Integrating Attributional and Distributional Information in a Probabilistic Model of Meaning Representation

Mark Andrews (mark@gatsby.ucl.ac.uk)
Gatsby Computational Neuroscience Unit, University College London
London, WC1N 3AR
Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
David Vinson (d.vinson@ucl.ac.uk)
Department of Psychology, University College London
London, WC1E 6BT

Abstract

In this paper we present models of how meaning is represented in the brain/mind, based upon the assumption that children develop meaning representations for words using two main sources of information: information derived from their concrete experience with objects and events in the world (which we refer to as attributional information) and information implicitly derived from exposure to language (which we refer to as distributional information). In the first part of the paper we present a model developed using self-organising maps (SOMs) starting from speaker-generated features (properties that speakers considered to be important in defining and describing the meaning of a word). This model captures meaning similarity between words based solely upon attributional information and has been shown to be successful in predicting a number of behavioural semantic effects. In the second part of the paper, we present a probabilistic model that goes beyond attributional information alone, integrating this information with distributional information derived from text corpora. The ability of this integrated model to learn semantic relationships is demonstrated with reference to comparable probabilistic models that use only attributional or distributional information.

1 Introduction

Successful communication implies exchanging information about concrete objects and events but also exchanging abstract ideas, for example about politics and religion. Such exchange of information and ideas requires that speakers convey their intended meanings through language, and that hearers understand them. The work we describe here addresses the question of how meaning is represented in the brain/mind and how it can be learnt from the environment.

Children come to the task of learning the meanings of words already equipped with a highly sophisticated repertoire of conceptual knowledge (i.e., mental representations for things, events, etc.) that sub-serves cognition in general. This repertoire, possibly grounded in innate biases

(Bloom 2002, Gleitman, Cassidy, Nappa, Papafragou & Trueswell 2005), develops by virtue of active interaction with the environment. Not surprisingly, words that refer to concrete entities and actions are among the first words being learned (Gleitman et al. 2005) as they can be grounded in perception and action in the environment.

Under the assumption that the meanings of words referring to concrete experience are grounded in our concrete interactions with the environment, the semantic representations for these words must be linked to perception and action. That is, the representations of these words must include **attributional information**: physical, emotional (non-linguistic) attributes associated with the real-world referents of words. More specifically, for example, when we hear words that refer to motion, we should retrieve attributes associated to movement and therefore engage some of the cognitive and neural processes that are also engaged in actual movement. There is important evidence that indicates that this is the case. Three recent imaging studies have shown activations in primary motor and/or premotor cortex when participants are presented with words (or sentences) that refer to movement (Hauk, Johnsrude & Pulvermüller 2004, Tettamanti, Buccino, Saccuman, Gallese, Danna, Scifo, Fazio, Rizzolatti, Cappa & Perani 2005, Vigliocco, Warren, Arciuli, Siri, Scott & Wise 2005). In the Vigliocco et al. (2005) PET study, Italian speakers were asked to simply listen to lists of words referring to motion (e.g., jump, skip) and lists of words referring to sensation (e.g., see, touch). Activation in primary motor cortex (and premotor cortex) was found when words referring to motion were contrasted to words referring to sensation. This finding suggests that in the most natural of all linguistic tasks (passive listening) we cannot help but engage knowledge that is not linguistic and in this case, used in the control and implementation of action.

However, learning words' meanings must also take into account information provided in linguistic input. In particular, words that are similar in meaning tend to behave similarly in terms of their distribution across different texts[1], thus information about the meaning of a word can be inferred from information about the contexts in which it occurs. This **distributional information** has long been recognised as playing a role in the learning of word meaning. For example, Firth (1957) has memorably summarised the importance of this information by the phrase "you shall know a word by the company it keeps". In other words, the type of linguistic contexts in which a word occurs can provide clues as to what that word might mean.

Within cognitive psychology and cognitive neuroscience, computational models have been developed either on the basis of attributional information, or on distributional information that can be extracted from text, but not both. In particular, in order to develop computational models of semantic representation that provide us with information about the semantic similarity among words and that take into account different types of experiential properties of concepts (motor, perceptual, emotional etc.), a number of researchers have used "speaker-generated features", obtained by asking participants to make decisions or to produce properties that they believe to be important in defining and describing given concepts (Farah & McClelland 1991, Hinton & Shallice 1991, McRae, de Sa & Seidenberg 1997). These features are then used as input to connectionist-style models, e.g. (McRae et al. 1997), or other models. In Section 2, we present a computational model of semantic representation developed from speaker-generated features and implemented using self-organising maps (SOMs), (Kohonen 1997). These models have been shown to provide a good fit to a num-

[1] Henceforth, we will use the term *text* in a technical sense to refer to any linguistic or textual utterance. While this could be an entire article, book, or transcribed conversation, we will usually use the term to refer to paragraphs, sentences, or strings of consecutive words in written documents or transcribed speech.

ber of behavioural findings concerning semantic effects in word recognition (semantic priming) and production (as in the picture-word interference paradigm). On the other hand, a number of researchers have developed models of semantic representation based on distributional information alone, e.g. (Landauer & Dumais 1997, Burgess & Lund 1997, Griffiths & Steyvers 2003). In Section 3, we present a probabilistic model that integrates these two types of information. We will describe how this model integrates these two sources of information to learn different semantic representations in comparison to models based on either attributional and distributional information alone.

2 The Featural and Unitary Semantic Space (FUSS) model

Here we describe FUSS, a model of semantic representation built from attributional information, in particular, speaker-generated features (see Vigliocco, Vinson, Lewis & Garrett (2004).

2.1 Collection of attributional information underlying FUSS

Speaker-generated features (following McRae et al. (1997)) were generated by native English speakers for a variety of words belonging to various semantic fields (not only concrete objects but also nouns and verbs referring to actions and events). 280 native speakers of English were asked to list properties that they considered to be important in defining and describing the meaning of a word, generating features for a total of 456 concepts (methods and complete materials appear in Vinson & Vigliocco (2002)). For each word, 20 participants generated features, and feature vectors were prepared by combining the features across participants for each word. Features were assigned weights according to the number of participants who listed a particular feature for a given word, thus providing us with a measure of featural salience across words (see Vinson, Vigliocco, Cappa & Siri (2003) for analyses and discussion of how features related to different modalities of experience are distributed across semantic fields).

2.2 Development of semantic similarity measures within FUSS

We used the speaker-generated features to develop a lexico-semantic similarity space, using self-organising maps (SOMs) (Kohonen 1997) to reduce the dimensionality of the featural representations on the basis of statistically relevant patterns of similarity which need not be specified in advance, but are nonetheless reflected in the emergent structure of the output space. The input consisted of the feature vectors and the output was a 25x40 bi-dimensional space (see Vinson & Vigliocco (2002) for complete details). By the end of training, the output layer had become topographically organised on the basis of the regularities of the featural input. For each map, each unit best responding to a given input vector was assigned a label corresponding to each word; this layer can be thought of as an expression of lexico-semantic representation (and thus would be linked to other linguistic information such as its phonology and orthography). 100 such maps were created, each from random input. We operationalised semantic similarity between two words in FUSS as the average Euclidean distance (across all 100 maps) between the two labelled units corresponding to those words.

2.3 Predicting behavioural performance using FUSS measures

We tested the validity of FUSS semantic similarity measures by assessing the extent to which they predict performance on a variety of online behavioural tasks sensitive to semantic similarity, and whether this similarity model exhibits the same quality of performance for the domains of objects and actions. More specifically, we assessed whether fine-grained semantic similarity effects could be observed in language production using picture naming tasks, including experimentally-induced semantic substitution errors, picture-word interference, and blocked cyclic naming; and in comprehension using primed lexical decision (for complete details see (Vigliocco et al. 2004, Vigliocco, Vinson, Damian & Levelt 2002). In all these tasks we found similar patterns of results: FUSS distances predicted fine-grained semantic similarity effects, and in all these tasks the performance of FUSS was comparable for object and action concepts.

For semantic substitution errors (Vigliocco et al. (2004), Experiments 1 and 2), FUSS distances significantly predicted the occurrence of errors beyond the contributions of form similarity between target and error words and of visual similarity between target and error pictures. In picture-word interference (Vigliocco et al. (2004), Experiments 3 and 4), FUSS measures also predicted fine-grained performance: the smaller the FUSS distance between the target and distracter words, the greater the semantic interference effect upon naming the target picture. In blocked cyclic naming (Vigliocco et al. 2002), the degree of FUSS similarity among items in a block (semantically very similar, moderately similar, dissimilar) also modulated the picture naming latencies. Finally, in primed lexical decision (Vigliocco et al. (2002), Experiments 5 and 6), FUSS distances predicted the extent to which a briefly presented word would facilitate lexical decision upon the target: greater proximity in FUSS led to faster response latencies in the decision task. These results provide converging evidence that semantic similarity measures derived from attributional information can offer valid insight into semantic representations, and that such information can be used as an approximation to semantic representations for different domains of knowledge (nouns referring to concrete objects; verbs referring to actions/events) without a cost to performance.

2.4 Limitations of FUSS

Although FUSS provides a plausible model of semantic representation it has two main limitations. First, it is unclear how such a model can be developed starting from speaker-generated features for words beyond words referring to concrete objects and events (a general problem for all approaches that take “embodiment” as a basic assumption, see e.g., Barsalou, Simmons, Barley & Wilson (2003)). Second, as we have argued in the introduction, it is intuitively obvious that children learn words’ meanings not just by associating words to attributes of referents in the world, but also via the distributional information provided in the linguistic input they receive. Thus, any model that only considers one type of information and not the other is necessarily limited. Below we show how these limitations can be overcome by developing probabilistic models based on both attributional and distributional information.

3 Integrating Attribution and Distributional Information

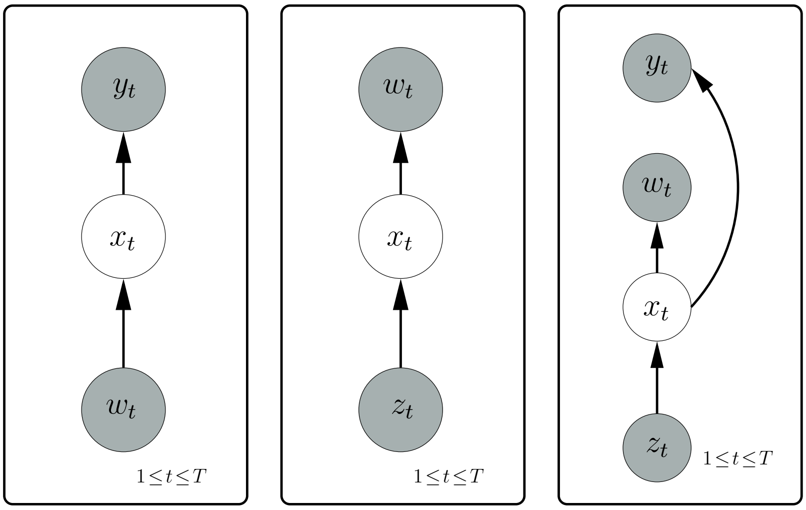
In this part of the paper, we discuss models that integrate attributional and distributional information to form semantic representations. The models we describe here are probabilistic generative models. They describe semantic representation in terms of explicit probabilistic relationships and variables. The nature of these relationships and variables are inferred or learned from data. These types of models seem well suited to integrating attributional and distributional information. Moreover, many of the previous models of semantic representation, like mentioned above, can be described as falling within this general class. For example, the FUSS model represents vectors of attributes in terms of points in a low-dimensional latent space, with the mapping to this latent-space preserving topological relationships between the data-points. While in FUSS, the mapping between an attribute vector and its latent representation is deterministic, more generally this mapping can be probabilistic, with attribute vectors now corresponding to probability distributions over latent space. Taking an explicitly probabilistic perspective introduces the language of probability theory which can often facilitate the analysis of learning and representation in these systems.

We describe three models that learn their semantic representations from attributional information, distributional information and both types of information, respectively. Attributional information is modelled by assuming that concrete terms are defined in terms of a distribution over latent attribute classes, and that these attribute classes are themselves probability distributions over binary properties or features. Distributional information is modelled by assuming that texts are multinomial distributions over latent semantic classes or topics, and that these latent topics are themselves multinomial distributions over words. In such a model, the semantic representation of a given word is defined in terms of its posterior probability distribution over the latent classes. Combining the two sources of information is straightforward. The data we observe consists of sets of words, and associated with each word is a distribution over binary non-linguistic attributes. As will be clarified, we can assume that a distribution over latent variables accounts for both the distribution of words in a text and the distribution of binary features associated with a given word. The semantic representation of a word is defined in terms of its posterior distribution over these latent topics. These semantic representations will be constrained to account for both the distributions of words in texts, and the distributions of binary attributes associated with given words.

3.1 Formal Specifications

We can make the preceding description more formal as follows. The observable data that we are modelling consist of both texts and the attributes associated with words. As mentioned, texts take the form of paragraphs, sentences and strings of consecutive words in a natural language corpus. If there are J texts in a corpus, we can label them as $\{z_1, z_2 \dots z_j \dots z_J\}$. The texts are, in general, of variable length, with text z_j of length T_j. Each text comprises words drawn from the vocabulary of word-types $\mathcal{V} = \{v_1, v_2 \dots v_k \dots v_K\}$. A subset of these words are *concrete words*, or words that refer objects, events, actions, etc., in the world. For each word in this set, we have a probability distribution over a set of L binary features $\mathcal{F} = \{f_1, f_2 \dots f_l \dots f_L\}$. Each feature is a property or attribute that could be associated with the physical referent of the concrete words.

The graphical models (or Bayesian Networks) for our three probabilistic models are shown in Figure 1. Taken together, we have the observable variables w_t, y_t and z_t that represent, respectively, the words, features and text occurring at a time t. In addition, we introduce the latent-variable



(a) Attributional Model (b) Distributional Model (c) Combined Model

Figure 1: The generative models that utilise a) attributional, b) distributional and c) combined information sources.

$x_t \in \{\xi_1, \xi_2 \ldots \xi_m \ldots \xi_M\}$. As a latent, or hidden, variable the value of x_t is unobserved. We see that in the attributional model, the binary attribute vector y_t is conditioned upon x_t, while x_t is conditioned upon the word label w_t. In the distributional model, we have the observable words conditioned upon the latent-variable x_t, which is then conditioned upon the text z_t. In the combined model, both the words w_t and binary attribute vector y_t are conditioned upon the latent-variable x_t, with x_t conditioned upon the text z_t.

The parametric forms of these conditional distributions is as follows. In the attributional model, the probability of the observed binary attributes y_t, conditioned upon w_t can be written

$$\mathrm{P}\big(y_t \big| w_t, \theta\big) = \sum_{m=1}^{M} \pi_m^{[w_t]} \prod_{l=1}^{L} p_{ml}^{(1-y_l^t)} (1 - p_{ml})^{(1-y_l^t)}, \tag{1}$$

where p_{ml} denotes the probability that y_l^t, takes the value 1 given that $x_t = \xi_m$, while $\pi_m^{[w_t]}$ denotes the probability that x_t takes the value ξ_m given the value of w_t. As such, the attributional model is a mixture of M multivariate Bernoulli distributions. In the distributional model, the probability of the observed word w_t, conditional upon observing text z_t can be written

$$\mathrm{P}\big(w_t \big| z_t, \theta\big) = \sum_{m}^{M} \pi_m^{[z_t]} \prod_{k=1}^{K} q_{mk}^{\mathbb{I}(w_t = v_k)}. \tag{2}$$

where q_{mk} denotes the probability of observing that $w_t = v_k$ given that $x_t = \xi_m$, the variable $\pi_m^{[z_t]}$ denotes the probability that $x_t = \xi_m$ upon observing that $z_t = j$ and $\mathbb{I}(a)$ is an indicator function, taking the value 1 if its argument a is true, and 0 otherwise. The distributional model is, as such, a mixture of multinomial distributions. In the combined model both the word label w_t and attribute vector y_t are conditioned upon the latent variable x_t, which is conditioned upon the text z_t. Integrating over the values of x_t, the conditional likelihood of the observables is

$$\mathrm{P}\big(y_t, w_t \big| z_t, \theta\big) = \sum_{m}^{M} \pi_m^{[z_t]} \prod_{k=1}^{K} q_{mk}^{\mathbb{I}(w_t = v_k)} \prod_{l=1}^{L} p_{ml}^{(1-y_l^t)} (1-p_{ml})^{(1-y_l^t)}. \tag{3}$$

As such, the combined model is a mixture of both multivariate Bernoulli distributions and multinomial distributions.

3.1.1 Model Learning

Given a set of training data $\mathcal{D}$ consisting of both texts and attributes associated with context-words, for each model above we would ideally wish to estimate $\mathrm{P}\big(\theta\big|\mathcal{D}\big)$, or the posterior probability of the parameters given the data. For present purposes, however, we will approximate $\mathrm{P}\big(\theta\big|\mathcal{D}\big)$ by its modal point θ_{mp}, which assuming a prior distribution over the parameters, is given by its maximum likelihood estimate θ_{mle}. The standard procedure for maximum-likelihood (or maximum posteriori) estimation in latent-variable models is Expectation-Maximisation (EM). This consists of iteratively computing a lower-bound on the likelihood of the data $\mathrm{P}\big(\mathcal{D}\big|\theta\big)$, and maximising this bound with respect to the parameters. This leads to set of parameter update rules that can be guaranteed to montonicaly increase the likelihood and converge to an (at least local) maximum. For example, in the case of the combined model above, the update rules for p_{ml}, q_{mk} and $\pi_m^{[z_t]}$ are

$$p_{ml} \propto \sum_{j=1}^{J} \sum_{t_j=1}^{T_j} \mathrm{P}\big(x_t = \xi_m, \big| w_t, y_t, z_t = j, \theta\big)\, y_l^t, \tag{4}$$

$$q_{mk} \propto \sum_{j=1}^{J} \sum_{t_j=1}^{T_j} \mathrm{P}\big(x_t = \xi_m, \big| w_t, y_t, z_t = j, \theta\big)\, \mathbb{I}(w_t = v_k), \tag{5}$$

$$\pi_m^{[z_t]} \propto \sum_{t_j=1}^{T_j} \mathrm{P}\big(x_t = \xi_m, \big| w_t, y_t, z_t = j, \theta\big)\,. \tag{6}$$

The update rules for the attributional model and distributional model are special cases of the above rules, with the appropriate changes having been made.

3.2 Simulations

For the attribute model, we used the feature set collected in Vigliocco et al. (2004). For the distributional and combined models, the text corpora used consisted of a fiction and non-fiction books publicly available at the Oxford Text Archive ($\approx 6.5 \cdot 10^6$ words) and Project Gutenberg ($\approx 11.6 \cdot 10^6$ words), a set Reuters Newswire texts ($\approx 2.5 \cdot 10^6$ words), and a set of Usenet articles ($\approx 5.25 \cdot 10^6$ words). We folded British into American spellings, folded affix-variations of words into their word-stems, and eradicated non-words and stop-words. This reduced the corpora to a total size of $\approx 7.7 \cdot 10^6$ words, with $16,956$ word-types. By further eradicating all word-types that appear with a frequency of greater than 10^4 or less than 10^2, we can reduce the total size to

NATION	MONEY	ALLAH	human	fruit	leg	CAR	PATIENT	WAR
AUTHORITY	HUNDRED	BIBLE	face	green	clothing	HORSE	MEDICAL	GUN
PRINCIPLE	COURT	BELIEF	hair	grow	wear	RIDE	DOCTOR	KILL
CENTURY	LAND	APOSTLE	eye	red	body	DRIVE	HEALTH	ATTACK
UNITE	PAY	CHURCH	shoulder	round	protect	DRIVER	MEDICINE	KNIGHT
GOVERNMENT	THOUSAND	DISBELIEVE	leg	sweet	cover	transport	nose	kill
SOCIETY	TAX	CHRIST	hand	juice	body	vehicle	body	weapon
RELIGION	CITY	JESUS	body	tree	warm	4-legs	human	anger
CONSTITUTION	SCIENCE	MARRY	foot	eat	long	wheel	eye	fear
POLITICAL	OFFICE	SPIRIT	mouth	food	humans	car	head	yell

Figure 2: Examples of the latent classes learned by (from left to right) the distributional, attributional and combined models. Capitals refer to words, while lower case refer to attributes.

$\approx 6.1 \cdot 10^6$ words, and $7,393$ word-types. This corpus was divided into a set of $51,160$ texts, each of which were ≈ 150 words long.

As described above, the latent variables in each model can be seen as distributions over words, attributes or both. In Figure 2, we provide examples of these learned distributions. The distributional model learns latent topics that are distributions over words in the corpus. In the examples shown, we see latent classes that could be labelled *government*, *business* and *religion*. For the attributional model, the attribute-classes that are learned could be labelled *human-body*, *fruit-vegetable* and *clothing*. The combined model learns latent classes that merge attributional and distributional classes. The classes learned could be labelled *transport*, *medical* and *war*, each defined both by clusters of words and clusters of attributes.

In each model, we can measure how much any given word predicts any other. The extent to which word v_j predicts v_i is given by

$$\mathrm{P}\big(w = v_i \big| w = v_j, \theta\big) = \sum_{\{x\}} \mathrm{P}\big(w = v_i \big| x, \theta\big)\, \mathrm{P}\big(x \big| w = v_j, \theta\big) \tag{7}$$

where $\mathrm{P}\big(x \big| w = v_j, \theta\big)$ is obtained from Bayes rule. This can be taken as a measure of semantic similarity that is, perhaps, more theoretically motivated than other measures, see Griffiths & Steyvers (2003). Below, we show words predicted by some example words in each of the three models. For comparison purposes between the models, and with human performance (described below), here we provide only prediction of words that were in both the text-based and attribute-based data-sets.

Dog

Distributional: growl bark chase lick whine cat tail paw wolf snap

Attributional: cat rabbit goat tail pig fox sheep horse bear fur

Combined: cat growl tail bark paw whine chase sheep lick wolf

Gun

Distributional: threat stab knife bomb kick kill argue snap knock murder

Attributional: murder sword dagger bomb threat knife shield threaten stab scream

Combined: threat knife murder stab bomb threaten kick snap knock argue

Ride

Distributional: motorcycle bicycle horse chase pant slide thumb truck clatter ankle

Attributional: carry drive travel train move pull approach walk push truck

Combined: motorcycle bicycle horse travel pull slide truck chase carry push

We can see from these demonstrations that the attributes associated with a word, and its distribution across texts both provide information about its the meaning. However, in using both sources of information, patterns in one source can interact with those of another. As a trivial example, if words v_a and v_b are related with respect to attributional information, while word v_b and v_c are related with respect to distributional information, then we might infer that v_a and v_c are related. Inferences may be made given the ensemble of correlations between and within the two data sources. From our examples, we can see that attributional information leads to knowledge of patterns of body parts, while distributional information leads to knowledge of patterns relating to medicine. Together both can lead to inference that realm of medicine and medical things are coupled with realm of body-parts and functions.

Ongoing studies are being carried to further the work done here. These studies aim to elaborate the phenomena described here by using more powerful models and larger data-sets. In addition, we are comparing these models with human semantic ratings. We speculate that semantic relationships formed when both attributional and distributional information are used in combination will be more accurate in describing human performance than models based on either source alone.

4 Conclusions

"Meaning is the goal of communication" (Glenberg & Robertson (2000), p. 379), therefore theories of how meaning is represented in the brain/mind and how it develops during childhood are central to cognitive science and cognitive neuroscience. In this paper we have argued that previous models based solely on attributional (as FUSS) or distributional information, e.g (Landauer & Dumais 1997, Burgess & Lund 1997, Griffiths & Steyvers 2003), capture some important facts about how words' meanings are learnt, but they also have important limitations. In particular, models based on attributional knowledge (operationalised in terms of speaker-generated features) such as FUSS ground conceptual knowledge in our interactions with the environment and have some neural plausibility in that the conceptual features (the building blocks for the model) are assumed to be organised following the organisation of the sensory-motor systems in the brain (an assumption that has received support from patients and imaging studies, as we have described in the introduction). Moreover, these models have been shown to be successful in fitting adult behavioural data, thus showing psychological plausibility. From a developmental perspective, these models provide an account of how words' meanings can be linked to conceptual knowledge that develops independent of language. Thus models based only on attributional information have a number of attractive properties. These models have been argued to be superior to models based solely on distributional information (operationalised in terms of co-occurrence in large corpora of texts). Although distributional models show at least some psychological plausibility, it is unclear how they can be linked to neural processes and to other cognitive functions (see Glenberg & Robertson (2000) for a discussion).

Nonetheless, models based solely on attributional information have important limitations. Although the conceptual information independent of language is crucial, it does not represent the only way in which children learn words' meaning. From early stages in their development, children are exposed to language and it is reasonable to assume that they implicitly use the distributional information that is provided in the linguistic input in order to learn new words, as well as to enrich the semantic representation for words referring to concrete referents (which we assume to be primarily learnt via attributional information). Using probabilistic models we have shown how these two different sources of information can be combined. This combined model, we argue, is more plausible in terms of what information is used in learning word meanings; it has some neural plausibility (i.e., in addition to maintaining the assumption underlying FUSS in which words' meanings are grounded in sensory-motor systems, the model developed here uses principles that have been shown to be neurally plausible in modelling other cognitive and neuro-computational domains, e.g. (Körding & Wolpert 2004, Yu & Dayan 2005, Knill & Pouget 2004, Deneve & Pouget 2004). Finally, although at the moment this is a prediction awaiting empirical scrutiny, we also expect that this model has psychological plausibility in fitting behavioural data.

References

Barsalou, L., Simmons, W., Barley, A. & Wilson, C. (2003), 'Grounding conceptual knowledge in modality-specific systems.', *Trends in Cognitive Science* **7**, 84–91.

Bloom, P. (2002), 'Mindreading, communication and the learning of the names for things.', *Mind and language* **17**, 37–54.

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

Deneve, S. & Pouget, A. (2004), 'Bayesian multisensory integration and cross-modal spatial links.', *Journal of Physiology (Paris)* **98**, 249–258.

Farah, M. J. & McClelland, J. (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 (Special Volume of The Philological Society, Oxford)', Blackwell, Oxford, pp. 1–32.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A. & Trueswell, J. (2005), 'Hard words', *Developmental Psychology* . (In Press).

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**(3), 379–401.

Griffiths, T. & Steyvers, M. (2003), Prediction and semantic association, *in* S. T. S. Becker & K. Obermayer, eds, 'Advances in Neural Information Processing Systems 15', MIT Press, Cambridge, MA, pp. 11–18.

Hauk, O., Johnsrude, I. & Pulvermüller, F. (2004), ‘Somatotopic representation of action words in human motor and premotor cortex.’, *Neuron* **22**, 301–307.

Hinton, G. & Shallice, T. (1991), ‘Lesioning an attractor network: Investigations of acquired dyslexia.’, *Psychological Review* **98**, 74–95.

Knill, D. & Pouget, A. (2004), ‘The bayesian brain: the role of uncertainty in neural coding and computation.’, *Trends in Neuroscience* **27**(12).

Kohonen, T. (1997), *Self-organizing maps.*, Springer, New York.

Körding, K. & Wolpert, D. (2004), ‘Bayesian integration in sensorimotor learning’, *Nature* **427**(244-247).

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.

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.

Tettamanti, M., Buccino, G., Saccuman, M., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. & Perani, D. (2005), ‘Listening to action-related sentences activates fronto-parietal motor circuits.’, *Cognitive Neuroscience* **17**, 273–281.

Vigliocco, G., Vinson, D., Damian, M. & Levelt, W. (2002), ‘Semantic distance effects on object and action naming’, *Cognition* **85**, 61–69.

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., Arciuli, J., Siri, S., Scott, S. & Wise, R. (2005), The role of semantics and grammatical class in the neural representation of words. *Manuscript Under Review*.

Vinson, D. & Vigliocco, G. (2002), ‘A semantic analysis of grammatical class impairments.’, *Journal of Neurolinguistics* **15**, 317–351.

Vinson, D., Vigliocco, G., Cappa, S. & Siri, S. (2003), ‘The breakdown of semantic knowledge along semantic field boundaries: Insights from an empirically-driven statistical model of meaning representation.’, *Brain and Language* **86**, 347–365.

Yu, A. J. & Dayan, P. (2005), Inference, attention, and decision in a bayesian neural architecture., *in* ‘Advances in Neural Information Processing Systems’, Vol. 17, MIT Press, Cambridge.