or from both sources independently. This results not from merely using more data but rather from the fact that experiential and distributional data are qualitatively distinct data types, while also being correlated with one another. As a result, the semantic representations learned from the combined data are based on statistical structures both within and between the two constituent data types. We believe that these demonstrations represent important progress in an understanding of the nature of semantic representations and how they are learned from statistical data.

Finally, treating experiential and distributional data as a joint data set allows us to overcome some of the fundamental shortcomings in recent theoretical accounts of semantic representations. Models based on experiential data alone are limited in that they cannot account for the learning of abstract words. By contrast, those based on distributional data alone do not hook words onto the physical world. We can address the problem of abstract words by postulating a division of labor in the acquisition of different kinds of words, with the mundane concrete words being acquired largely through experiential data and the nonconcrete words being acquired largely through distributional data. We can address the problem of hooking onto the world as we can make plausible inferences about the referents of words from knowing about their distributions in language, thus allowing knowledge acquired from language to be related to knowledge acquired from the physical world.

References

Anderson, J. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology, 16,* 1818–1823.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 129–163). Cambridge, England: Cambridge University Press.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron, 34,* 149–159.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Bock, J. K. (1991). A sketchbook of production problems. *Journal of Psycholinguistic Research, 20,* 141–160.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioural study. *Cognitive Brain Research, 24,* 355–363.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 177–210.

Burnard, L. (2009). *What is the BNC?* Retrieved June 1, 2009, from http://www.natcorp.ox.ac.uk/corpus/index.xml

Butterworth, B. (1989). Lexical access and representation in speech production. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 108–135). Cambridge, MA: MIT Press.

Carey, S. (1985). *Conceptual change in infancy.* Cambridge, MA: MIT Press.

Chao, L. L., Haxby, J., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2,* 913–919.

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *NeuroImage, 12,* 478–484.

Chao, L. L., Weisberg, J., & Martin, A. (2002). Experience-dependent modulation of category-related cortical activity. *Cerebral Cortex, 12,* 545–551.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin and use.* New York: Greenwood Press.

Collins, A., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 12,* 240–247.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33*(A), 497–505.

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

Damasio, H., Grabowski, T. J., Tranel, D., Ponto, L. L. B., Hichwa, R. D., & Damasio, A. R. (2001). Neural correlates of naming actions and of naming spatial relations. *NeuroImage, 13,* 1053–1064.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391–407.

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review, 93,* 283–321.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience, 10,* 77–94.

Farah, M., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120,* 339–357.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford, England: Blackwell Publishers.

Garrett, M. F. (1984). The organization of processing structure for language production: Applications to aphasic speech. In D. Caplan, A. R. Lecours, & A. Smith (Eds.), *Biological perspectives on language* (pp. 172–193). Cambridge, MA: MIT Press.

Garrett, M. F. (1992). Lexical retrieval processes: Semantic field effects. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields and contrasts: New essays in semantic and lexical organization* (pp. 377–395). Hillsdale, NJ: Erlbaum.

Garrett, M. F. (1993). Errors and their relevance for models of language production. In G. Blanken, J. Dittman, H. Grim, J. Marshall, & C. Wallesch (Eds.), *Linguistic disorders and pathologies: An international handbook* (pp. 72–91). Berlin, Germany: Walter de Gruyter.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Chapman & Hall.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23,* 183–209.

Gerlach, C., Law, I., & Paulson, O. B. (2002). When action turns to words: Activation of motor-based knowledge during categorisation of manipulable objects. *Journal of Cognitive Neuroscience, 14,* 1230–1239.

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Erlbaum.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton, Mifflin.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics, 41,* 337–348.

Glaser, W. R., & Düngelhoff, R. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 640–654.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43,* 379–401.

Glenberg, A. M., Sato, M., & Cattaneo, L. (2008). Use-induced motor plasticity affects the processing of abstract and concrete language processing. *Current Biology, 18,* 290–291.

Grabowski, T. J., Damasio, H., & Damasio, A. R. (1998). Premotor and prefrontal correlates of category-related lexical retrieval. *NeuroImage, 7,* 232–243.

Griffiths, T., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray & C. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 381–386). Mahwah, NJ: Erlbaum.

Griffiths, T., & Steyvers, M. (2003). Prediction and semantic association. In S. T. S. Becker & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 11–18). Cambridge, MA: MIT Press.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, USA, 101,* 5228–5235.

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

Harley, T. A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science, 8,* 191–219.

Harley, T. A., & MacAndrew, S. B. G. (1993). Modelling paraphasias in normal and aphasic speech. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 378–383). Hillsdale, NJ: Erlbaum.

Harley, T. A., & MacAndrew, S. B. G. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research, 30,* 395–418.

Harris, Z. (1954). Distributional structure. *Word, 10,* 146–162.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigation of acquired dyslexia. *Psychological Review, 93,* 74–95.

Hofmann, T. (1999a). Probabilistic latent semantic analysis. In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 289–296). San Francisco: Morgan Kaufmann.

Hofmann, T. (1999b). Probabilistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval* (pp. 50–57). New York: ACM.

Howell, S., & Becker, S. (2001). Modelling language acquisition: Grammar from the lexicon? In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 457–462). Mahwah, NJ: Erlbaum.

Howell, S., Becker, S., & Jankowicz, D. (2001). Modelling language acquisition: Lexical grounding through perceptual features. In R. Pfeifer & G. Westermann (Eds.), *Proceedings of the 2001 Workshop on Developmental Embodied Cognition (DECO-2001)* (pp. 36–40). Edinburgh, Scotland: [no publisher named].

Howell, S., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve grammar learning. *Journal of Memory and Language, 53,* 258–276.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences, USA, 96,* 9379–9384.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Keil, F. C. (1989). *Concepts, kinds and cognitive development.* Cambridge, MA: MIT Press.

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh, Scotland: University Press.

Landau, B., & Gleitman, L. (1985). *Language and experience.* Cambridge, MA: Harvard University Press.

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

Landauer, T., Laham, D., & Foltz, P. (1998). Learning human-like knowledge by singular-value decomposition: A progress report. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 45–51). Cambridge, MA: MIT Press.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the nineteenth annual conference of the cognitive science society* (pp. 412–417). Mahwah, NJ: Erlbaum.

Lee, P. M. (2004). *Bayesian statistics: An introduction.* London, England: Hodder Arnold.

Levelt, W. J. M., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1–28.

Locke, J. (1975). *An essay concerning human understanding*. Oxford, England: Clarendon Press. (Original work published 1689)

Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review, 15,* 838–844.

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

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 660–665). Mahwah, NJ: Erlbaum.

Lupker, S. (1979). Semantic nature of response competition in the picture-word interference task. *Memory & Cognition, 7,* 485–495.

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development, 6,* 17–46.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognitive Development, 1,* 2–36.

Mandler, J. M., Bauer, P. J., & McDonough, L. (1991). Separating the sheep from the goats. *Cognitive Psychology, 23,* 263–298.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11,* 194–201.

Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., & Ungerleider, L. G. (1995, October 6). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science, 270,* 102–105.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996, February 15). Neural correlates of category-specific knowledge. *Nature, 379,* 649–652.

McClelland, J., & Rogers, T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4,* 310–322.

McClelland, J., & Rumelhart, D. (1985). Distributed memory model and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 558–572.

McRae, K., de Sa, V., & Seidenberg, M. (1997). On the nature and scope of featural representation of word meaning. *Journal of Experimental Psychology: General, 126,* 99–130.

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234.

Nelson, D., McEvoy, C., & Schreiber, T. (2004). The University of South Florida word association, rhyme and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 408–420.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Methodological, 56*(B), 3–48.

Norman, D. A. (1988). *The psychology of everyday things.* New York: Basic Books.

Oliveri, M., Finocchiaro, C., Shapiro, K., Gangitano, M., Caramazza, A., & Pascual-Leone, A. (2004). All talk and no action: A transcranial magnetic stimulation study of motor cortex activation during action word production. *Journal of Cognitive Neuroscience, 16,* 374–381.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Paivio, A. (1990). *Mental representations: A dual-coding approach.* New York: Oxford University Press. (Original work published 1986)

Phillips, J., Noppeney, U., Humphreys, G. W., & Price, C. J. (2002). Can segregation within the semantic system account for category-specific deficits? *Brain, 125,* 2067–2080.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.

Poesio, M., & Almuhareb, A. (2004). Feature-based vs. property-based KR: An empirical perspective. In A. C. Varzi & L. Vieu (Eds.), *Formal ontology in information systems: Proceedings of the third conference (FOIS-2004)* (pp. 177-184). Amsterdam: IOS Press.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences, 22,* 253–336.

Pulvermüller, F. (2001). Brain reflections of words and their meanings. *Trends in Cognitive Sciences, 5,* 517–524.

Pulvermüller, F., & Hauk, O. (2005). Category-specific conceptual processing of color and form in left fronto-temporal cortex. *Cerebral Cortex, 16,* 1193–1201.

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience, 21,* 793–797.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12,* 410–430.

Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM, 12,* 459–476.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1–20.

Rogers, T., & McClelland, J. (2005). *Semantic cognition: A parallel distributed processing account.* Cambridge, MA: MIT Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Rosinski, R. (1977). Picture-word interference is semantically based. *Child Development, 48,* 643–647.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26,* 113–146.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Schütze, H. (1992). *Dimensions of meaning.* Washington, DC: IEEE Computer Society Press.

Schriefers, H., Meyer, A., & Levelt, W. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language, 29,* 86–102.

Sivia, D. S., & Skilling, J. (2006). *Data analysis: A Bayesian tutorial.* Oxford, England: Oxford University Press.

Smith, E., Shoben, E., & Rips, L. (1974). Structure and process in semantic memory: Featural model for semantic decisions. *Psychological Review, 81,* 214–241.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 333–338.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 8,* 291–309.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., et al. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience, 17,* 273–281.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75,* 195–231.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48,* 422–488.

Vigliocco, G., Warren, J., Siri, S., Arciuli, J., Scott, S., & Wise, R. (2006, December). The role of semantics and grammatical class in the neural representation of words. *Cerebral Cortex, 16,* 1790–1796.

Vinson, D., & Vigliocco, G. (2002). A semantic analysis of noun-verb dissociations in aphasia. *Journal of Neurolinguistics, 15,* 317–351.

Vuilleumier, P., Henson, R. N., Driver, J., & Dolan, R. J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience, 5,* 491–499.

Wauters, L. N., Tellings, A. E., Van Bon, W. H., & Van Haaften, A. W. (2003). Mode of acquisition of word meanings: The viability of a theoretical construct. *Applied Psycholinguistics, 24,* 385–406.

Wittgenstein, L. (1997). *Philosophical investigations*. Oxford, England: Blackwell Publishers. (Original work published 1953)

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18,* 414–420.

(*Appendixes follow*)

Appendix A

Formal Description of Models

The models[A1] used in this article are based on the latent Dirichlet allocation (LDA) model introduced by Blei et al. (2003) and also used, for example, in Griffiths and Steyvers (2002, Griffiths and Steyvers, 2003) and Griffiths et al. (2007). For both the experiential and distributional models, we use the basic form of the LDA model as presented in Blei et al. For the combined model, we use a simple extension of this basic form. In all cases, as in Griffiths and Steyvers and in Griffiths et al., we use Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution over the parameters of the models, rather than using the variational approximation to the maximum-likelihood estimate as described in Blei et al. Note, however, that the MCMC methods we use differ from those used in Griffiths and Steyvers and in Griffiths et al..

In what follows, we present the LDA model we use for the case of the distributional model. We present this case first as LDA models are most commonly presented as bag-of-words language models. We then show how an identical model can be used for the experiential model. Finally, we show how this basic form can be simply extended for the case of the combined model.

Distributional Model

A language corpus $\mathbf{w}_{1:J}$ consists of a set of J texts $\{\mathbf{w}_1 \ldots \mathbf{w}_j \ldots \mathbf{w}_J\}$. The jth text, that is, $\mathbf{w}_j$, consists of n_j words $\{w_{j1} \ldots w_{ji} \ldots w_{jn_j}\}$, with each word being an element of a vocabulary v of size V. The bag-of-words assumption discards the sequential order of these words. It assumes that each text is an unordered set of words, and so, the only pertinent information is the frequency of occurrence of each word type within each text. The LDA model is a hierarchical mixture model that specifies a generative model for data of this nature. In particular, it specifies that each word w_{ji} is drawn from one of K discrete probability distributions $\varphi = \{\varphi_1 \ldots \varphi_k \ldots \varphi_K\}$. Which one of these K distributions is chosen is indicated by the value of the latent variable $x_{ji} \in \{1 \ldots k \ldots K\}$ that corresponds to w_{ji}. The latent variable x_{ji} is drawn from a discrete probability distribution π_j of length K that is specific to text j. In turn, each π_j is drawn from a symmetric Dirichlet distribution with parameters α, that is,

$$P(\pi_j|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha^K)}\prod_{k=1}^{K}\pi_{jk}^{\alpha-1}. \quad (A1)$$

The parameters in the model are φ and α, while each π_j and x_{ji} are latent, that is, unobserved, variables. A vague prior on α is an inverse gamma function with shape and scale parameters both equal to 1, that is,

$$P(\alpha) = \exp\left(-\frac{1}{\alpha}\right)\alpha^{-2}. \quad (A2)$$

A prior for φ is a symmetric Dirichlet prior with hyperparameter β, that is,

$$P(\varphi|\beta) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta^V)}\right)^K \prod_{k=1}^{K}\prod_{v=1}^{V}\varphi_{kv}^{\beta-1}. \quad (A3)$$

The hyperparameter β itself can be given an inverse gamma prior, that is,

$$P(\beta) = \exp\left(-\frac{1}{\beta}\right)\beta^{-2}. \quad (A4)$$

The posterior distribution over the parameters and hyperparameters given the data is

$$P(\varphi, \alpha, \beta|\mathrm{w}_{1:J}) \propto P(\mathrm{w}_{1:J}|\varphi, \alpha)P(\varphi|\beta)P(\alpha)P(\beta), \quad (A5)$$

where $P(\mathbf{w}_{1:J}|\varphi, \alpha)$ is the likelihood of observing the data $\mathbf{w}_{1:J}$ given φ and α, that is,

$$P(\mathrm{w}_{1:J}|\varphi, \alpha) = \prod_{j=1}^{J}\int d\pi_j \prod_{i=1}^{n_j}P(\pi_j|\alpha) \sum_{k=1}^{K}P(w_{ji}|x_{ji} = k, \varphi)P(x_{ji} = k|\pi_j). \quad (A6)$$

This posterior is analytically intractable, but a Gibbs sampling MCMC method may be used to draw samples from it. Denoting the latent variables by $\mathbf{x}_{1:J}$ and the set of J mixture proportions by $\pi_{1:J}$, we can draw samples from the joint posterior

$$P(\mathrm{x}_{1:J}, \varphi, \pi_{1:J}, \alpha, \beta|\mathrm{w}_{1:J}) \quad (A7)$$

by initializing the parameters, hyperparameters, and latent variables at arbitrary values and iteratively drawing samples from $P(\mathbf{x}_{1:J}|\varphi, \pi_{1:J}, \mathbf{w}_{1:J})$, $P(\varphi|\mathbf{x}_{1:J}, \mathbf{w}_{1:J}, \beta)$, $P(\pi_{1:J}|\mathbf{x}_{1:J}, \alpha)$, $P(\alpha|\pi_{1:J})$, and $P(\beta|\varphi)$. The first three of these conditional posterior distributions are analytically tractable—

$$P(\mathrm{x}_{1:J}|\varphi, \pi_{1:J}, \mathrm{w}_{1:J}) \propto \prod_{j=1}^{J}\prod_{i=1}^{n_j}P(w_{ji}|x_{ji} = k, \varphi)P(x_{ji} = k|\pi_j) \quad (A8)$$

is a discrete distribution,

$$P(\varphi|\mathrm{x}_{1:J}, \mathrm{w}_{1:J}, \beta) \propto \prod_{k=1}^{K}\prod_{v=1}^{V}\varphi_{kv}^{\beta+r_{kv}-1}, \quad (A9)$$

where $r_{kv} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbb{I}(x_{ji} = k, w_{ji} = v)$, is a Dirichlet distribution (the function $\mathbb{I}(\cdot)$ is an indicator function, taking the value 1 if its argument is true, and zero otherwise), and likewise

$$P(\pi_{1:J}|\mathrm{x}_{1:J}, \alpha) \propto \prod_{j=1}^{J}\prod_{k=1}^{K}\pi_{jk}^{\alpha+s_{jk}-1}, \quad (A10)$$

[A1] Python and Fortran code implementing these models can be found at http://www.mjandrews.net/code.

where $s_{jk} = \sum_{i=1}^{n_j} \mathbb{I}(x_{ji} = k)$, is a Dirichlet distribution—and samples are easily drawn from them. Both $P(\alpha|\pi_{1:J})$ and $P(\beta|\varphi)$ are log-concave univariate distributions and can be sampled by adaptive rejection sampling (e.g., Gilks & Wild, 1992).

In practice, initialization of the Gibbs sampler is important, particularly for large data sets. We have used the variational method described in Blei et al. (2003) to initialize φ and α to their approximately maximum-likelihood values. We have monitored convergence using simple trace-plots of the log-likelihood of the model (given the samples at time t) against time. Upon convergence, samples may be drawn at lagged (e.g., every 10 iterations) intervals. Samples of φ, α, and β will be approximately distributed as $P(\varphi, \alpha, \beta|\mathbf{w}_{1:J})$.

Finally, we can choose the optimal value of the number of latent component distributions, namely, K, by calculating the approximate value of the marginal likelihood of each model for a range of values of K and choosing the model with highest value. Following Griffiths and Steyvers (2004), we approximate the marginal likelihood by the harmonic mean of the model likelihood given samples from the posterior (see also Kass & Raftery, 1995; Newton & Raftery, 1994).[A2] The optimal number of latent components for the experiential, distributional, and combined models were $K = 120$, $K = 250$, and $K = 350$, respectively.

Experiential Model

The experiential model is identical in form to the distributional model, with the only difference being the nature of the data. In what follows, we reuse some of the notations for the case of the distributional model, with the hope that their meanings will be unambiguous from the context.

A feature corpus consists of a set of J objects or events, $\mathbf{y}_{1:J} = \{\mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_j \ldots \mathbf{y}_J\}$. Each $\mathbf{y}_j$, which may be labeled by a single word, is an unordered set of n_j observations, with each observation being of one of F primitive or elementary features, that is, $\mathrm{y}_j = \{y_{j1} \ldots y_{j2} \ldots y_{ji} \ldots y_{jn_j}\}$ with $y_{ji} \in \{1 \ldots F\}$. As in the case of bag-of-words models, the only pertinent information in this set is the frequency of occurrence of each individual feature for each individual word/event.

An LDA model serves as an appropriate generative model for this data. As in the case of the distributional model, each y_{ji} is sampled from one of K discrete probability distributions $\psi = \{\psi_1 \ldots \psi_k \ldots \psi_K\}$, with each ψ_k of length F. Which of these K components is chosen is determined by the value of the latent variable $x_{ji} \in \{1 \ldots K\}$ that corresponds to y_{ji}. Each x_{ji} is sampled from a mixing distribution π_j specific to word/event j. This is turn is drawn from a Dirichlet distribution with parameter α.

The likelihood of the data given ψ and α is

$$P(\mathrm{y}_{1:J}|\psi, \alpha) = \prod_{j=1}^{J} \int d\pi_j \prod_{i=1}^{n_j} P(\pi_j|\alpha) \sum_{k=1}^{K} P(y_{ji}|x_{ji} = k, \psi) P(x_{ji} = k|\pi_j). \quad \text{(A11)}$$

Using a vague inverse gamma prior on α, a symmetric Dirichlet prior with hyperparameter β on ψ, and a vague inverse gamma prior on β, the posterior distribution

$$P(\psi, \alpha, \beta|\mathrm{y}_{1:J}) \propto P(\mathrm{y}_{1:J}|\psi, \alpha) P(\psi|\beta) P(\alpha) P(\beta) \quad \text{(A12)}$$

may be approximated using an identical Gibbs sampling procedure as described for the distributional model. Likewise, the optimal value of K may be determined using an identical marginal-likelihood-based method.

Combined Model

The combined model is based upon a straightforward extension of the basic LDA model. A combined corpus consists of a set of documents, each being a bag of words from a vocabulary v. A subset v_f of these words is concrete words for which we have collected features. Each time one of these words occurs in a document, it is paired with one of its elementary feature (as explained in the Models section in the main text, this is done by drawing from the frequency distribution over elementary features that corresponds to that word). As such, the combined model's training set is a set of J texts, each comprising n_j words, and each word (if it is a word for which we have features) coupled to an elementary feature, that is,

$$(\mathrm{w} \cup \mathrm{y})_{1:J} = \{(w_{j1}, y_{j1}), (w_{j2}, y_{j2}) \ldots (w_{ji}, y_{ji}) \ldots (w_{jn_j}, y_{jn_j})\}_{j=1}^{j=J}, \quad \text{(A13)}$$

where if $w_{ji} \in v_f$, then $y_{ji} \in \{1 \ldots F\}$, otherwise $y_{ji} = \emptyset$.

A generative model for these data is an extension of the LDA model where each w_{ji} is drawn from one of K distributions $\varphi = \{\varphi_1, \varphi_2 \ldots \varphi_k \ldots \varphi_K\}$ and each y_{ji} (if $w_{ji} \in v_f$) is drawn from one of K distributions $\psi = \{\psi_1, \psi_2 \ldots \psi_k \ldots \psi_K\}$. In both cases, which of these components is chosen is determined by the value of $x_{ji} \in \{1 \ldots K\}$. In other words, the components φ and ψ are coupled such that if $x_{ji} = k$, then φ_k and ψ_k are chosen from which to sample w_{ji} and y_{ji}, respectively. As before, x_{ji} is sampled from a document-specific discrete distribution π_j, which in turn is sampled from a symmetric Dirichlet distribution with parameters α.

(*Appendixes continue*)

[A2] Although the harmonic mean method is not ideal as a means to calculate marginal likelihoods, we have found, from simulations with data sets where the correct number of components in the LDA model is known, that this method consistently provides reliable estimates of the optimal number of components.

The likelihood of the combined data given the parameters φ, ψ, and α is

$$P((w \cup y)_{1:J}|\varphi, \psi, \alpha) = \prod_{j=1}^{J} \int d\pi_j \prod_{i=1}^{n_j} P(\pi_j|\alpha) \sum_{k=1}^{K} P(w_{ji}|x_{ji} = k, \varphi)$$

$$P(y_{ji}|x_{ji} = k, \psi) P(x_{ji} = k|\pi_j). \quad \text{(A14)}$$

As before, we place a vague inverse gamma prior on α and symmetric Dirichlet priors, with hyperparameters β and γ, on φ and ψ. The hyperparameters β and γ can themselves have vague inverse gamma priors.

As with both the distributional and experiential models, posterior inference in the combined model is accomplished by Gibbs sampling. The posterior over the parameters and hyperparameters is

$$P(\varphi, \psi, \alpha, \beta, \gamma|(w \cup y)_{1:J}) \propto P((w \cup y)_{1:J}|\varphi, \psi, \alpha)$$

$$P(\varphi|\beta) P(\psi|\gamma) P(\alpha) P(\beta) P(\gamma). \quad \text{(A15)}$$

It can be approximated by iteratively sampling from $P(\mathbf{x}_{1:J}|\varphi, \psi, (\mathbf{w} \cup \mathbf{y})_{1:J})$, $P(\varphi|\mathbf{x}_{1:J}, \mathbf{w}_{1:J}, \beta)$, $P(\psi|\mathbf{x}_{1:J}, \mathbf{y}_{1:J}, \gamma)$, $P(\pi_{1:J}|\mathbf{x}_{1:J}, \alpha)$, $P(\alpha|\pi_{1:J})$, $P(\beta|\varphi_{1:J})$, and $P(\gamma|\psi_{1:J})$. These conditional posterior distributions have identical forms to those encountered already:

$$P(x_{1:J}|\varphi, \psi, \pi_{1:J}, (w \cup y)_{1:J}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j}$$

$$P(w_{ji}|x_{ji} = k, \varphi) P(y_{ji}|x_{ji} = k, \psi) P(x_{ji} = k|\pi_j) \quad \text{(A16)}$$

is a discrete distribution,

$$P(\varphi|x_{1:J}, w_{1:J}, \beta) \propto \prod_{k=1}^{K} \prod_{v=1}^{V} \varphi_{kv}^{r_{kv}+\beta-1}, \quad \text{(A17)}$$

where $r_{kv} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{I}(x_{ji} = k, w_{ji} = v)$, is a Dirichlet distribution,

$$P(\psi|x_{1:J}, y_{1:J}, \gamma) \propto \prod_{k=1}^{K} \prod_{f=1}^{F} \psi_{kf}^{u_{kf}+\gamma-1}, \quad \text{(A18)}$$

where $u_{kf} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbb{I}(x_{ji} = k, y_{ji} = f)$, is a Dirichlet distribution, and

$$P(\pi_{1:J}|x_{1:J}, \alpha) \propto \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{jk}^{s_{jk}+\alpha-1}, \quad \text{(A19)}$$

where $s_{jk} = \sum_{i=1}^{n_j} \mathbb{I}(x_{ji} = k)$, is also a Dirichlet distribution. The distributions $P(\alpha|\pi_{1:J})$, $P(\beta|\varphi_{1:J})$, and $P(\gamma|\psi_{1:J})$ are univariate log-concave and can be sampled by adaptive rejection sampling, as before. Likewise, convergence monitoring, lagged sampling, and optimization of K in the combined model proceed as with the cases of the distributional and experiential models.

Semantic Knowledge, Semantic Representation, Interword Similarity

As mentioned in the main text, we can regard the component distributions in each model as that model's semantic knowledge. For example, in the distributional model, each φ_k is a probability distribution over the V words in the vocabulary. In practice, most of the probability mass of φ_k will be confined to a small number of interrelated words, and as such, each φ_k will represent a discourse topic. By direct analogy, each ψ_k in the experiential model will represent a cluster of interrelated features. In the combined model, each pair φ_k and ψ_k will represent a coupling of a discourse topic and a related feature cluster. In Tables 2, 3, and 4 in the main text, we show examples of some of these components in the trained models. As Bayesian learning does not lead to a point estimate of each component, in the tables, we display the mean distribution from a posterior sample. For example, if $\{\tilde{\varphi}_k^0, \tilde{\varphi}_k^1, \ldots \tilde{\varphi}_k^N\}$ is a posterior sample of component φ_k, then the mean is $\frac{1}{N} \sum_{n=1}^{N} \tilde{\varphi}_k^n$.

If the component distributions are the model's semantic knowledge, any given word's semantic representation is its distribution over these component distributions. In the distributional model and the combined model, if the parameters are known, the distribution over the components corresponding to word w is simply

$$P(x = k|w, \varphi, \alpha) \propto P(w|x = k, \varphi) \int d\pi \; P(x = k|\pi) P(\pi|\alpha). \quad \text{(A20)}$$

As the parameters are unknown but our knowledge of them is described by a posterior sample, we average over these samples as in

$$\frac{1}{N} \sum_{n=1}^{N} P(x = k|w, \tilde{\varphi}^n, \tilde{\alpha}^n), \quad \text{(A21)}$$

to arrive at the semantic representation of w in these two models. In the experiential model, the situation is different. Each word w_j corresponds to a grouping of observations $\mathbf{y}_{j.}$ The probability distribution over the components corresponding to the group labeled by w_j is $\pi_{j.}$ As this distribution is not known, we infer its value by averaging over a posterior sample mean, that is, $\frac{1}{N} \sum_{n=1}^{N} \tilde{\pi}_j^n$.

Given that in each model, each word's semantic representation is a probability distribution, we can compare probability distributions using a suitable metric. A common choice for this is the symmetrized Kullback-Leibler divergence. For two probability discrete distributions p and q, each of length K, this is given by

$$d(p, q) = D_{KL}(p\|q) + D_{KL}(q\|p), \quad \text{(A22)}$$

where $D_{KL}(p\|q)$ is the Kullback-Leibler divergence

$$D_{KL}(p\|q) = \sum_{k=1}^{K} p_k \log \left(\frac{p_k}{q_k} \right). \quad \text{(A23)}$$

Appendix B

Bayesian Data Analysis

The data analyses we use are based on Bayesian methods rather than classical or sampling-based ones. These methods have now become part of the mainstream in statistics (see, e.g., Box & Tiao, 1973; A. Gelman, Carlin, Stern, & Rubin, 2003; Lee, 2004; Sivia & Skilling, 2006, for introductions) yet are not yet widely practiced in psychology and related disciplines. We cannot provide a thorough introduction to this topic, but, in what follows, we aim provide a brief description of the methods we employ, and we suggest interested readers consult a standard introduction, such as those just mentioned, for full details and background.

High Posterior Density Regions

An important concept on which we rely is that of high posterior density (HPD) regions. These can be seen as the Bayesian analogue of the classical confidence interval. The HPD describes regions of high probability for the values of a given parameter or set of parameters. As such, we should have high confidence that the value of the unknown parameter of interest lies within the HPD region. More formally, following Box and Tiao (1973), if $P(\theta|D)$ is the posterior over θ given data D, then the HPD region R with mass $(1 - \alpha)$ is such that $P(\theta \in R|D) = 1 - \alpha$, and for every $\theta_0 \in R$ and $\theta_1 \in R$, $P(\theta_0|D) \geq P(\theta_1|D)$.

Bayesian Analysis of Variance

The common one-way analysis of variance (ANOVA) is used to test whether $J > 2$ normally distributed random variables differ in their means. This procedure is normally described and interpreted from the perspective of classical or sampling-theory-based statistics. It can also be interpreted from the point of view of Bayesian statistics.

The ANOVA is appropriate for data that consists of J groups $\mathbf{x}_{1:J} = \mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_j \ldots \mathbf{x}_J$ where $\mathrm{x}_j = x_{j1}, x_{j2} \ldots x_{ji} \ldots x_{jn_j}$. For each group j, we assume that each element $x_{ji} \sim \mathcal{N}(\mu_j, \sigma^2)$. In other words, the n_j elements of each group j are assumed to be independently and identically distributed as a Gaussian random variable with mean μ_j and variance σ^2. Across the J groups, this variance σ^2 is assumed to be identical.

The probability distribution over the possible values of the parameters $\mu_1, \mu_2 \ldots \mu_J$ given the observed data $\mathbf{x}_{1:J}$ is attained by integrating over the possible values of the common variance parameter σ^2, that is,

$$P(\mu_1, \mu_2 \ldots \mu_J|\mathrm{x}_{1:J}) = \int d\sigma^2 \, P(\mu_1, \mu_2 \ldots \mu_J, \sigma^2|\mathrm{x}_{1:J}), \tag{B1}$$

where

$$P(\mu_1, \mu_2 \ldots \mu_J, \sigma^2|\mathrm{x}_{1:J}) \propto P(\sigma^2) \prod_{j=1}^{J} P(\mu_j) \prod_{i=1}^{n_j} P(x_{ji}|\mu_j, \sigma) \tag{B2}$$

is the full posterior distribution. The priors $P(\mu_1), P(\mu_2) \ldots P(\mu_J)$ and $P(\sigma^2)$ describe the prior knowledge about the values of the parameters. In the case of no prior knowledge, noninformative priors should be used. For each μ_j, a suitable noninformative prior is uniform. For σ^2, a noninformative prior is uniform on $\log(\sigma)$. Under these circumstances, it can be shown (see Box & Tiao, 1973, for full derivation) that the full posterior (Equation B2) is closed form, as is the marginal posterior (Equation B1). In particular, the latter is

$$P(\mu_1, \mu_2 \ldots \mu_J|\mathrm{x}_{1:J}) \propto \left[1 + \frac{1}{\nu} \sum_{j=1}^{J} n_j \frac{(\mu_j - \bar{\mathrm{x}}_j)^2}{S^2} \right]^{-\frac{1}{2}(\nu+J)}, \tag{B3}$$

where

$$\bar{\mathrm{x}}_j = \frac{1}{n_j} \sum_i x_{ji}, \quad \nu = \sum_j n_j - J, \quad S^2 = \frac{1}{\nu} \sum_{j,i} (x_{ji} - \bar{\mathrm{x}}_j)^2.$$

This is a J-dimensional t distribution, $\mathrm{t}_\nu(\hat{\mathrm{x}}, \Sigma)$, with location parameters $\hat{\mathrm{x}} = \bar{\mathrm{x}}_1 \ldots \bar{\mathrm{x}}_j \ldots \bar{\mathrm{x}}_J$, scale parameters

$$\Sigma = S^2 \begin{bmatrix} \frac{1}{n_1} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{n_J} \end{bmatrix},$$

and shape (i.e., degrees of freedom) ν.

Having inferred the posterior probability over the possible values of $\mu_1 \ldots \mu_J$, the question of primary interest now is whether $\mu_1 = \ldots = \mu_J$, that is, whether the means of the J groups are identical. To assess this, we introduce the variables $\varphi_1, \varphi_2 \ldots \varphi_{J-1}$, where $\varphi_j = \mu_j - \bar{\mu}$ and $\bar{\mu} = \frac{1}{j} \sum_j \mu_j$ It can be easily verified that $\mu_1 = \ldots = \mu_J$ if and only if $\varphi_1 = \ldots = \varphi_{J-1} = 0$. We can assess whether $\varphi_1 = \ldots = \varphi_{J-1} = 0$ by inferring the posterior $P(\varphi_1 \ldots \varphi_{J-1}|\mathbf{x}_{1:J})$ and determining whether the $(J - 1)$-dimensional zero vector $\mathbf{0}$ is in the HPD region. The posterior is

$$P(\varphi_1, \varphi_2 \ldots \varphi_{J-1}|\mathrm{x}_{1:J}) \propto \left[1 + \frac{1}{\nu} \sum_{j=1}^{J} n_j \frac{[\varphi_j - (\bar{\mathrm{x}}_j - \bar{\mathrm{x}})]^2}{S^2} \right]^{-\frac{1}{2}(\nu+J-1)}, \tag{B4}$$

where $\bar{\mathrm{x}}_j$, S^2, and ν are as above, $\bar{\mathrm{X}} = \frac{\sum_j n_j \bar{\mathrm{X}}_j}{\sum_j n_j}$, and $\varphi_j = \mu_j - \bar{\mu}$. It can be shown (Box & Tiao, 1973) that $\mathbf{0}$ is in the $(1 - \alpha)$ HPD region of this posterior if

(*Appendixes continue*)

$$\frac{\sum_{j=1}^{J} n_j(\bar{\mathrm{x}}_j - \bar{\mathrm{x}})^2}{(J-1)S^2} < \mathrm{F}(J-1, \nu, \alpha), \tag{B5}$$

where $\mathrm{F}(J-1, \nu, \alpha)$ is the value of an F variable with $(J-1, \nu)$ degrees of freedom above which is α of the probability mass.

Finally, to compare any pair of means, we can apply a linear transformation to the posterior in (Equation B3). In particular, to obtain the posterior over the difference between μ_i and μ_j, we define a linear operator A that is a row vector of zeros, with element i equal to 1 and element j equal to -1. The posterior distribution is simply

$$\mathrm{P}(\mu_i - \mu_j|\mathrm{x}_{1:J}) = \mathrm{t}_\nu(A\hat{\mathrm{x}}, A\Sigma A').$$

Bayesian Correlation Analysis

Pearson's product–moment correlation coefficient between two variables x and y is defined as

$$\rho(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}, \tag{B6}$$

where $\mathrm{Cov}(x, y)$ is the covariance of x and y, and σ_x and σ_y are their standard deviations. If x and y are normally distributed, given a sample $x_1, x_2 \ldots x_n$ and $y_1, y_2 \ldots y_n$, the posterior distribution over ρ is closely approximated by

$$\mathrm{P}(\rho|\{x_i, y_i\}_{i=1}^n) \propto \frac{(1-\rho^2)^{\frac{n-1}{2}}}{(1-\rho r)^{n-\frac{3}{2}}}, \tag{B7}$$

where r is the sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{B8}$$

with $\bar{x} = \sum_i x_i/n$ and $\bar{y} = \sum_i y_i/n$. This posterior is under the assumption of a uniform prior on ρ and standard noninformative priors on the means and variances of x and y (see Lee, 2004, for details).

Received May 31, 2007
Revision received March 25, 2009
Accepted March 27, 2009 ■