A critique of connectionist semantics

JONATHAN A. WASKAN

Philosophy Department, William Paterson University,
300 Pompton Road, Wayne, NJ 07470, USA
email: waskanj@wpunj.edu

Abstract. Amongst modellers of natural language comprehension, the suspicion that explicit semantic representations are inherently biased has led many to rely more heavily on the ability of networks to form their own internal semantic representations over the course of training. The concern over explicit semantics, however, betrays a lack of appreciation for the manner in which insidious biases can and cannot creep into models of comprehension. In fact, the trend of relying on networks to form their own internal semantic representations has done little to curtail one common form of insidious bias. Where models of natural language comprehension are concerned, the cause of inappropriate biases has everything to do with the manner in which regularities find their way into sentence/meaning pairs and little or nothing to do with the degree to which semantic information is made explicit. This is fortunate, as there may be drawbacks to relying too heavily on the ability of networks to form their own internal semantic representations.

Keywords: natural language, comprehension, semantics, linguistics, microfeatures.

1. Introduction

Natural language comprehension has long been taken to involve a mapping from linguistic (e.g. orthographic or phonological) representations to semantic representations (Geschwind 1965, Craik and Lockhart 1972, Hyde and Jenkins 1973). This process seems to involve the satisfaction of multiple syntactic and semantic constraints (Fillmore 1968, Fodor et al. 1974). It is therefore not surprising that researchers have recently been attempting to model the comprehension process by harnessing the powerful constraint-satisfaction capabilities of connectionist systems (McClelland and Kawamoto 1986, St. John and McClelland 1990, Miikkulainen 1993, Harris 1994, Kintsch 1998).

Amongst such researchers, there has lately been a trend of avoiding the explicit encoding of information concerning the properties of, and relationships amongst, the objects described by a given sentence. Researchers have instead been relying more heavily on the ability of connectionist networks to detect semantic information that is implicit in a training corpus and, thereby, to form their own internal semantic representations over the course of training. This trend is in no small measure a reaction to the worry that explicit semantic representations are inherently and insidiously biased (Lachter and Bever 1988, Harris 1990, St. John and McClelland 1990, Miikkulainen 1993). This worry, I argue, is ill-founded. There is, however, a subtler way in which a modeller’s encoding activities can foster misgivings about the putative achievements of a given model.
In what follows, I clarify the manner in which insidious biases can and cannot creep into models of comprehension. I also discuss the drawbacks of relying too heavily upon the capacity of networks to form their own internal semantic representations. In short, I hope to exonerate, and perhaps even provide a mandate for, the explicit encoding of semantic information for purposes of modelling comprehension.

2. Comprehension through microfeatures
McClelland and Kawamoto (1986) designed one of the early connectionist models of sentence comprehension. They were chiefly concerned with a particular aspect of sentence meaning: the assignment of roles (e.g. agent, patient, instrument, etc.) to sentence constituents. As illustrated by (1)–(5) below, syntactic and semantic factors figure in the determination of which roles are filled by which entities. For example, sentences (2) and (3) suggest that the inclusion of a grammatical object bears, in certain contexts, on the role (viz. instrument or patient) assigned to an inanimate subject. The different role assignments for (4) and (5) suggest that lexical semantics are also very important.

(1) The boy broke the window.
(2) The rock broke the window.
(3) The rock broke.
(4) The boy broke the window with the rock.
(5) The boy broke the window with the curtain.

The primary goal McClelland and Kawamoto set for their model was to learn the complex syntactic and semantic constraints governing role assignment. McClelland and Kawamoto’s (1986) model (figure 1) is a two-layer, feed-forward network that maps sentence structure representations onto role filler representations. The network consists of four sets of input units and four sets of output units. Each set of input units encodes a sentence constituent in terms of an appropriate set of semantic ‘microfeatures’. Each semantic microfeature (e.g. fragile) falls into one of several categories. Noun categories include SOFTNESS, VOLUME, BREAKABILITY and GENDER. The microfeatures constituting each category (soft–hard, small–medium–large, fragile–unbreakable, male–female–neuter, respectively) are mutually exclusive. So, for instance, one might encode the term ‘window’ in terms of such microfeatures as hard, medium, fragile, neuter. Verb microfeatures fall into categories like TOUCH (specifying which entities, if any, are in contact with the patient), NATURE OF CHANGE (specifying what kind of change, if any, occurs in the patient) and AGENT MOVEMENT (specifying the nature of the movement, if any, undergone by the agent).

Following 50 cycles of training on a subset of possible sentence/meaning pairs, the model’s performance (when examined for appropriate activity of each of the 2500 output units) reached approximately 99% accuracy for both familiar and novel sentences (McClelland and Kawamoto 1986). The model thus seemed to exhibit an impressive degree of sensitivity to the complex constraints imposed by syntax and semantics on role assignment. There is clearly more to sentence comprehension than role assignment. One of the nice features of McClelland and Kawamoto’s model is that it represents, in terms of microfeatures, a large number of properties and relationships over and above those
implicated in the pigeon-holing of objects into roles. Consider, for example, the model’s output following an input of ‘The boy broke the window’ (see McClelland and Kawamoto 1986 figure 2). In addition to indicating that the boy is the agent of breaking, the output pattern indicates a causal interaction instigated by the agent, a change in the state of the patient to one of being in pieces, agent/patient contact, partial agent motion and patient immobility. The model was also able to provide a contextually appropriate interpretation of ambiguous words (e.g. ‘bat’ in ‘The boy broke the vase with the bat’), to represent shades of meaning (e.g. the trained model interpreted balls which were instruments of breaking as being hard) and to generalize appropriately when presented with novel words. Because it learned to generate semantic representations that were both rich in detail and faithful to the meanings of the corresponding sentences, McClelland and Kawamoto’s model marked a major advance for natural language research.

3. Explicit semantics and modeller bias
Many worries have since been voiced about the explicit encoding of semantic information through the use of microfeatures (Lachter and Bever 1988, Harris 1990,
St. John and McClelland (1990, Miikkulainen 1993). Miikkulainen (1993: 280), for instance, complains: ‘Hand-coded representations are always more or less ad hoc and biased. In some cases it is possible to make the task trivial by a clever encoding of the input representations’. Lachter and Bever (1988) contend that just such a clever encoding has, in fact, taken place in the case of McClelland and Kawamoto’s sentence comprehension model. They argue as follows:

‘semantic role features’ . . . are just descriptors of roles themselves . . . Hence, any ‘learning’ that occurs is trivial. The learning does not involve isolating independently defined semantic features which are relevant to roles, but rather an accumulation of activation strengths from having role-features available, and being given correct instances of words (feature matrices) placed in particular role positions. (Lachter and Bever 1988: 222–223)

Lachter and Bever claim, in other words, that the model was trained on a corpus containing reliable indicators of role assignments (under the guise of microfeatures) and, as a result, was effectively tipped off to the fact that particular sentence constituents denote the fillers of particular roles. The presence of such role indicators would simplify the model’s task, because it would merely have to detect a few I/O regularities in order to learn how to handle a wide range of familiar and novel sentences.

What generally worries researchers about the explicit encoding of semantic information (e.g. through the use of microfeatures) is that modellers may, through their encoding activities, end up doing too much of a model’s work for it. That theoretical biases often creep into psychological models is, of course, an ineliminable, and even integral part of the modelling process. Yet, if these complaints about microfeatures are warranted, those who use this encoding strategy do so at the risk of biasing performance in a most undesirable manner.

4. Assigning roles without microfeatures
In order to alleviate worries about possible modeller biases (as well as to increase encoding efficiency), modellers have been relying more heavily on the ability of networks to form their own internal semantic representations over the course of training (see Harris 1990, St. John and McClelland 1990, Miikkulainen 1993).

St. John and McClelland’s (1990) Sentence Gestalt model provides a clear illustration of this trend.³ Like McClelland and Kawamoto, the goal that St. John and McClelland set for their model was to learn the complex determinants of role assignment. Other goals set for their model necessitated a more complex architecture than McClelland and Kawamoto’s. One such goal was to handle the sequential presentation of sentence constituents. This allows their model to generate expectations about subsequent inputs and to modify event representations as new evidence is encountered. Notice, for example, how the role filled by ‘hammer’ is uncertain in the partial utterance ‘The hammer broke—’ until followed by ‘into pieces’, ‘the window’, and so on. St. John and McClelland hoped to simulate even those major revisions of sentence interpretation known as ‘garden path effects’. These goals were realized by incorporating multiple hidden unit layers and through the recurrent use of hidden unit representations (figure 2).
Other advances over McClelland and Kawamoto’s model include the ability to represent prepositions, adverbs and auxiliary verbs (which enables the comprehension of passive sentences) and the representation of a larger set of roles at the output layer. While this adds to the repertoire of syntactic forms that the model is able to handle, it also means that the model must learn to respect a larger set of constraints. St. John and McClelland abandon the use of distributed semantic microfeature representations.
(for the reasons stated above) in favour of local input representations of sentence constituents and local output representations of both roles and fillers.\(^4\)

As discussed above, the meaning of a lexical item will sometimes determine what role is filled by the entity/entities that it picks out. Although the Sentence Gestalt model encodes lexical items locally, it is nevertheless able to rely upon the fact that certain properties of their referents are reflected in the training corpus by input/output regularities. Notice, for instance, how terms that denote humans should show up in similar sentence contexts and denote fillers of the same roles in those contexts. By picking up on such regularities, the Sentence Gestalt model is able to learn the semantic constraints governing role assignment—this, despite the fact that the explicit encoding of semantic information has, by design, been severely curtailed.

It is worth drawing attention to the fact that these input/output regularities play an essential role in the model’s capacity to generalize. If no such regularities were present in the training data (e.g. were it the case that any given noun phrase could designate an agent, patient, instrument, or modifier), the model would be unable to learn (save by rote memorization) the distinct role assignments corresponding to, for example, (6) and (7).

(6) The bus driver ate the spaghetti with the sauce.
(7) The bus driver ate the spaghetti with the knife.

And the Sentence Gestalt model does generalize appropriately (i.e. it does not rely upon mere rote learning). This accomplishment can be attributed to the fact that it represents semantically similar items in terms of similar weight vectors. This, in turn, is because the properties of lexical referents give rise to detectable I/O regularities.

The success of models that generate their own internal semantic representations naturally leads to speculation that semantic representation might be effected by the human cognitive system in a similar manner. Before passing judgement on this proposal, it will be helpful if we first reconsider the above-mentioned worries concerning inappropriate modeller biases in light of the difference between the two types of model considered thus far.

5. Statistical regularities and modeller bias

With regard to McClelland and Kawamoto’s model, Lachter and Bever worry that reliable indicators of role assignment have been provided under the guise of semantic microfeatures. One consideration whose full importance Lachter and Bever fail to appreciate, however, is that features only prove informative about role assignments provided other contextual features are present. For example, in (1) and (2) above, whether or not the subject denotes an animate object seems to be relevant to whether that item plays the role of agent or instrument. If, however, no grammatical object is specified, as in (3), then the subject may specify the patient. Yet this rule has its own exceptions, as is plainly shown by (8). In this case, the fact that there is no grammatical object is not a reliable indicator that the subject specifies a patient.

(8) The man ate.

What these examples suggest is that particular properties are only indicative of particular role assignments when they occur in the context of certain syntactic forms
and when the other terms present denote items with specific properties. But the properties in terms of which McClelland and Kawamoto represent lexical items are present in every context in which those lexical items appear. It is therefore up to the model to figure out, so to speak, which of the many features used to encode a given lexical item are relevant in which syntactic/semantic contexts. In other words, the extremely context-sensitive nature of role assignment precludes any one feature from functioning as the kind of reliable indicator that Lachter and Bever envision. Admittedly, McClelland and Kawamoto’s model does pick up on the fact that when a particular feature is present in a certain word position and certain other features are present in the other positions, then that feature is indicative of a certain role assignment. Yet this hardly squares with Lachter and Bever’s assessment that features ‘are just descriptors of roles themselves’.

Perhaps, however, there are other ways of making a model’s task easier, ways that have nothing to do with whether or not the properties of items are explicitly represented. Indeed, when we examine how structure is imposed on sentence/meaning pairs through the use of sentence frames (a popular technique for generating such I/O pairs), we see that insidious biases persist despite the shift away from microfeature representations. By examining another connectionist model of role assignment (i.e. Miikkulainen 1993), the structuring effect of sentence frames can be brought into sharp relief.5

Miikkulainen’s (1993) model is like the Sentence Gestalt model in so far as words are input sequentially and the explicit encoding of semantic information has been severely curtailed. Like McClelland and Kawamoto and St. John and McClelland, Miikkulainen uses a set of sentence frames in order to generate a corpus of sentence/role assignment pairs. Sentence frames are constructed around a particular action, and licensed combinations of roles and fillers are specified for that action. Miikkulainen’s frames (much like McClelland and Kawamoto’s) incorporate placeholders that specify the kind of item (e.g. animal, human, utensil) that can fill a given role in a particular type of event. For example, one sentence frame used by Miikkulainen was the following:

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Corresponding roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>The <strong>human</strong> ate the <strong>food</strong> with the <strong>utensil</strong>.</td>
<td>Agent, Patient, Instrument (respectively)</td>
</tr>
</tbody>
</table>

Frames such as this one can be used to generate I/O pairs by substituting for each placeholder (italicized) an item from the corresponding category and assigning roles accordingly.6

Some of the lexical items used by Miikkulainen were members of multiple categories. For example, ‘fork’ was a member of the categories utensil and object. Furthermore, certain sets of items were members of all and only the same categories. For example, ‘spoon’ was a member of exactly the same categories as ‘fork’. Terms with identical category memberships occurred in all and only the same sentence contexts and filled all and only the same roles within those contexts. There were also a few terms that were members of a unique set of categories. For example, ‘chicken’ was the only item that was a member of the categories food and prey. Over the course of training, Miikkulainen’s model came to represent items that shared identical category memberships with nearly identical activation vectors. Items such as ‘chicken’, which had a unique set of category memberships, were represented in terms of a unique activation vector.
As with the Sentence Gestalt model, the activation vectors generated by Miikkulainen’s model seem to encode information about the properties of items—this, despite the fact that those properties were never explicitly encoded. Also like the Sentence Gestalt model, Miikkulainen’s model generalizes appropriately (i.e. it does not engage in mere rote learning)—an accomplishment that can be attributed to the fact that it represents semantically similar items with similar weight vectors. For instance, because the vector representations for each utensil are nearly identical, in so far as the model knows that ‘spoon’ designates an instrument in a particular context it will know that ‘fork’ does as well.

Clearly, however, we have a strong reason to suspect that the categorization that Miikkulainen’s model effected in order to perform the mapping task is an artifact of the sentence frame technique. After all, the trained model classifies terms in a manner that neatly reflects Miikkulainen’s own categorization scheme. Given the method of data generation, this is to be expected. A sentence frame specifies the roles filled by all exemplars of the categories it incorporates, and items with identical category memberships will always fill the same roles. By picking up on the resulting I/O regularities, the model learns the categorization scheme used to create the data and, thereby, learns to process a wide variety of familiar and novel sentences. In other words, by using this popular data-generation technique, Miikkulainen infused the data with regularities of the sort that guide learning and facilitate appropriate generalization. This seems to be a case in which too much of a model’s work has been done for it beforehand—a case, in other words, of inappropriate modeller bias.

Lachter and Bever were worried that microfeatures supply reliable indicators of role assignment which simplify the task for a model. That is to say, they were worried that models merely have to pick up on some modeller-supplied I/O regularities in order to learn how to handle a wide range of familiar and novel sentences. This worry, we saw, was misguided. What the present case makes clear, however, is that one can structure data in a manner that facilitates learning and generalization without the use of microfeatures. In this case, it is not the explicit encoding of item properties that makes the model’s task easier, rather it is the structure that is imposed on data through the use of sentence frames.

For a slightly different take on the matter, recall that, in order to succeed at role assignment, a model has to learn that certain properties of lexical referents bear on role assignment in certain contexts. Admittedly, Miikkulainen’s model did seem to classify items into categories based upon properties that seem salient to role assignment. Imagine, however, that someone were to claim that, because the model came to represent two items with very similar vectors, this provides an independent justification for the claim that those items share properties that are salient to role assignment. Surely Miikkulainen’s model possesses no such corroborative powers. The model merely picks up on the antecedent categorization effected by the modeller. Yet, as will be explained presently, connectionist models can and do have this corroborative potential. Clarifying this fact serves to highlight further the insidious biases that result from the use of sentence frames, and it also suggests a way to avoid them.

6. Getting over sentence frames
One alternative to sentence frames is the overgenerate-and-screen strategy. One begins, quite simply, with a number of substitution instances of one or more highly
schematic sentence forms (e.g. NP VP, NP VP [NP], NP VP [PP [NP]], etc.). The sentences generated in this way are then screened for grammaticality and intelligibility (see Elman 1991, Harris 1990).

We can, for example, generate the following substitution instances of the sentence form NP VP NP PP:

(9) The dog ate the ball to the window.
(10) The fork hit the curtain on cheese.
(11) The man moved the rock with the hammer.
(12) Pasta broke the girl with sauce.

Of these, only (11) is both grammatical and intelligible. Harris (1990) uses a similar sentence-generation technique in order to model the various contextually appropriate meanings of ‘over’.

According to standard cognitive linguistic analyses, ‘over’ can have any of several different meanings, depending upon the context in which it occurs (Lakoff 1987). Consider the different senses of ‘over’ evoked in the following sentences:

(13) The plane flies over the bridge.
(14) The helicopter hovers over the city.
(15) The carpet stretches over the wall.

In (13), ‘over’ seems to denote the path taken by a trajector (the plane) above and across a landmark (the bridge). In (14), ‘over’ simply means the trajector is above the landmark. The meaning of ‘over’ in (15) is used to specify the covering of the landmark by the trajector.

Linguistic analyses have revealed a number of semantic considerations that seem to bear on the meaning of ‘over’ in particular contexts. Important properties seem to include whether trajectors are two-dimensional (e.g. carpet and fog), one-dimensional (e.g. road or cable), or zero-dimensional (e.g. helicopter or bird); whether or not verbs specify trajector/landmark contact and/or trajector motion; the relative height of landmarks; and whether or not they provide a surface that can support a trajector.

Some of the goals Harris set for her model coincide with those set by St. John and McClelland for their own Sentence Gestalt model. In particular, Harris’s model was charged with: (a) learning to map ordered sentence constituents on to a particular aspect of sentence meaning (in this case, ‘over’-interpretation); and (b) extracting the meanings of locally encoded sentence constituents from their occurrence in various sentence/meaning contexts. Harris’s model is a four-layer, non-recurrent, feed-forward network with three sets of input units and a single set of output units (figure 3). As input it takes sentences of the form NP VP PP, and as output it delivers local encodings of the appropriate sense of ‘over’.

Harris was also interested in determining whether or not the trained model’s internal semantic representations of lexical items would reflect those properties (e.g. dimensionality, height, etc.) thought by linguists to bear on ‘over’-interpretation. It was thus necessary for Harris to generate sentence/meaning pairs in such a way that certain dimensions of lexical meaning would be reflected by input/output regularities while at the same time minimizing the influence of her own theoretical leanings. Harris used the schematic sentence form ‘trajector verb over landmark’ and a lexicon of...
possible trajectors, verbs and landmarks in order to generate a set of 8100 sentences. Of these, only 1600 were deemed to be semantically and grammatically unexceptional. For each of the sentences that remained, Harris determined which sense of ‘over’ was appropriate (for details, see Harris 1990: 23–24)—that is, like McClelland and Kawamoto, St. John and McClelland and Miikkulainen, she drew upon her own intuitions concerning the meanings of sentences.

After training, it was found that the model’s internal representations could be interpreted as categorizing lexical items based upon some of those dimensions of semantic variation thought by linguists to bear on ‘over’-interpretation. What is, in fact, more interesting than this particular finding is what it teaches us about proper data-generation techniques.

Like Miikkulainen’s model, Harris’s model has no direct access to the properties of lexical referents. Therefore, in so far as certain of those properties constrain ‘over’-interpretation, the model has to rely on the fact that such properties engender regularities in the training data. For instance, if standard analyses are correct, dimensionality is a property of trajectors that bears on ‘over’-interpretation in certain contexts. Because of this fact (presuming that it is a fact), there will be regularities that manifest across a suitably large set of sentence/‘over’-interpretation pairs.

Notice that Harris could have generated input/output pairs with the help of sentence frames such as:

<table>
<thead>
<tr>
<th>trajector</th>
<th>verb</th>
<th>landmark</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-dimensional</td>
<td>b/t contact</td>
<td>over</td>
<td>covering</td>
</tr>
</tbody>
</table>

But had she used such frames, and had the model subsequently learned to classify certain items as two-dimensional, others as trajector-supporting, and so on (much as
Miikkulainen’s model learned to classify certain items as utensils, others as food, etc., this would hardly count as corroborating an independent linguistic analysis.

Instead, Harris was able, by using the overgenerate-and-screen strategy, to minimize the effect of her own theoretical leanings. Whatever regularities were present in the data resulted from simple yes/no judgements concerning intelligibility and grammaticality and, for those sentences surviving the cut, intuitions about the meaning of ‘over’ on a sentence-by-sentence basis. Thus, when the model learned to classify items in terms of properties like dimensionality, we could be reasonably certain that the model picked up on the actual semantic constraints governing ‘over’-interpretation rather Harris’s theory about those semantic constraints.

The overgenerate-and-screen strategy may not be the only way to generate a corpus of sentences in a manner that minimizes the influence of a modeller’s prior theoretical leanings. What we are concerned with minimizing, in particular, are the effects of a modeller’s preconceptions concerning the semantic factors that are relevant to a given type of mapping. An alternative means of satisfying this condition would be to utilize naturally occurring discourse, such as informal electronic conversations (e.g. chat room conversations) (Burgess and Lund 1998). However, it will, in so far as one’s goal is to model comprehension, probably be necessary to cull such discourse for those sentences containing just the terms comprising a limited lexicon. One way to facilitate this screening process would be to start by harvesting discourse that is already restricted to a specific content area and tailoring one’s lexicon accordingly.⁸

Whatever strategy one opts for, what we should like to find is that the regularities present in a given corpus of sentences are not the product of a modeller’s preconceptions concerning the semantic factors that are salient to the mapping task under consideration. In other words, we should like to be able to say that the semantic classification of items effected by a given model constitutes a reasonable source of evidence that the dimensions of semantic variance reflected in that scheme of classification really are salient to the corresponding sentence/semantics mapping, whether it be a sentence/role assignment mapping, a sentence/‘over’-interpretation mapping, or what have you. Such evidence, as already noted, may corroborate (or falsify) independent hypotheses concerning the semantic constraints that govern such mappings.

It seems likely that the Sentence Gestalt model and Miikkulainen’s model would, had they been trained on data created in accordance with one of the techniques just described, also have learned to classify lexical items based upon the salient (i.e. to role assignment) properties of their referents. It is therefore worth revisiting the suggestion that these models, which must generate their own internal semantic representations, might shed some light on the nature of semantic representation in humans.

7. Output representations: The indispensability of intuition?

As a preliminary, we should bear in mind that the explicit representation of semantic information has by no means been eliminated in these models. Although the properties of the lexical referents are not explicitly represented, these models still explicitly encode, at the output layer, such relational properties as who is doing what to whom (role assignment) and where one object is situated relative to another over time (‘over’-interpretation). Indeed, if we are correct in viewing comprehension as a process that involves a mapping between natural language representations and
semantic representations, then the very nature of the problem seems to mandate the explicit encoding of some semantic information at the output layer.

At present, it may be that the only way to meet this demand is to rely upon human semantic intuitions. Specifically, modellers must decide what properties and relationships are to be represented and devise a corresponding encoding strategy. From there, the meanings of particular sentences (relative to the chosen properties and relationships) must be settled upon and appropriate vector representations generated.

It has been suggested that a possible alternative to the reliance upon human intuitions would be to utilize statistical techniques such as discrete multi-dimensional scaling (see Clouse and Cottrell 1996) and latent semantic analysis (LSA) (Landauer and Dumais 1997, also see Kintsch 1998). These techniques convert into vector representations certain aspects of meaning that show up as statistical regularities across a set of natural language sentences (not to be confused with the regularities that show up across a set of sentence/meaning pairs). LSA vectors, for instance, are informative about the semantic relatedness of words, or even about word–sentence relatedness, sentence–sentence relatedness, word–paragraph relatedness, and so on. To see how LSA works, notice that if one were to examine those paragraphs of an encyclopedia containing the word ‘termite’, one would probably find that they contain the word ‘tree’ more often than the word ‘toothbrush’. This regularity, or co-occurrence effect, is a consequence of the meanings of the terms involved, and it can be detected through LSA. What LSA delivers is a set of vectors (corresponding to words or larger grammatical units) that collectively represent facts about semantic relatedness in terms of vector similarity/dissimilarity (i.e. $n$-dimensional proximity). Because such representations as LSA vectors carry a great deal of information about semantic relatedness, they are quite useful for certain modelling tasks (e.g. for modelling priming effect phenomena). However, while co-occurrence vectors display some sensitivity to syntax (e.g. the partitioning of $n$-dimensional space may reflect a distinction between nouns and verbs) and may even effect a categorization of terms on the basis of the properties of their referents (Burgess and Lund 1998), these vectors seem not to constitute the kinds of semantic representations that would be ideal for modelling sentence comprehension. In particular, representations of semantic relatedness do not suffice to capture information about the many spatial, causal and functional relationships that the objects described in a given sentence bear to one another. That is to say, LSA vectors are not well suited for representing such aspects of meaning as who is doing what to whom (role assignment) and where one object is situated relative to another over time (‘over’-interpretation).

Ultimately, we should like to find that the state of the art is such that perceptual mechanisms are able to ground, directly or indirectly, the contents of the semantic representations harboured by models of comprehension, thus effectively freeing modellers from the responsibility of generating semantic representations. Until such a time, however, there may be no alternative but to rely, at least in part, upon human semantic intuitions. This is not to suggest that modellers should bear the entire burden by themselves. One possible method for determining what objects, properties and relationships to represent would be to base these decisions on the testimony of theoretically-disinterested subjects. Such a strategy would be reminiscent of the one used by McRae et al. (1997) in order to generate feature representations for individual words, the crucial difference being that feature lists for entire events would be generated. For a very large corpus of sentence/meaning pairs, this technique might
ease the workload on modellers. It should, at the very least, contribute to an overall sense of propriety.

8. **The limitations of internal semantics**

Even granting that a connectionist model of comprehension will need to be supplied with some explicit semantic information at the output layer, it nevertheless remains to be determined how much semantic information is appropriate. Perhaps those models of role assignment forced to create their own internal representations of lexical meanings achieve the proper balance. It might be argued, for instance, that while modeller-specified representations of role assignments provide a schematic rendering of the manner in which various objects are related to one another, the remaining details of a given event can be filled in with the help of a model’s internal semantic representations of lexical items. However, even granting that important information concerning the properties of lexical referents is carried by the internal representations that result from exposure to a large corpus of sentence/role assignment pairs, these representations are ill-equipped to fill in all of the requisite details.

As explained earlier, lexical items whose referents share certain properties tend to occur in similar sentence contexts and denote the fillers of the same roles in the events described by those sentences. An appropriately configured connectionist system (viz. one that has too few hidden units to learn by rote) will learn the mapping task by picking up on such regularities and, as a result, those aspects of meaning that bear on role assignment will find their way into that model’s vector representations.

Nevertheless, it needs to be recognized that *only* those aspects of meaning that bear on role assignment will make their way into such a model’s internal vector representations. There are, however, a great many more properties that should be represented by any adequate model of comprehension. To be sure, a viable model of sentence comprehension ought to be capable of representing the fact that both ‘spaghetti’ and ‘steak’ are food, and (arguably) models forced to generate their own internal representations of lexical meanings are capable of this feat. But there are also many properties that distinguish spaghetti from steak. These distinguishing properties will, moreover, give rise to further differences at the level of sentential semantics—differences, for example, in terms of the how certain activities, like cooking and eating, are carried out with respect to the two types of food. It seems reasonable, for instance, to demand of models of comprehension that they not only represent whatever overlap there is between the meanings of (6) and (16), but also the differences.

(6) The bus driver ate the spaghetti with the sauce.
(16) The bus driver ate the steak with the onions.

A model will be unable to represent the relevant differences, however, if its access to semantic information is limited to role assignments and whatever properties happen to engender regularities across a corpus of sentence/role assignment pairs.

This limitation can, of course, be remedied by relying less on the ability of networks to form their own internal semantic representations and by making more semantic information explicit at the output layer. Many of the benefits of rich semantic output representations have, in fact, already been demonstrated by McClelland and Kawamoto’s model which, as noted above, did far more than learn to pigeon-hole
items into roles. To recap briefly, it was also able to represent numerous object properties over and above those that impose constraints on role assignment, a rich array of relational properties over and above role assignments themselves, the contextually appropriate meanings of ambiguous words, and certain contextually appropriate shades of meaning.

With regard, then, to the goal of supplying a thoroughgoing model of the comprehension process, it may be a mistake to rely too heavily on the ability of networks to learn the meanings of terms through exposure to sentence/role assignment pairs. If, on the other hand, one is interested in corroborating hypotheses concerning the semantic constraints governing a particular sentence/semantics mapping, then it clearly does make sense to curtail severely direct access to certain kinds of semantic information (viz. information concerning those properties hypothesized to impose constraints).

There is also the question of how to represent lexical inputs when one’s goal is to provide a thoroughgoing model of the comprehension process. To start with, notice that an interesting property of McClelland and Kawamoto’s model is that it is able to deal sensibly with novel terms (i.e. terms encoded on the basis of a conjunction of microfeatures that it has not yet encountered). This may be a desirable trait—after all, we humans seem to have a similar ability. If you were informed, for example, that ‘flad’ means ‘a tart Indonesian syrup’, you would have little trouble comprehending (17).

(17) The girl ate the bread with the flad.

Unlike McClelland and Kawamoto’s model, those models deprived (at the input layer) of access to information about the properties of lexical referents (e.g. the Sentence Gestalt model and Miikkulainen’s model) are unable to deal quite so sensibly with terms that they have not been trained on, though the context in which they occur will narrow down the space of possibilities.

Admittedly, it does, on the face of things, seem a bit unnatural to represent lexical items (as McClelland and Kawamoto do) in terms of microfeatures. This strategy might look a good deal more sensible, however, if one keeps in mind that McClelland and Kawamoto’s intent was merely to model the second stage of a hypothesized two-stage process. McClelland and Kawamoto envision the first stage as involving a mapping between orthographic (or phonological) representations and semantic representations. This mapping is supposed to be effected on a term-by-term basis and in a context-insensitive manner—thereby delivering representations of the meanings of individual sentence constituents. The proposal that lexical meanings are initially accessed in this context-insensitive manner does enjoy a certain amount of empirical backing (Swinney 1979, Rayner et al. 1994, Kintsch 1998), though the matter is far from settled. At any rate, from the standpoint of one concerned about the potential for inappropriate modeller biases, there is (as I have taken pains to establish) nothing inherently suspect about the explicit representation of semantic information.

9. Conclusion
The main point of this analysis has been to clarify the manner in which inappropriate biases can and cannot creep into connectionist models of natural language comprehension. As the trend of eschewing explicit semantic representations has in no small
measure been a reaction to the misguided concern that such representations are inherently biased, it is hoped that the present analysis will foster a renewed interest in models that are provided with explicit access to semantic information. Indeed, when a model of comprehension is deprived of such information, the internal representations that it generates will tend to reflect just those dimensions of semantic variance that impose constraints on the type of mapping under consideration. In other words, an over-reliance upon the ability of connectionist systems to generate their own semantic representations may come at the expense of their capacity to comprehend the full meanings of sentences.

Acknowledgements
I would like to thank to Bill Bechtel, Mark Rollins, Rick Grush, Gary Cottrell, Daniel Clouse, Mark St. John and an anonymous reviewer for their many helpful comments on earlier drafts of this manuscript.

Notes
1. See Harris for an excellent discussion of connectionist models of comprehension from the perspective of cognitive linguistics. It also is worth noting that connectionists have begun to model the reverse mapping as well—that is, from a semantic or perceptual medium to a linguistic one (Zlatev 1992, Nenov and Dyer 1994, Regier 1995).
2. I use the term ‘role’ here as a theory-neutral way of expressing the common thought behind such terms as ‘thematic case role’ (McClelland and Kawamoto 1986: 272), ‘theta role’ (Haegeman 1994: 49–55) and ‘role archetype’ (Langacker 1991: 236).
3. See p. 249.
4. There were also 13 feature units at the output layer. The presumption seems to have been that the feature units would enable the model to capture some contextually appropriate shades of meaning while not significantly affecting its capacity to perform the pigeon-holing task (St. John and McClelland 1990: 226).
5. Because St. John and McClelland make some events more probable than others, there are practical difficulties in isolating the effect of sentence frames on the vector representations generated by their model.
6. St. John and McClelland do not use such placeholders, rather they enumerate possible fillers (and their probabilities) within the sentence frames themselves. The introduction of statistical regularities seems to necessitate the use of a modified sentence-frame procedure such as theirs. By removing the statistical constraints and adding placeholders, the sample sentence frame (corresponding to the action hit) provided by St. John and McClelland (1990: 253–255) can be reduced from 75 lines to 11.
7. The use of placeholders actually generates more events than St. John and McClelland’s frames, because St. John and McClelland neglect many possible filler combinations. For instance, of the many items that might fall under the object placeholder in the ‘hit’ frame, only balls were hit by baseball bats.
8. As Harris (1990) notes, the form ‘trajectory verb over landmark’ is less schematic (i.e. more semantically determinant) than ‘NP VP (over NP)’. It might be argued that the use of the placeholders trajectory and landmark imposed certain theoretically-biased regularities on the training data. However, because this particular sentence form is still very schematic, the lexical items involved were (in terms of the properties of their referents) quite diverse and an overgenerate-and-screen strategy was used, there seems little cause for concern (see Harris 1990: 18–19 for details).
9. Mark St. John suggested this very clever strategy.
10. Although there are ways (discussed later) of easing the burden on modelers and of alleviating any lingering worries about modeler biases.
11. Because McClelland and Kawamoto use sentence frames, their model’s achievements have been tainted to some extent. Keep in mind, however, that the blame does not rest with explicit encoding of semantic information.
12. The ability to represent shades of meaning is an especially noteworthy accomplishment, because the meaning of a sentence is often more than a simple concatenation of individual lexical meanings. Consider, for instance, the different meanings of ‘chicken’ in the following sentences (see also McClelland and Kawamoto 1986: 314–315): (a) The woman ate the chicken. (b) The wolf ate the chicken. (c) The chicken lay dead at the side of the road. For some experimental results that seem to support of this view, see Barsalou et al. (1999).
References
Harris, C. L., 1990, Connectionism and cognitive linguistics. Connection Science, 2: 7–33.