UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Evaluating the Contribution of Intra-Linguistic and Extra-Linguistic Data to the Structure of Human Semantic Representations

Permalink

Journal

ISSN

Authors
Andrews, Mark
Vigliocco, Gabriella
Vinson, David

Publication Date
2007

Peer reviewed

Evaluating the Contribution of Intra-Linguistic and Extra-Linguistic Data to the Structure of Human Semantic Representations

Mark Andrews (m.andrews@ucl.ac.uk)
Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
David Vinson (d.vinson@ucl.ac.uk)
Department of Psychology, University College London,
26 Bedford Way
London, WC1H 0AP
United Kingdom

Abstract

We describe Bayesian models that learn semantic representations from either extra-linguistic data or intra-linguistic data, or from both in combination. We evaluate the validity of these models using three human-based measures of semantic similarity. The results provide strong evidence for the hypothesis that human semantic representations are the product of the statistical combination of extra- and intra-linguistic sources of data.

Introduction

For the purposes of this paper, we use the term *semantic representation* to refer to a language user's mental or cognitive representation of the meaning of words. We informally define this as the knowledge that allows the language user to infer, amongst other things, which words are similar or identical in meaning, what are the semantic or ontological categories to which a word belongs, what (if anything) are the referents of a word. Our general aim is to consider how both *extra-linguistic* and *intra-linguistic* data can be used to acquire this knowledge. Extra-linguistic, or *attributional* data, is data that is derived from our perception and interaction with the physical world, and in particular, from the perceived physical attributes or properties associated with the referents of words[1]. In contrast, intra-linguistic, or *distributional* data, is derived from the statistical characteristics within a language itself, or how a given word is distributed across different spoken or written texts[2]

In previous literature, it has been repeatedly demonstrated that semantic representations can be learned from either attributional data alone, e.g. McRae, Sa, and Seidenberg (1997); Vigliocco, Vinson, Lewis, and Garrett (2004); McClelland and Rogers (2003), or distributional data alone, e.g. Lund and Burgess (1996); Landauer and Dumais (1997); Griffiths and Steyvers (2002). However, in previous work of our own (Andrews, Vigliocco, & Vinson, 2005), we described how, for the most part throughout this literature, the contribution of

any one of these data types had been considered independently and to the exclusion of the other. To address this concern, we considered the combined effects of both sources of data and introduced a probabilistic model that learns semantic representations on the basis of both attributional and distributional data simultaneously. We then compared this model with probabilistic models that learn semantic representations from each data source independently.

In our above mentioned work, we did not provide an analysis of how well the semantic representations learned by our model predict human data. The primary aim of this paper is to address this issue. For this purpose, we have also found it necessary to elaborate and extend upon the models that we previously used. As such, in what follows, we provide Bayesian models of semantic representations that are learned from either attributional data or distributional data, or from both in combination. We then evaluate the validity of these models using three human-based measures of semantic similarity: word-association norms, semantic-priming results from a lexical decision task, and interference patterns from a picture-word interference task.

Model Description

We provide Bayesian models that learn semantic representations from examples of attributional data, or from distributional data, or from both combined. The probabilistic models we employ for each of the various data types are described graphically in Figure 1. The attributional model (leftmost) describes any given word w_f as a probability distribution over a set of binary attributes, such that $\{\mathbf{y}_{m[f]}\colon 1 \leq m \leq M^{[f]}\}$ is a set of bit vectors, each being an instance of the referent of the word w_f. These probability distributions are compositions of a basic repertoire of latent distributions $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1 \ldots \boldsymbol{\psi}_{K_{\mathsf{Att}}}\}$ that intuitively correspond to clusters of interrelated attributes each describing basic characteristics of the attributional data. The distributional model (second left) describes texts as multinomial distribution over words, such that $\{w_{n[t]}\colon 1 \leq n \leq N^{[t]}\}$ is a sample of words from text t. These distributions are compositions of latent distributions $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1 \ldots \boldsymbol{\phi}_{K_{\mathsf{Dist}}}\}$ that intuitively correspond to discourse-topics in a corpus of text. The combined model (second right) describes texts as probability distributions over words, and words as distributions over attributes. These distribu-

---

[1] 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, etc

[2] 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 email message, an essay, a speech, etc.
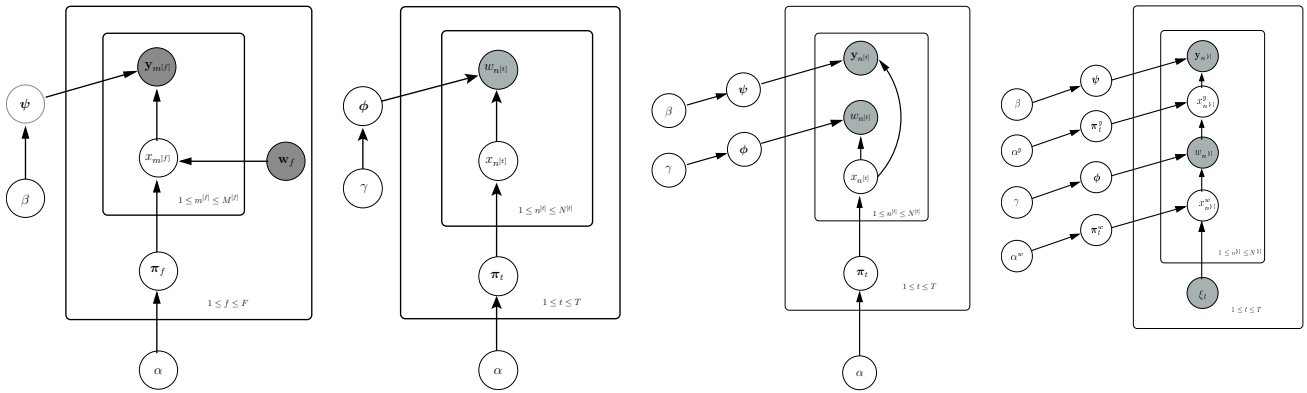
Figure 1: Bayesian Networks for the (from left) attributional model, distributional model, combined model, independent model. Each node corresponds to a variable, parameter or hyper-parameter, with observed variables being shaded. Note that in the rightmost diagram, the variable ξ_t is an indicator variable denoting the identity of the text at time t. All other variables are described in the main text.

tions are compositions of coupled latent distributions $\psi, \phi = \{\psi_1, \phi_1 \ldots \psi_{K_{\mathsf{Comb}}}, \phi_{K_{\mathsf{Comb}}}\}$, each intuitively corresponding to an attribute cluster coupled to a discourse-topic. Finally, our so-called independent model (rightmost) acts as a experimental control model to our combined model. Like the combined model, the independent model describes texts as probability distributions over words, and words as distributions over attributes. However, these distributions are de-coupled and independent.

Each of the four models can be seen as a hierarchical mixture model. Each observed data point is sampled from a single latent distribution that is indicated by an unobserved indicator variable (this is denoted by the indexed x variables in each diagram in Figure 1). These indicator variables are themselves sampled from multinomial distributions (denoted by the indexed π variables in the diagrams). We treat the latent distributions, as well as the multinomial distributions over the indicator variables, as the parameters of the model. The objective of learning is to infer the posterior distribution of these parameters. This was accomplished using the Markov Chain Monte Carlo method of Gibbs sampling. For this, conjugate prior distributions in the form of Dirichlet distributions for ψ and π, and Beta distributions for ψ were used. Each of these distributions had controlling hyper-parameters (the α, β and γ variables in the diagrams). Model selection by marginal likelihood optimization was used to find optimal values for these hyper-parameters, as well as for the total number of latent distributions in each model.

The data used for model training were as follows: Following common practice, we obtained the attributional data for 456 words by way of speaker-generated attribute norms collected in Vigliocco et al. (2004). The distributional data was 2245 texts[3] taken from the British National Corpus (BNC).

Semantic Representations

The latent distributions in each model intuitively correspond to that model's semantic knowledge. We can provide examples of this knowledge by drawing samples from the mean of the posterior distribution over the latent distributions. Examples for the cases of the attributional, distributional and combined models are shown in Table 1(a). From these examples, it can be seen that the latent-distributions are clusters of inter-related attributes (in the case of the attributional model), or words (in the distributional model), or both (in the combined model)[4]. Importantly, in the case of the combined model, attribute clusters align with discourse-topics that are consistent with the same general meaning.

Within each model, each word can be expressed as a distribution over that model's latent distributions. From this we can measure the correspondence between any pair of words in each model. In general, in a model whose unobserved indicator variable is denoted by x the correspondence between words w_i and w_j is given by $\mathrm{P}(w_j|w_i) = \sum_{\{x\}} \mathrm{P}(w_j|x)\mathrm{P}(x|w_i)$. In Table 1(b), using this formula, and averaging over samples from the posterior over the parameters, we provide examples of the near-neighbors of a set of example words according to each of our four models.

Model Evaluation

We evaluate each model by comparing its set of inter-word similarities with human-based measures of semantic similarity. There is, of course, no flawless means by which to measure human semantic representations or the inter-word similarities implied by them. In light of this, we have used a collection of methods that will hopefully lead to converging evidence. These are the Nelson word-association norms[5], and semantic-priming re-

---

[3]Each text was approximately 200-250 words in length. In total there were 7818 unique word types.

[4]We do not display examples from the independent model as these will, by design, be identical to the independent product of the attributional and distributional models' latent distributions.

[5]http://w3.usf.edu/FreeAssociation/

Table 1: Semantic Knowledge and Inter-word Similarities in the Models

(a) Examples of latent distributions learned by the attributional model (upper left), distributional model (lower left) and combined model (right). The latent distributions in the combined model are coupled distributions over both attributes and words.

mouth	transport	foot	food			read	fast	make	explode
tongue	vehicle	leg	oven			write	leg	construct	danger
taste	wheel	ball	heat			story	move	tool	destroy
food	fly	force	cook		Attributes	pencil	exercise	building	action
eat	drive	arm	eat			communicate	walk	build	war
throat	passenger	pain	prepare			question	destination	wood	kill
taste-bud	seat	game	consume			pen	foot	house	fire
sense	motor	win	mouth			knowledge	speed	fix	demolish
hospital	party	market	rate			book	run	build	killed
death	election	price	cut			film	team	house	fire
died	political	prices	interest			draw	race	repair	attack
operation	national	sales	rates		Words	page	next	fix	war
treatment	opposition	stock	figures			star	winner	building	security
injuries	elections	sold	economic			written	grand	equipment	women
medical	held	sale	trade			series	field	construction	shot
cancer	seats	buy	industry			television	running	steel	bomb

(b) Examples of the near neighbors of a set of five words (boldface) according to the attributional, distributional, combined and independent models. The five words were chosen so as to highlight the differences between the four models.

	ankle	**exchange**	**punch**	**knife**	**threat**		**ankle**	**exchange**	**punch**	**knife**	**threat**
Attributional	elbow	pay	punch	knife	threat	**Combined**	knee	buy	hit	knife	threat
	ankle	buy	chin	axe	warn		ankle	pay	punch	kill	attack
	knee	exchange	slap	saw	threaten		elbow	sell	down	blood	threaten
	toe	sell	hit	scissors	argue		toe	exchange	slap	assault	attacks
	cut	trade	pound	hatchet	bark		cut	sales	knock	axe	killed
	ache	acquire	bark	chisel	growl		leg	selling	chin	dead	armed
	shoulder	donate	shoulder	razor	hit		injury	trade	injury	dagger	murder
	thumb	loan	knock	chop	challenge		ran	money	pound	arrest	military
	leg	borrow	face	dagger	fear		broken	billion	hitting	charges	kill
	twist	accept	neck	drill	kill		walked	markets	fell	murder	killings
	walk	donation	break	sword	murder		injured	loan	blood	illegal	violence
Distributional	injury	market	fight	court	threat	**Independent**	injury	market	fight	court	northern
	fit	stock	world	blood	terrorist		fit	stock	world	saw	threat
	squad	price	title	case	violence		squad	price	title	case	violence
	coach	prices	punch	knife	northern		side	prices	punch	knife	community
	knee	sales	heavyweight	murder	community		cup	sales	chin	man	fear
	fitness	financial	boxing	alleged	province		ankle	exchange	slap	then	violence
	ankle	customer	knock	prison	loyalist		knee	sold	knock	heard	warn
	side	exchange	manager	trial	army		elbow	sale	hit	axe	threaten
	cup	business	battle	appeal	forces		draw	sell	manager	door	argue
	season	markets	lost	judge	fear		break	business	boxing	razor	clash

sults from lexical decision tasks and interference patterns from picture-word interference task, both obtained from Vigliocco et al. (2004). These three methods were chosen so as to provide complementary measures of human semantic representations. In particular, it is arguable that word-association norms are primarily a measure of *syntagmatic* relationships. While, by contrast, behavioral measures like semantic priming and picture-word interference data are arguably based more upon *paradigmatic* rather than syntagmatic relationships. Syntagmatic relations are said to hold between two words that commonly co-occur within a sentence, often when both are of different parts of speech. Examples are easy to come by: *sit-chair, drink-wine rain-wet.* On the other hand, paradigmatic relations hold between words that have similar roles with respect to the other words, or syntactic structures, within sentences. Examples of paradigmatically related pairs would include *eat-drink, sit-stand, wet-dry.* By using multiple human based measures of semantic relationships that are either more influenced by syntagmatic over paradigmatic relationships, or vice versa, we can hopefully provide a more general or unbiased picture of human semantic representation against which to compare our models.

For the purposes of comparison we also include in our analysis what we refer to as a unigram model. The un-igram probability of a word from the vocabulary in the corpus is simply the relative frequency of occurrence of that word. We can use this probability distribution as a null model of the extent to which word v_i predicts v_j as it specifies that $\forall i,\ \mathrm{P}(v_j|v_i) \triangleq \mathrm{P}(v_j)$. In effect, this means that the highest probability words predicted by any words will be simply the highest frequency words.

Note that the attributional model contains only a subset of the words (i.e. concrete words for which attributes exist) that also occur in the distributional, combined, independent and unigram models. Accordingly, we divided our analysis in such a way that we compare all the models with one another using the subset of words that they all share, and we compare the larger vocabulary models with one another using all available words.

Bayes Factors Based Hypothesis Testing

In keeping with the Bayesian nature of the models, we explore the use of Bayesian hypothesis tests rather than the more commonly used classical, or sampling-theory, approaches. The relative merits of these two approaches is subject to (sometimes intemperate) debate, but it is beyond the scope of this paper to either review or contribute to this debate. Suffice it to say that the Bayesian approaches are ideally suited to the analysis we wish to pursue.

$\log \lambda$	Evidence for $\mathcal{M}_1$
$\log \lambda < 0$	Negative
$0 \leq \log \lambda < 1$	Weak
$1 \leq \log \lambda < 2.5$	Positive
$2.5 \leq \log \lambda < 5$	Strong
$\log \lambda \geq 5$	Very strong

Table 2: Interpretation of λ in the Bayes Factor Test.

The analysis we will pursue is often referred to as the Bayes factor test. Given any test data-set $\mathcal{D}_{\text{test}}$ and any two alternative models $\mathcal{M}_0$ and $\mathcal{M}_1$ (parameterized by θ_0 and θ_1, respectively) the Bayes factor for $\mathcal{M}_1$ relative to $\mathcal{M}_0$ is given by

$$\lambda = \frac{\mathrm{P}(\mathcal{D}_{\text{test}}|\mathcal{M}_1)}{\mathrm{P}(\mathcal{D}_{\text{test}}|\mathcal{M}_0)} = \frac{\int d\theta_1 \ \mathrm{P}(\mathcal{D}_{\text{test}}|\theta_1)\mathrm{P}(\theta_1|\mathcal{M}_1)}{\int d\theta_0 \ \mathrm{P}(\mathcal{D}_{\text{test}}|\theta_0)\mathrm{P}(\theta_0|\mathcal{M}_0)}. \quad (1)$$

The term λ is a measure of evidence for the superiority of $\mathcal{M}_1$ over $\mathcal{M}_0$. Jeffreys (1961) provides a scale of interpretation for λ as shown in Table 2. Clearly, this test is easily applied to our model comparisons, whereby we integrate over the posterior probabilities of the parameters for each model, evaluating the probability of data set for each parameter value. In our case, however, we must replace the integral with a sum over samples from the posteriors.

Word Association Norms The Nelson word association norm data-set is a collection of the close word-associates of 5019 English words. These have been collected from human participants under controlled circumstances, and each word associate is assigned a probability indicating the relative frequency of its being paired with the target word. Of the 5019 words, a subset of 2824 also occurred in our text-corpus vocabulary.

The word association norms data-set can be re-described as a (sparse) $V \times V$ matrix $\mathbf{W}$, where V is the number of unique words in our text-corpus (i.e. 7818), and W_{ij} is the probability that word w_j is associated with target word w_i. If either w_i or w_j do not occur in the association norms set, then W_{ij} is set to 0. From $\mathbf{W}$ we can define $\mathbf{W}'$ as the $V \times V$ matrix with $W'_{ij} = 1$ if $W_{ij} > 0$ and zero otherwise.

The likelihood of the Nelson norms for any one of our models, with specific parameter values denoted by θ, is given by

$$\mathrm{P}(\mathbf{W}'|\theta) = \prod_{\{i,j\}} \mathrm{P}(v_i|v_j, \theta)^{W'_{ij}}, \quad (2)$$

By sampling from each model's posterior distributions over its parameters, and averaging over these samples, we can thus calculate how how probable the norms' word-pairs are according to each model. In other words, given the set of word-associate pairs in the Nelson norms, how likely are these data according to each of four models' predictions of word relationships.

The log of these predictive likelihoods for each model are shown in the left column of Figure 2. The upper left shows the results for all available words. The lower left shows the results for the subset of words that all models share (see above note). Note that the differences in the log of these probabilities is equivalent to the log of the ratio of the probabilities. Hence to evaluate $\log \lambda$ for any pair of models, simply subtract the log probability of one from the other. Upon inspection of the graphs, it is evident that there is very strong evidence (according to the Jeffreys definition of the term) for the superiority of the combined model's predictiveness of the word association norms. In particular, the order of performance of the models (proceeding from best to worst) is the combined model, independent model, distributional model, unigram model (when all words are used) and the combined model, attributional model, independent model, distributional model and unigram model (when the subset of words is used). In both cases, there is very strong evidence for these orderings[6].

Lexical Decision Based Priming Semantic priming using a lexical decision task is one of the most commonly used behavioral measures of the semantic relationship between pairs of words. In the study carried out by Vigliocco et al. (2004), priming data for a set of prime-target word-pairs, all of which occur in our data-sets, were collected. The speed of response to the target word, given the presence of the prime, is compared to the speed of response of the target word in the presence of an obviously unrelated baseline word (matched to the prime word on salient characteristics such as length, frequency, etc.). This allows each prime-target pair to be represented in terms of the relative speed up of response to the target in the presence of the prime.

In order to assess how well each model predicts this data, we used a separate Bayesian linear regression model for each model. In each case, we regressed relative speed-up (msec) for target-word v_i given prime-word v_j on the log of $\mathrm{P}(v_i|v_j)$ derived from each model. The quantity $\mathrm{P}(v_i|v_j)$ was obtained by averaging over the posterior of the parameters in each model. The outcome of the Bayesian regression is a posterior distribution over the parameters of the linear-Gaussian regression model. From this, we can calculate the marginal likelihood of the priming-data for the case of each model, and compare these in a Bayes factor test as in Equation 1. We have plotted the log of these marginal likelihoods in the upper right sub-figure of Figure 2. Note that, as before, the differences between any pair of log probabilities will be equal to the log of the ratio of these probabilities. As can be seen in this figure, there is very strong evidence (using the Jeffreys' definition) in favor of the superiority of the combined model as a model of P. The exact ordering of the models' performance is (from best to worse): combined model, distributional model, independent model, attributional model and unigram model. This ordering is also strongly supported by the results of the Bayes factor test.

---

[6]For the purposes of comparison, we performed an analysis of these data using non-parametric statistics and sampling based null hypothesis tests and the relative ordering of the models' performances was identical.
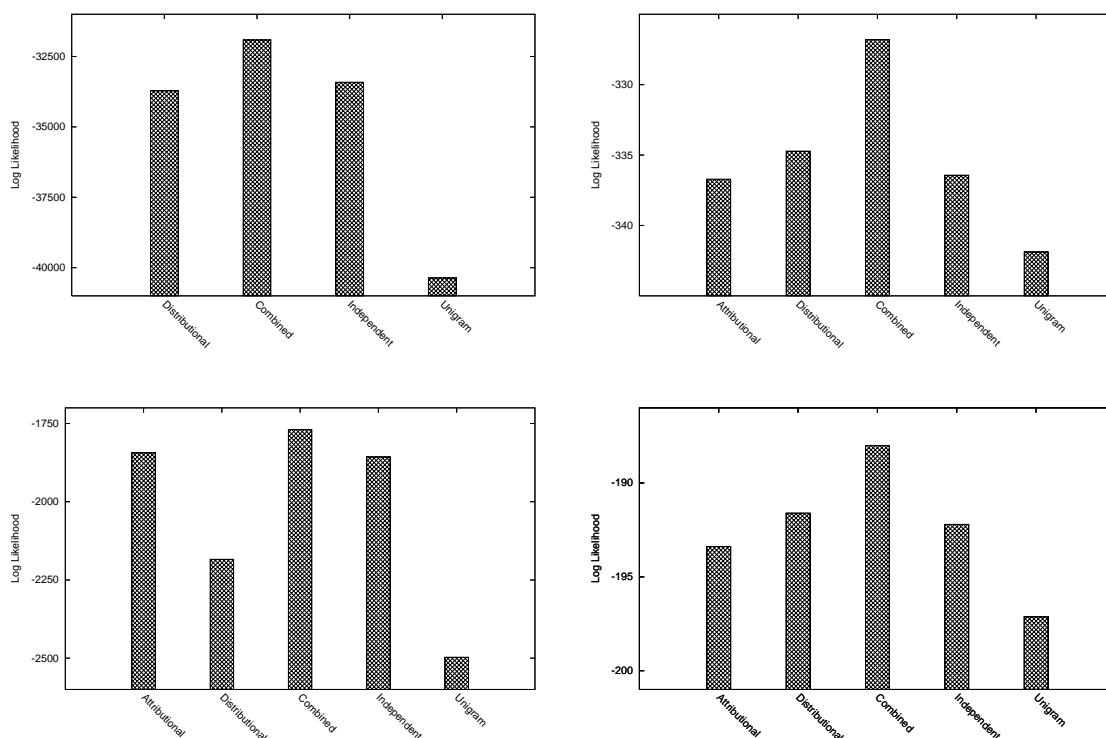
Figure 2: Log Likelihoods for three human-based data-sets for each of the models under investigation. The left-most column provides results for the word-association norm data. The right-most column provides the results for priming and picture-word interference data. See text for details.

For the purpose of comparison, it is also useful to consider the results from a standard, or non-Bayesian, linear regression test. Commonly used measures from this type of analysis include a measure of the strength of the linear relationship between the variables R, the amount of variance in the dependent variable accounted for by the independent variable R^2, and the p-value significance of these statistics p. These are as follows:

Model	R	R^2	p-value
Attributional	.31	.09	.006
Distributional	.29	.086	.008
Combined	.39	.16	.0002
Independent	.22	.05	.04
Unigram	.078	.006	.49

Picture Word Interference In a picture-word interference task, naming latencies of drawings of objects (or actions/events) are recorded. When these pictures are presented simultaneously with a word, and if that word is semantically related to the picture, naming latencies increase. This increase resembles the Stoop phenomenon whereby the semantically related word interferes with the activation of the picture's name. In Vigliocco et al. (2004), picture-word interference data was collected for a set of word pairs (all of which occur in all our models). If the picture depicts word v_i and the distractor word is v_j, the slow-up for naming the picture as v_i (relative to

a baseline distractor) can be used as a measure of the semantic similarity between v_i and v_j.

As in the case of the priming data, we used separate Bayesian linear regression models, regressing naming latency against the log of $P(v_i|v_j)$ in each model (averaging over parameters). From this, we can calculate the marginal likelihood of the picture-word interference data according to each model. The marginal likelihoods can be compared in a Bayes factor test, as before. We have plotted the log of these marginal likelihoods in the lower right sub-figure of Figure 2. The relative pattern of results is almost identical to that seen in the priming data case. The combined model shows the strongest predictive power with the ordering from strongest to weakest model is combined model, distributional model, independent model, attributional and unigram models. These results are strongly supported by the Bayes factor test. As with the case of the priming data, for the purposes of comparison, we can mention standard measures from non-Bayesian regression analysis, i.e. R, R^2 and p:

Model	R	R^2	p-value
Attributional	$-.24$	.06	.09
Distributional	$-.35$	.12	.01
Combined	$-.38$	.14	.009
Independent	$-.26$	.06	.08
Unigram	$-.19$	.03	.19

Discussion

The general aim in this paper has been to consider how semantic representations are acquired. To answer this we have identified two major types of data from which semantic information can be attained. We have referred to these as *attributional* and *distributional* data types. These represent data types that are, respectively, extra-linguistic and intra-linguistic in their origin. Of particular concern to us has been the question of how these two distinct data types can be combined to learn coherent semantic representations. We have provided a model of the semantic representations that are learned from attributional and distributional data taken in combination, and compared this to the representations learned from either source taken independently. Our specific aim has then been to evaluate these models against human-based measures of semantic representations.

Although the relative performance of each model to predict the human data is not identical across the three different data-sets there are obvious and compelling general trends. For example, and unsurprisingly, all four of the attributional, distributional, combined and independent models outperform the null model on all data-sets. While superior performance against a null-model is not surprising, it does serve as a worthwhile sanity check, effectively corroborating the impression given by Table 1 that each of these models is providing (at the very least) a modest description of the meaning of words.

If the unigram model represents a lower-bound on the models' predictive performances, then it appears as if the combined model represents an upper-bound. The combined model outperforms all other models consistently across all three sets of human-based measures. This corroborates the impression given by Table 1(b) that the combined model provides a more comprehensive and valid account of the meanings of words than do either the attributional, distributional or independent models. As such, we can take this as direct evidence in favor of our primary hypothesis that human semantic representations are the product of the statistical combination, and not simply the sum or average, of attributional and distributional data-types.

Conclusion

The results imply a certain picture of how word-meanings are learned. This can be described by reference to following scenario: A child learning his or her native language will regularly experience words referring to, for example, everyday objects in the context of one or more of their referents. On the other hand, the words that the child is learning are not necessarily heard in isolation, but rather will regularly occur in the context of meaningful sentences. From this, the data from which the child can learn word-meanings occur in two forms simultaneously: There is the set of attributes associated with a given word, and the set of textual contexts in which that word occurs. While it has been repeatedly shown in previous literature that either one of these sources can provide information from which word-meanings can be learned, learning from both data-types in combination

would allow the correspondences between the two data-types to be apparent, and to be exploited. For example, if the child learned that the word *cat* refers to creatures with claws and whiskers and tails, etc. and that it also co-occurs with terms like *dog*, *pet*, *owner*, etc., it may also infer that creatures with claws and whiskers and tails, etc., are conceptually related to the words *dog*, *pet*, *owner*, etc. From this, we can see that while using either extra-linguistic or intra-linguistic data can allow semantic representations to be learned by discovering the correlations *within* that specific data-type, using the combination of both allows the discovery of correlations both *within* and *between* these data-types.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2005). The role of attributional and distributional information in semantic representation. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society*.

Griffiths, T., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive science society*.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Clarendon Press.

Landauer, T., & Dumais, S. (1997). A solutions to Plato's problem: The Latent Semantic Analyis theory of acquistion, induction and representation of knowledge. *Psychological Review*, *104*, 211-240.

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

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

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

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.