ELSEVIER

Cognitive Science 28 (2004) 89–104

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Modelling the effects of semantic ambiguity in word recognition

Jennifer M. Rodd [a,*], M. Gareth Gaskell [b], William D. Marslen-Wilson [c]

[a] *Centre for Speech and Language, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK*
[b] *Department of Psychology, University of York, UK*
[c] *MRC Cognition and Brain Sciences Unit, Cambridge, UK*

Received 21 August 2002; received in revised form 5 June 2003; accepted 15 August 2003

Abstract

Most words in English are ambiguous between different interpretations; words can mean different things in different contexts. We investigate the implications of different types of semantic ambiguity for connectionist models of word recognition. We present a model in which there is competition to activate distributed semantic representations. The model performs well on the task of retrieving the different meanings of ambiguous words, and is able to simulate data reported by Rodd, Gaskell, and Marslen-Wilson [J. Mem. Lang. 46 (2002) 245] on how semantic ambiguity affects lexical decision performance. In particular, the network shows a disadvantage for words with multiple unrelated meanings (e.g., *bark*) that coexists with a benefit for words with multiple related word senses (e.g., *twist*). The ambiguity disadvantage arises because of interference between the different meanings, while the sense benefit arises because of differences in the structure of the attractor basins formed during learning. Words with few senses develop deep, narrow attractor basins, while words with many senses develop shallow, broad basins. We conclude that the mental representations of word meanings can be modelled as stable states within a high-dimensional semantic space, and that variations in the meanings of words shape the landscape of this space.

Keywords: Word recognition; Connectionist networks; Semantic ambiguity; Lexical ambiguity; Distributed representations; Lexical decision

* Corresponding author. Tel.: +44-20-7679-1096; fax: +44-20-7436-4276.
E-mail address: jrodd@ucl.ac.uk (J.M. Rodd).

0364-0213/$ – see front matter © 2003 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/j.cogsci.2003.08.002

1. Introduction

1.1. Semantic ambiguity

Recent connectionist models of word recognition characterize the process of retrieving a word's meaning as a mapping between an input representation (orthography or phonology) and a semantic representation (e.g., Gaskell & Marslen-Wilson, 1997; Hinton & Shallice, 1991; Joordens & Besner, 1994; Plaut, 1997; Plaut & Shallice, 1993). Typically, the semantic representations are distributed, such that the meaning of each word is represented as a pattern of activation across a large set of units, with each unit corresponding to some aspect of its meaning. Most models make the simplifying assumption that a word's meaning can be characterised as a single pattern of activation across these units. In other words, they assume that words have a single, well-defined meaning, and that there is a one-to-one mapping between the form of a word and its meaning. For most words, however, this assumption is incorrect. Here, we explore the implications of semantic ambiguity for connectionist models of word recognition.

The most straightforward examples of words that do not have a one-to-one mapping between form and meaning are homonyms. These are words that have a single spoken and written form that refers to multiple unrelated concepts; for example, the word form *bark* can refer either to a part of a tree, or to the sound made by a dog. These two meanings are entirely unrelated, and it is a historical accident that they share the same form. A second (more frequent) case where a single form can map onto multiple semantic representations is the case of polysemous words. These have a range of systematically related senses. For example, *twist* has several dictionary definitions, including to make into a coil or spiral, to operate by turning, to alter the shape of, to misconstrue the meaning of, to wrench or sprain, and to squirm or writhe. Although these definitions are clearly related, there are also important differences between them. For example if you were to *twist the truth* you would not expect the truth to be injured or to feel pain, which would both be appropriate for the phrase *twist an ankle*.

This distinction between unrelated word meanings and related word senses is respected by all standard dictionaries; lexicographers decide whether different usages of a word correspond to separate lexical entries, or to different senses within a single entry. This allows us to measure the frequency of these two types of ambiguity. Of the 4930 entries in the Wordsmyth dictionary (Parks, Ray, & Bland, 1998) with word-form frequencies greater than 10 per million in the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993), 7.4% correspond to more than one entry in the dictionary, and are therefore classified as homonyms. However, 84% of the dictionary entries have multiple senses, and 37% have five or more senses. This shows that most words in English are ambiguous in some way. For a model of word recognition to be applicable to the majority of words in the language, it must, therefore, be able to explain how words with different meanings and different senses are represented and recognised, as well as accounting for experimental data concerning the effect of these different types of ambiguity on word recognition.

The task that has been used most often to investigate effects of ambiguity on word recognition is lexical decision. Several studies have reported faster lexical decision times for ambiguous words, compared with unambiguous words. Early reports of an ambiguity advantage came from Rubenstein, Garfield, and Millikan (1970) and Jastrzembski (1981), who found faster

visual lexical decisions for ambiguous words compared with unambiguous words matched for overall frequency. Gernsbacher (1984) discussed a possible confound with familiarity in these experiments, since ambiguous words are typically more familiar. She found no effect of ambiguity over and above familiarity. Since then, however, several papers have reported an ambiguity advantage in visual lexical decision experiments using stimuli that were controlled for familiarity (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Hino & Lupker, 1996; Kellas, Ferraro, & Simpson, 1988; Millis & Button, 1989; Pexman & Lupker, 1999). Although these studies vary in the robustness of the effects reported, their cumulative weight seemed to establish the ambiguity advantage as an important constraint on theories of lexical representation and lexical access.

More recently, the view that there is a processing advantage for semantic ambiguity has been strongly challenged. Rodd, Gaskell, and Marslen-Wilson (2002) argue for a distinction between words like *bark* which, by chance, have two unrelated meanings, and words like *twist* that have multiple, related senses. It is likely that the mental representations of these two types of words will differ significantly. In a set of visual and auditory lexical decision experiments, Rodd et al. (2002) replicate the ambiguity advantage for words that have multiple related word senses (e.g., *twist*), but found that for ambiguous words that have multiple unrelated meanings (e.g., *bark*) the effect of ambiguity is reversed—multiple meanings delay recognition. Previous studies reporting an ambiguity advantage have done so, we claim, because they have confounded ambiguity between multiple meanings with ambiguity between multiple senses. The stimulus sets that show an ambiguity advantage primarily contrast words with multiple senses, and not words with multiple meanings. The challenge for models of word recognition is to explain how these two apparently contradictory effects of ambiguity can emerge from a single architecture. We discuss them in turn in the next two sections.

1.2. Modelling the effect of ambiguity between unrelated meanings

The recent results of Rodd et al. (2002) suggest that different forms of ambiguity have different effects on word recognition. This renders problematic the earlier accounts of ambiguity, which focused on explaining the apparent ambiguity advantage for words with multiple meanings. Traditional accounts (e.g., Rubenstein et al., 1970) assumed that ambiguous words benefit from having more than one competitor in a race for recognition. More recently, there have been attempts to show that the ambiguity advantage can emerge from connectionist models of word recognition, although this has been far from straightforward. The following section explains why it has proved difficult to get distributed connectionist models to show an ambiguity advantage.

For connectionist models that characterize word recognition as a mapping between orthographic and semantic representations, this process is straightforward for unambiguous words; there is a one-to-one mapping between the two patterns, and a network can easily learn to produce the appropriate semantic pattern in response to orthographic input. In contrast, for words with multiple meanings, the situation is more complicated; a single orthographic pattern is paired during training with more than one semantic pattern. Many connectionist models of word recognition use strictly feed-forward connections, trained using the deterministic back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). These models

cannot cope with the one-to-many mapping between form representations and semantics. Given such an ambiguity, the best solution that back-propagation can achieve is a compromise, where the semantic activation produced in response to the form of an ambiguous word is a blend between the two possible meanings (see Movellan & McClelland, 1993 for a detailed discussion of this issue). This blend will be biased by the frequency of the different meanings, such that it most closely resembles the more frequent meaning. Such blends between unrelated meanings do not correspond to coherent meanings of real words.

One possible solution is to allow interaction within the semantic representations, typically implemented by including recurrent connections between semantic units. In this type of network, although the initial pattern of activation of the semantic units generated by an ambiguous input may be a blend between different meanings, the recurrent connections within the semantic layer are able to "clean up" this pattern. Activation is continuously modified until the pattern of activation becomes stable—known as an *attractor state*. If the learning algorithm has correctly adjusted the recurrent weights within the semantic layer, these weights can ensure that all the patterns seen in the training set correspond to stable attractor states (and also increase the probability that blends between these states are not stable). In other words, although the initial state of the set of semantic units may correspond to a blend state, the recurrent connections between the semantic units can force a shift towards a stable attractor corresponding to one of the word's meanings. The frequencies of the meanings and the initial activation of the network determine which of the attractors is settled into on any given trial. This additional process of resolving the ambiguity, and thus moving from a blend state to a meaningful semantic representation, is likely to delay the process of settling into a stable attractor for ambiguous words. It therefore seems likely that a network of this type would show an ambiguity disadvantage. Although this is inconsistent with the conventional pattern of results, it is consistent with the findings of Rodd et al. (2002).

Despite the apparent natural tendency for such models to show an ambiguity disadvantage, there have been several attempts—in response to the earlier results—to show the reverse effect of ambiguity. Kawamoto (1993) and Kawamoto, Farrar, and Kello (1994) simulated an ambiguity advantage in a network of this type by assuming that lexical decisions are made on the basis of orthographic representations. This effect arises because of the error-correcting nature of the learning algorithm that was used; in order to compensate for the increased error produced by the ambiguous words in the semantic units, stronger connections were formed between the orthographic units, which were being used as the index of performance.

Joordens and Besner (1994) and Borowsky and Masson (1996) trained a two-layer Hopfield network (Hopfield, 1982) to learn a mapping between orthography and semantics for words that are either unambiguous or ambiguous. In contrast to our claim that competition between word meanings should normally delay processing of ambiguous words, these models showed an advantage for the ambiguous words. The authors argue that this advantage arises because of a "proximity advantage". When the orthography of a word is presented to the network, the initial state of the semantic units is randomly determined. The network must move from this state to a valid finishing state corresponding to the meaning of the word. For ambiguous words, there are multiple valid finishing states and, on average, the initial state of the network will be closer to one of these states than for an unambiguous word, where there is only one valid finishing state.

These results are the reverse of what is predicted on the assumption that the additional processing associated with ambiguous words will increase the time that it takes to produce a coherent semantic representation. In these simulations, the advantage for ambiguous words that results from a proximity advantage seems to be sufficiently large to overcome any disadvantage produced by competition between different meanings. One limitation of these models is that the settling performance of the networks is poor. Joordens and Besner (1994) report an error rate of 74%. These errors result from the network frequently settling into blend states, which are a mixture of the ambiguous word's different meanings. In Borowsky and Masson (1996), these blend states are not considered errors; the authors argue that lexical decision does not require resolution of the ambiguity, and that a blend state would be sufficiently familiar to support a lexical decision. This idea that lexical decisions can be made on the basis of activating a blend state was extended by Piercey and Joordens (2000), who suggested that the reason that ambiguity produces an advantage in lexical decision but a disadvantage in text comprehension is that activating a blend state is enough to make a lexical decision, while for text comprehension a specific meaning must be retrieved.

Although blend states may possibly be sufficient to make lexical decisions, the fact that the networks used by Joordens and Besner (1994) cannot escape from these blend states is problematic for two reasons. First, the failure of these models to retrieve individual word meanings severely limits their value as general models of word recognition. It is clear that, given an ambiguous word in isolation, we can retrieve one of its meanings. For example, if participants are given an ambiguous word in a word-association task, they can easily provide an associate of one of its meanings (Twilley, Dixon, Taylor, & Clark, 1994). In contrast, these models predict that without a contextual bias, we would get stuck in a blend state. The second problem with these models' tendency to remain in blend states is that the ambiguity advantage observed in these networks may be an artefact of this tendency to settle into blend states. Indeed, Joordens and Besner (1994) report that as the size of their network is increased, and performance improves, the ambiguity advantage is eliminated.

The simulation reported here investigates the effects of different kinds of semantic ambiguity in a two-layer network trained on the mapping between orthographic and semantic representations. The network is based on the Hopfield network (Hopfield, 1982) used by Joordens and Besner (1994) and Borowsky and Masson (1996), but was modified to improve the network's performance on ambiguous words, such that it does not tend to settle into blend states. While Hopfield networks are known to have limited capacity, the networks used by Joordens and Besner (1994) and Borowsky and Masson (1996) are performing well below the theoretical capacity limit. Hopfield (1982, p. 2556) stated that "About 0.15 N states can be simultaneously remembered before error in recall is severe", where N is the number of units in the network. Therefore, the Joordens and Besner (1994) network should be able to learn 45 patterns, and yet the network cannot reliably learn four words. This poor performance arises because the patterns corresponding to the different meanings of ambiguous words are correlated—they share the orthographic part of their pattern. Hopfield (1982) noted that these networks have a particular difficulty with correlated patterns. Therefore, the simple Hebbian learning rule used in Hopfield networks, which captures the correlational structure of the training set, may not be suitable for learning ambiguous words. The simulation reported here uses instead the least mean-square error-correcting learning algorithm, which adjusts the weights between units to

reduce any error in the activation patterns produced by the current sets of weights. This has been shown by Rodd et al. (2001) to alleviate the problem of blend states; the learning algorithm changes the weights such that blend states are not stable.

In summary, prior research into how ambiguity between unrelated meanings affects both human lexical decisions and network performance has been rather inconclusive. The traditional view was that this ambiguity produced a benefit in lexical decision. This has been simulated in distributed connectionist networks, but only when the networks settle into blend states on a high proportion of the trials (Borowsky & Masson, 1996; Joordens & Besner, 1994), or when lexical decisions are made on the basis of orthographic representations (Kawamoto, 1993; Kawamoto et al., 1994). In contrast, Rodd et al. (2002) have shown that this form of ambiguity produces a *disadvantage* in lexical decision; this seems more consistent with the idea that in these networks interference between different meanings delays recognition. The model reported here investigates the effect of ambiguity between unrelated meanings in a model that uses an error-correcting learning algorithm to develop attractor structure that overcomes the problem of blend states. In addition, it investigates the effect of ambiguity between related word senses.

1.3. Modelling the effect of ambiguity between related senses

The models described above investigate the effect of ambiguity between unrelated meanings, where a single orthographic pattern is paired with two uncorrelated semantic patterns. Here, we investigate for the first time the effect of ambiguity between related word senses. In contrast to the disadvantage for words with semantically unrelated multiple meanings, lexical decisions to words that are ambiguous between related word senses are faster, compared with unambiguous words (Rodd et al., 2002). Although it seems straightforward to explain the ambiguity disadvantage in terms of interference between the alternative meanings of a word, it is less clear how an advantage for semantic ambiguity could arise in such a system. One difference between these two forms of ambiguity is the degree of semantic overlap between the alternative semantic patterns. However, although an increase in the similarity of the two meanings of an ambiguous word might reduce the level of semantic competition (and therefore the ambiguity disadvantage), this can only improve performance to the level of the unambiguous words; it cannot produce a benefit; two related meanings might interfere less than two unrelated meanings, but they would still interfere (Rodd, 2000).

In this simulation, we explore the hypothesis that variation in the meanings of words such as *twist*, which are listed as having many word senses, should be viewed not in terms of ambiguity, but in terms of flexibility. We assume that the multiple senses of these words are not distinct, but that their meaning is flexible and therefore variable, such that it has a slightly different interpretation in different contexts. Therefore, rather than being ambiguous between unrelated meanings that correspond to distinct attractor basins in different parts of semantic space, these words have a range of possible meanings that fall within a single large attractor basin. We can almost think of these words as “noisy”, such that semantic patterns corresponding to their different senses are noisy versions of some kind of “core” or average meaning. The reason that we might expect this characterization of word senses to facilitate

network performance is that, if an identical pattern is repeatedly presented to the network, it can develop a very deep attractor basin that can be difficult for the network to settle into when it is given only the orthographic input (e.g., Rodd et al., 2001). It is possible that adding a small amount of noise to the network (corresponding to a degree of variability in meaning) might allow the network to develop broader attractor basins that are easier for the network to enter.

This variability in meaning can be captured by generating a base semantic pattern for each word, and then adding random variation to this pattern. However, this is not a realistic characterization of how the senses of words differ; it is not the case that a new sense of a word can be created from its core meaning simply by changing arbitrary features. For example, the word *twist* may in some contexts not activate the features relating to pain, but it will never arbitrarily gain a feature such as "has legs". Rather, these words have sets of possible semantic features that are sometimes, but not always, present. In this simulation, we assume that each word has a range of possible semantic features, and that for each sense some (but not all) of these features are turned on.

Although this idea that words with many senses can be characterised as words whose meaning can vary within a region of semantic space does not reflect the categorical way in which these words are listed in dictionaries, there is support for this idea that the classification of the meanings of such words into distinct senses is artificial. For example, Sowa (1993) states that "for polysemous words, different dictionaries usually list different numbers of meanings, with each meaning blurring into the next". However, it is an oversimplification to assume that the variation between word senses is entirely random; the relationships between different word senses can be highly systematic (Klein & Murphy, 2001), and some semantic features will be more likely to coexist than others. However, despite these limitations, using random variation allows us to investigate the general properties of this network without making any assumptions about the structure of this variation.

2. Simulation: the ambiguity disadvantage and sense benefit

2.1. Introduction

This simulation contrasts the two forms of ambiguity used by Rodd et al. (2002) in a factorial design. Ambiguity between unrelated meanings was simulated using training items in which a single orthographic pattern was paired with two uncorrelated semantic patterns. Words with many related senses were generated from a core pattern by selecting a random subset of its semantic features.

2.2. Method

2.2.1. Network architecture

The network had 300 units: 100 (orthographic) input units and 200 (semantic) output units. Each unit in the network was connected to all other units. All units were bivalent; either on [+1] or off [0].

2.2.2. Learning algorithm

Connection strengths were initially set to zero. During each learning trial, the network was presented with a single training pattern, i.e., the activation of every unit in the network was clamped at either [+1] or [0] according to its target value for that pattern. An error-correcting learning algorithm was then used to change the connection strengths for all the forward connections from orthographic units to semantic units, and the recurrent connections between semantic units. The change in connection strength from a given unit i to a unit j was proportional to the difference between the current activation for unit j (which was set to its target value) and the total activation that this unit received from all the other units in the network (which were also set to their target values). This weight change is given by

$$\Delta w_{ij} = S\frac{x_i(x_j - \sum_k w_{kj}x_k)}{n}, \quad i \neq j,\ k = 1 \ldots n \tag{1}$$

where w_{ij} is the connection strength between units i and j, x_i is the target activation for unit i, and n is the total number of units in the network. The learning rate parameter (S) was set to 5, to provide good performance after relatively little training.

2.2.3. Training

The orthographic and semantic training patterns were sparse, such that only 10% of the units were set to [+1] and the remainder were set to [0]. The network was trained on 32 words, of which half had a single meaning: a single input pattern paired with a single, randomly generated output pattern. The remaining words were ambiguous between two meanings: a single input pattern paired with two different, randomly generated output patterns. Further, to generate sets of few- and many-sense words, half of each group had noise added to the semantic representations during training, and half did not. The patterns for few-senses words without added noise had 20 of the 200 semantic units turned on. For the many-sense words, 25 of the 200 semantic units were selected for the base pattern, and each training exemplar was generated by randomly selecting 20 of these units. Therefore, the actual patterns presented to the network for two types of words were equated for sparseness. This reflects the intuition that individual senses of the different types of word do not differ in terms of the amount of associated semantic information. Each of the 32 training items was presented to the network 128 times, such that for the ambiguous items, each of its meanings would be presented 64 times. This ensured that the presentation frequencies of the ambiguous and unambiguous words were equated in terms of the frequency of their orthographic form. The network was trained and tested using 100 independently generated sets of training items.

2.2.4. Testing

Each input pattern was presented to the network, with the output units initially set to zero. Retrieval of the semantic patterns was the result of an asynchronous updating procedure, which consisted of a series of updates in which a randomly selected semantic unit was updated by summing the weighted input to that unit. If this input was greater than 0.5, then the unit was set to [+1]; otherwise, the unit was set to zero. This updating continued for 2000 updates.

To increase the sensitivity to any effects that might occur early in the settling of the network, noise was added to the activation of units during testing; this has the effect of slowing the

settling of the network. The noise acted such that when a unit was updated on the basis of the activation of the other units, there was a 15% probability that a unit that was set to [+1] would be set to zero. A set of simulations (not reported in detail here) showed that if this noise was not included, the overall pattern of results was the same, although the differences between conditions were smaller, mainly due to a ceiling effect for the unambiguous words, which were typically retrieved as soon as all units had been updated once.

To check that the model can appropriately discriminate words from nonwords, each network was also tested on novel items that shared 50% of the orthographic features of the real words.

2.3. Results

Before assessing the network's performance in terms of its ability to simulate lexical decision performance, we need to check that the network is settling into the correct meanings. In other words, we need to ensure that when the network is not under any time pressure it retrieves an appropriate meaning for all the types of words. Deciding what constitutes an "appropriate" meaning is not straightforward. For words with no noise added during training, we would expect all 20 semantic features to be activated. However, the words with noise added during training had a maximum of 25 possible semantic features, of which 20 were present on any training presentation. For these words there is no reason to assume that exactly 20 of these units should be activated on any given test trial. Therefore the network was considered to have produced an appropriate response if it activated at least 15 appropriate semantic features, and if it did not incorrectly activate any semantic features. The same criterion was used for all types of words and all levels of noise. For the ambiguous words, the network was considered to have settled appropriately if it settled into either meaning of the word. Using these criteria, the overall performance of the network was very good; the network settled appropriately on 99.8% of trials. All the errors were made on ambiguous words with few senses, and in all cases were caused by the network failing to activate sufficient semantic features.

Although we have not implemented a full model of lexical decision, in the following analysis, we assume that lexical decision latencies reflect the time taken to activate sufficient semantic features of a word to distinguish it from meaningless nonwords. In line with other researchers (cf. Grainger & Jacobs, 1996), we assume that "no" decisions are made using a response deadline that can be modified on the basis of the overall degree of semantic activation (i.e., greater semantic activation will lead to an extended deadline).

Fig. 1 shows the total number of semantic units that were switched on during the settling of the network. The first thing to note is that there was relatively little semantic activation in response to the nonwords. When these nonword responses were looked at in detail, it was clear that most of the semantic activation seen in response to nonwords consisted of relatively transient peaks of activation which were quickly turned off. The apparently stable plateau of activation seen in Fig. 1 is the result of averaging together many of these transient peaks of activation.

Looking at the responses to the real words shows two main effects. First, the ambiguous words took longer to activate semantic units than the unambiguous words did. Second, for both ambiguous and unambiguous words, the network activated the semantic units more quickly for words with noise added during training, that is, the many-sense words.[1] To investigate these

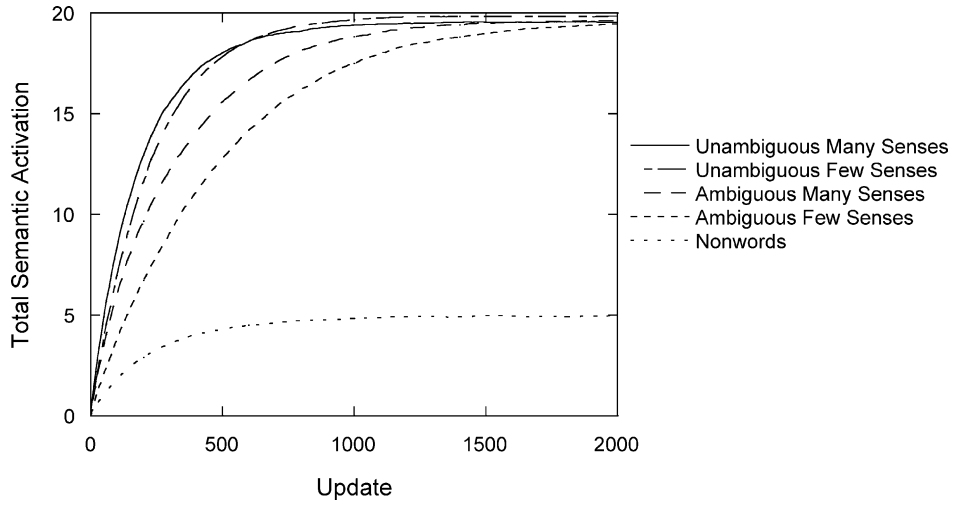

Fig. 1. Activation of semantic units during settling.

effects in more detail, we applied a decision criterion such that the network would make a positive lexical decision once 10 semantic units were activated. The number of updates taken to reach this criterion for the different conditions in shown in Table 1. An ANOVA on these data with ambiguity and senses entered as factors confirmed that both these main effects were significant (Ambiguity: $F(1, 3189) = 980$, $p < .001$; Senses: $F(1, 3189) = 470$, $p < .001$). There was also a significant interaction between these variables, such that the effect of multiple senses was larger for the ambiguous words than the unambiguous words ($F(1, 3245) = 136$, $p < .001$). A similar tendency was seen in the lexical decision data reported by Rodd et al. (2002). Further analyses showed that the effect of number of senses was significant for both ambiguous ($t(1591) = 19$, $p < .001$) and unambiguous words ($t(1598) = 13$, $p < .001$).

Importantly, although selecting a decision threshold of 10 was relatively arbitrary, Fig. 1 shows that these differences between conditions emerge very early (before the words can reliably be differentiated from nonwords), and continues until about 18 of the 20 semantic features have been retrieved. Therefore, if we assume that lexical decision latencies reflect the speed with which semantic information about words becomes available, but that it is not necessary for the representation to be completely stable before a response is made, this network can simulate the pattern of lexical decision data reported by Rodd et al. (2002). However in the latter stages of settling, there is a cross-over for the unambiguous words such that there is an

Table 1
Number of updates to activate 10 semantic units

Ambiguity	Senses	Mean	S.E.
Ambiguous	Few	391	5.0
Unambiguous	Few	157	5.0
Ambiguous	Many	205	5.0
Unambiguous	Many	127	5.0

advantage for the few-sense words. This suggests that in a different task in which decisions are made only once semantic representations are completely stable, there may be a reverse sense effect, such that words with few senses are responded to more quickly.

To investigate the dynamics of this network further, we looked at the stress level of the semantic units within the network. Stress provides a measure of the stability of the network, which in turn represents how deep into a stable attractor the network has moved. Plaut (1997), defined stress as

$$\text{Stress} = \sum((a_i \log_2 a_i) + (1 - a)\log_2(1 - a) - \log_2(0.5)), \quad (3)$$

where a_i is the activation of semantic unit i. In the current network, the activations of the units were thresholded and set to [0] or [+1]; however, the stability of the network can be evaluated by setting a_i to be the input to unit i before it has been thresholded. This measure is maximal when the inputs to all the units are close to the target values of either [0] or [+1], and decreases for all intermediate values. Fig. 2 shows the stress of the network for the different types of words during settling. (Minus semantic stress is plotted, so that the lower points correspond to more stable points in semantic space.)

The contrast between ambiguous and unambiguous words in Fig. 2 shows that at all stages in settling, the network is more stable for the unambiguous compared with the ambiguous words. This suggests that for the ambiguous words, the network is slower to move into an attractor basin, and that its final stable state is in an attractor basin that is less deep than the basins that correspond to unambiguous words. The contrast between words with few and many senses is more complex: early in settling, the network is more stable for the words with many senses. However, this benefit for the words with noise added during training then reverses, and is replaced by a benefit for the unambiguous words. The early benefit for the words with many senses shows that, for these words, the network is entering the appropriate attractor basin faster than is the case for the few-sense words. The later disadvantage for these words reflects the

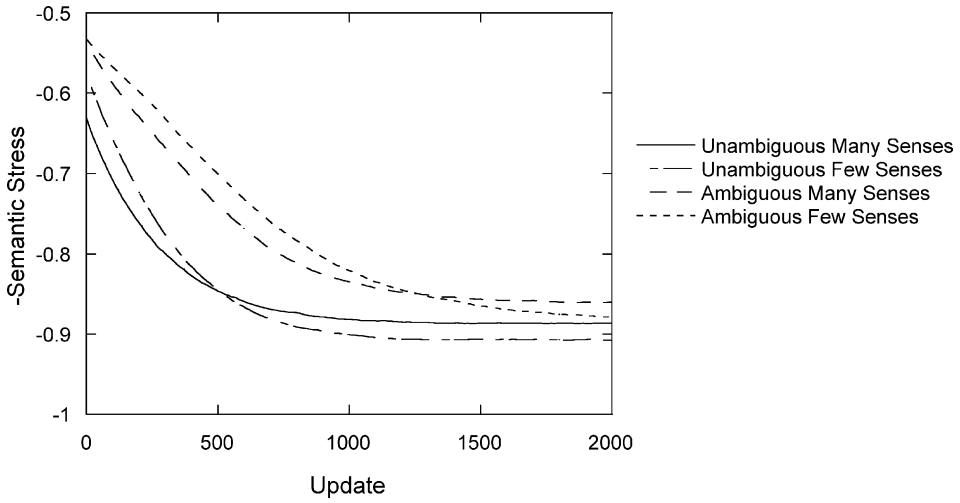

Fig. 2. Stability of network during settling.

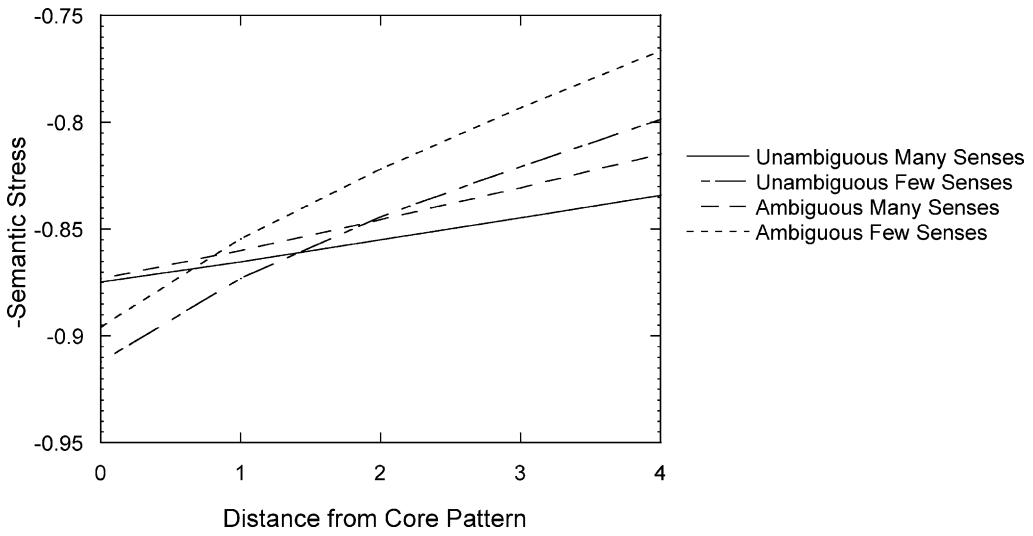

Fig. 3. Stability of semantic activation near to learned patterns.

fact that the attractor basins for these words are less deep than for the few-sense words; the words with many senses become relatively stable faster, but do not settle into a state that is as stable as the few-sense words.

To confirm the differences in the attractor structures for these words, we obtained estimates for their average attractor basin shapes by measuring the network's stability near the centre of an attractor basin by turning off between 0 and 4 of its semantic features. Fig. 3 shows a plot of the stability of the network as a function of the distance from the core training pattern for the two types of words. This shows that at the centre of the attractor basins, the pattern of activation is more stable for the words that have had no noise added during training, but that the attractor basins are wider for the words with many senses. In other words, the area of relatively stable semantic space surrounding the core pattern is larger for the words with many senses. This supports our argument that it is the difference between the structures of the attractor basins that accounts for the difference in performance for the two types of words.

3. General discussion

In this paper, we have argued that the phenomenon of semantic ambiguity provides important constraints on models of word recognition. Most words in English are to some extent semantically ambiguous, and this ambiguity is known to affect performance on tasks such as lexical decision. Although preceding sentential context is often highly valuable in the resolution of ambiguity, there are also many cases in which sentential context is insufficiently constraining, or even absent altogether. Although even in the absence of a constraining sentence context, it is possible that residual activation in the system might bias participants towards one meaning rather than another, it seems unlikely that this context would ever be strong enough to prevent the system settling into a blend state. In the model presented here, although factors like context

and the frequency of a word's different meanings would influence which meaning is settled into on any given trial, it is the attractor structure that prevents the system from remaining in a blend state, and that ensures that the final state of the network after settling corresponds to a real meaning of the word.

The aim of the simulation reported here was to investigate whether models of word recognition in which there is competition to activate distributed semantic representations can not only overcome this problem of blend states, but can also accommodate the pattern of semantic ambiguity effects reported by Rodd et al. (2002). They reported that the effect of semantic ambiguity on performance on a lexical decision task depends on the nature of the ambiguity: ambiguity between unrelated meanings delays recognition, while ambiguity between related word senses improves performance.

Here we have shown that these opposite effects can indeed emerge from a model that incorporates distributed semantic representations, and where the interaction between semantic units allows an attractor structure to develop within the semantic layer of representation. The ambiguity advantage and the sense benefit can be explained in terms of the structure of their semantic representations. The ambiguity disadvantage emerges because words such as *bark* have separate meanings that correspond to separate attractor basins in different regions of semantic space. For these words, the orthographic input is ambiguous, and in the early stages of the network's settling, a blend of these meanings will be activated. The connections within the semantic units constrain the activation, such that features relating to only one of the possible meanings are activated; the network moves away from the blend state and settles in one the different meanings. This is in contrast to the claim made by Piercey and Joordens (2000) that there is an advantage for these words because they settle into a blend state quickly, and that this blend state is sufficient to make a lexical decision. Under our account it is the process of moving away from a blend state that makes these words harder to recognise.

In contrast, the different possible semantic representations of words with multiple senses do not correspond to separate regions in semantic space; the distributed semantic representations of the different senses of these words are highly overlapping, and thus correspond to neighbouring points in semantic space. The simulation showed that if a single input orthographic pattern is repeatedly presented with a semantic pattern that varies within a small area of semantic space, then the attractor basin for this word becomes broader. In other words, there is a larger area of semantic space that corresponds to the meaning of the words than would be the case if there were no variation. This simulation also showed that this broadening of the attractor basin can produce faster activation of semantic features. It is important to note that this explanation predicts that the sense benefit should be restricted to tasks such as lexical decision in which the activation of *any* semantic information is sufficient to support performance.[2] In a task that requires a particular sense of a word to be retrieved, or where a decision can only be made once the semantic representation is completely stable, it is likely that the different word senses will compete with each other and produce a sense disadvantage (e.g., Klein & Murphy, 2001).

In summary, the model produces the two opposing effects of ambiguity because unrelated meanings are represented in different parts of semantic space and compete with each other for activation, while multiple senses are represented within a single region of semantic space and combine to form a single large attractor basin. We interpret the success of this model in simulating the apparently opposite effects of semantic ambiguity as evidence to support the

claim that the meanings of words are represented as distributed patterns of activation across a large set of semantic units, where there is also interaction within the semantic representations, and that the speed with which these semantic representations can be activated plays an important role in even an apparently nonsemantic task such as lexical decision.

An alternative account of these data would be that these effects emerge from pre-semantic lexical representations. For example, the ambiguity advantage reported by Borowsky and Masson (1996), Millis and Button (1989), and Azuma and Van Orden (1997) can be interpreted in terms of models of word recognition in which words compete to activate localist, abstract lexical representations; ambiguous words are assumed to have multiple entries in the race for recognition, and are therefore recognised more quickly (see, for example, Jastrzembski, 1981). These accounts could be adapted to accommodate the ambiguity disadvantage for words with unrelated meanings reported by Rodd et al. (2002). It is possible that the disadvantage in lexical decision for words like *bark* is a nonsemantic effect that arises because of differences in the frequency of the individual word meanings, or as a result of lateral inhibition between the nodes corresponding to the different meanings of an ambiguous word.

It is less clear how one might explain the word sense benefit as a nonsemantic lexical effect. One alternative might be to assume that individual word senses are represented by individual localist "sense nodes". Within such a framework, we would need to assume that while the nodes corresponding to different meanings compete with each other and delay recognition, the nodes corresponding to different senses do not. In other words, the sense benefit could be interpreted in a similar way to the original accounts of the ambiguity advantage; words with many senses benefit from having multiple entries in the race for recognition. Workable though such an account might be, it is both ad hoc and inconsistent with our intuitions about word senses. Lexicographers have great difficulty in determining the boundaries between different word senses, and in deciding which of the different usages of a word are sufficiently frequent and distinct to warrant their separate inclusion in the dictionary. Any model that includes word sense nodes would have to make similar arbitrary divisions. This framework also seems to undermine the fundamental nature of word senses. Word senses allow us to express new ideas by extending the meanings of known words in systematic ways, and the variation in word meanings allows them to be used in a variety of contexts, with subtly different meanings. The idea that word senses should be represented as discrete, nonsemantic representations, rather than by an intrinsic flexibility in distributed lexical semantics, is inconsistent with this widely accepted view of word senses.

We argue that in order to capture the natural variability in meaning of words with many senses, we must move away from the approach of capturing a word's meaning with a single, static representation, towards a theory of semantic representation that can capture the flexible, productive and expressive nature of lexical semantics. This is likely to require us to assume that the meanings of words are represented in a distributed manner.

Notes

1. To investigate the implications of our decision to characterise the variation in the meanings of many-sense words as selecting a set of features from a finite set of possible features, we conducted a control simulation in which all the words had 20 semantic

features, and the variation of the many-sense words was entirely random. Under these conditions, the effect of word senses was small and variable: there was a small sense benefit for the ambiguous words but a sense disadvantage for the unambiguous words.

2. The size of ambiguity effects in lexical decision tend to increase as the nonwords become more word-like (e.g., Piercey and Joordens, 2000). This is likely to reflect a shift towards relying on semantic representations. When the nonwords look very different to words the orthographic information can be used to discriminate words form nonwords (and so effects of semantic variables are small), whereas when highly word-like nonwords are used participants reply more on semantic cues to make the discrimination.

Acknowledgments

This work was supported by a Research Fellowship from Peterhouse, Cambridge to J.R. and by the UK Medical Research Council.

References

Azuma, T., & Van Orden, G. C. (1997). Why safe is better than fast: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*, 484–504.

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). The CELEX lexical database.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*, 63–85.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*, 254–281.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518–565.

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming—An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1331–1356.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America: Biological Sciences*, *79*, 2554–2558.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence and the lexicon. *Cognitive Psychology*, *13*, 278–305.

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank—Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning Memory and Cognition*, *20*, 1051–1062.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*, 474–516.

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1233–1247.

Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 601–609.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*, 259–282.

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory & Cognition*, *17*, 141–147.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, *17*, 463–496.

Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English dictionary-thesaurus*. University of Chicago. Retrieved February 1, 1999, from: http://www.wordsmyth.net/Chicago.

Pexman, P. M., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology*, *53*, 323–334.

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition*, *28*, 657–666.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*, 765–805.

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

Rodd, J. M. (2000). *Semantic representation and lexical competition: Evidence from ambiguity*. Cambridge: University of Cambridge.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2001). For better of worse: Modelling effects of semantic ambiguity. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 863–868). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*, 245–266.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*, 487–494.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. I, pp. 318–364). Cambridge, MA: MIT Press.

Sowa, J.F., (1993). Lexical structure and conceptual structures. In J. Pustejovsky (Ed.), *Semantics and the lexicon* (pp. 223–262). Dordrecht/Boston/London: Kluwer Academic Publishers.

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*, 111–126.