Opinion TRENDS in Cognitive Sciences Vol.8 No.7 July 2004

Full text provided by www.sciencedirect.com

An alternative view of the mental lexicon

Jeffrey L. Elman

Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92093-0515, USA

An essential aspect of knowing language is knowing the words of that language. This knowledge is usually thought to reside in the mental lexicon, a kind of dictionary that contains information regarding a word's meaning, pronunciation, syntactic characteristics, and so on. In this article, a very different view is presented. In this view, words are understood as stimuli that operate directly on mental states. The phonological, syntactic and semantic properties of a word are revealed by the effects it has on those states.

> For a first approximation, the lexicon is the store of words in long-term memory from which the grammar constructs phrases and sentences.
>
> Ray Jackendoff [1]

> My approach suggests that comprehension, like perception, should be likened to Hebb's (1949) paleontologist, who uses his beliefs and knowledge about dinosaurs in conjunction with the clues provided by the bone fragments available to construct a full-fledged model of the original. In this case the words spoken and the actions taken by the speaker are likened to the clues of the paleontologist, and the dinosaur, to the meaning conveyed through these clues.
>
> David Rumelhart [2]

What does a word mean? The usual answer assumes that, whatever that meaning is – and there is considerable debate about this – it mostly resides in what is called the mental lexicon. Determining meaning thus involves, at the very least and as a first step, retrieving that meaning from the lexicon. Here, I describe another possibility, which involves a very different view of word meaning and of the lexicon. It is a view very similar to the one proposed by David Rumelhart 25 years ago, in which words don't *have* meaning as much as they provide *clues* to meaning.

In early generative theories, the lexicon was not considered to be a particularly interesting place. Rules were where the action lay. In the past two decades, however, the lexicon has come into its own [1,3–5], with some theories going so far as to place even grammatical patterns in the lexicon. Similar trends have occurred in psycholinguistics [6–8], accounts of child language acquisition [9–11], and in many connectionist models of learning [12].

What does the lexicon actually look like? The metaphor of lexicon-as-dictionary is inviting, and has led to what Pustejovsky [13] has called the 'sense enumeration model'. In that view, a lexical entry is a list of information. Just what information actually goes into the lexicon is a matter of some debate [1,14–16], although most linguists agree that lexical entries contain information regarding a word's semantic, syntactic and phonological properties. Some accounts of language processing have argued that lexical entries contain very fine-grained information about, for example, grammatical usage [6,7]. The common thread in the vast majority of linguistic theories is to see the lexicon as a type of passive data structure that resides in long-term memory.

But one can imagine thinking about words, what they mean, and how they are stored, in a very different way. Rather than putting word knowledge into a passive storage (which then entails mechanisms by which that knowledge can be 'accessed', 'retrieved', 'integrated', etc.), words might be thought of in the same way that one thinks of other kinds of sensory stimuli: they act directly on mental states. This by no means is to deny that the nature of this interaction is complex or systematic. Indeed, it is in the precise nature of their causal effects that the specific properties of words – phonological, syntactic, semantic, pragmatic, and so forth – are revealed.

To make this proposal a bit more tangible, let me offer what can serve as a working metaphor for how word knowledge might be instantiated through the *word as 'operator'* rather than *as 'operand'*. The metaphor involves a particular connectionist model, the Simple Recurrent Network (SRN; Figure 1) [17], but similar behaviors are found in the broader class of networks that have dynamical properties (see, for example, [18,19]). Thus, the role of the SRN in the following discussion is merely to help illustrate the alternative view proposed here rather than advance it as the model of choice. Finally, it should be noted that many core ideas in what follows have been anticipated in earlier work (e.g. the Sentence Gestalt Model [18]). The goal here is to distill and present these proposals in as consise and clear a form as possible.

Computation through dynamics

The SRN (Figure 1) is designed to process events and behaviors that unfold over time. The key to the architecture is recurrence. Outputs are driven by internal states (patterns across the hidden-unit layer). These internal states are themselves the product of an external input

Corresponding author: Jeffrey L. Elman (jelman@ucsd.edu).
Available online 4 June 2004

www.sciencedirect.com 1364-6613/$ - see front matter © 2004 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2004.05.003

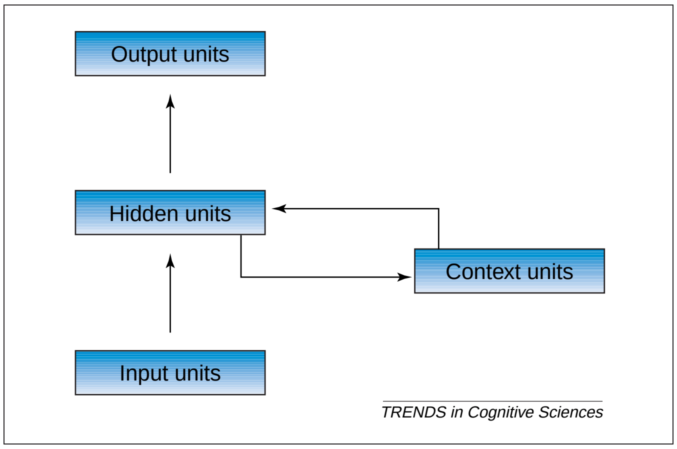

Figure 1. Simple Recurrent Network. Each layer is composed of one or more units. Information flows from input to hidden to output layers. In addition, at every time step t, the hidden-unit layer receives input from the context layer, which stores the hidden-unit activations from time t−1.

(on any given time step; it could be nil), as well as – crucially – the network's own previous internal state, cycled through context units. Counter-intuitively, even though only the last internal state is accessible to the network, the memory capacity is potentially very powerful. The previous hidden-unit state is itself affected by the hidden-unit state before that, and so on. Thus, it is possible for the network to learn to retain information over many (in principle, an infinite number) of time steps.

Learning to predict

For a variety of reasons, prediction has been a useful task in training SRNs:

(i) Prediction of the next word in a sentence fragment cannot rely solely on linear relationships [20]. The network is therefore forced to discover the more abstract relationships between constituents.

(ii) No special oracle is required for training. The success or failure of prediction can be verified when the next item is processed. As a model of human learning, all that is required is the ability to compare what was anticipated with what actually occurs.

(iii) Although language is clearly not mainly about prediction, there is good evidence that expectancy generation plays a role in language comprehension (e.g. [21,22]). Prediction thus has psychological plausibility.

Analyzing the network's 'mental space'

In one study, an SRN was trained on the predict-the-next-word task, given a corpus of sentences that had been generated by an artificial (but natural-like) grammar [17]. Each word was assigned a unique vector consisting of 0s and a single 1. This so-called 'localist' representation deliberately deprives the network of any information about grammatical category, meaning, inflection, and so on. (This is obviously a worst case scenario. In real life, morphology provides important cues for all of these things.) The network then saw a succession of sentences, each presented one word at a time. As a new word was entered, the network's task was to predict the next word. Rather than memorizing the corpus, the network learned to predict, in a context-appropriate manner, all the words that were grammatically possible, with activation levels corresponding to the probability that each word might occur.

This discovery of these classes of network is interesting, because the information predicted is not carried by the form of the input. Rather, the network used distributional information to induce categories such as Noun, Verb, or Animate. Analysis of the network's hidden-unit patterns revealed that the patterns evoked in response to each word did indeed reflect the word's category membership. This can be seen in the hierarchical clustering diagram shown in (Figure 2), which shows the similarity structure of the words' hidden-unit activations, measured in terms of Euclidean distance in the hidden-unit space.

In the past 10 years, there have been several empirical studies (not involving neural networks) that demonstrate that distributional information that could be used to induce categories does in fact exist in speech addressed to young children [23–25]. More direct confirmation that infants are sensitive to the statistical structure of the language around them comes from research in which infants learn patterns presented to them in artificial languages [26–30].

More intuitively, one can understand the SRN's behavior in terms of the hidden-unit vectors lying in the network's 'mental' space. A schematic of this, in three (of the 70 actual) dimensions is shown in Figure 3. The network has learned to partition this space into major categories (Noun, Verb), with each category subdivided into smaller regions corresponding to additional distinctions (among the Nouns: animate and inanimate, human and non-human, edible, etc; among the Verbs: transitive, intransitive, optionally transitive). Thus, the network has learned about important properties of each word, and uses these to generate expectancies about words that are grammatically possible successors. And grammar is undoubtedly important. In simulations involving complex sentences, SRNs have learned agreement relations that span multiple embeddings, long-distance filler-gap dependencies, and so on (see, for example, [31–33]).

Compare this with the traditional view, in which words are assumed to be tokens of the abstract entities contained in the lexicon. When we process a word, we know its meaning by virtue of its entry in the lexicon. But in the network, there is no lexicon, at least, not in the usual sense of the term. A better way of thinking about how this system works is that lexical knowledge is implicit in the effects that words have on internal states.

The mental states that are produced in response to words are clearly non-arbitrary. Thus, in Figure 3 the states reflect both grammatical as well as semantic properties of the words. This is what we mean when we say that a word's properties are revealed by its effects on internal states. Importantly, these effects are always and unavoidably modulated by context. Thus, words are only partial (but obviously very important) determinants of internal states.

What implications does this approach have for other theories?

Types versus tokens

The distinction between types and tokens is central in symbolic theories. The lexicon is said to instantiate 'types';

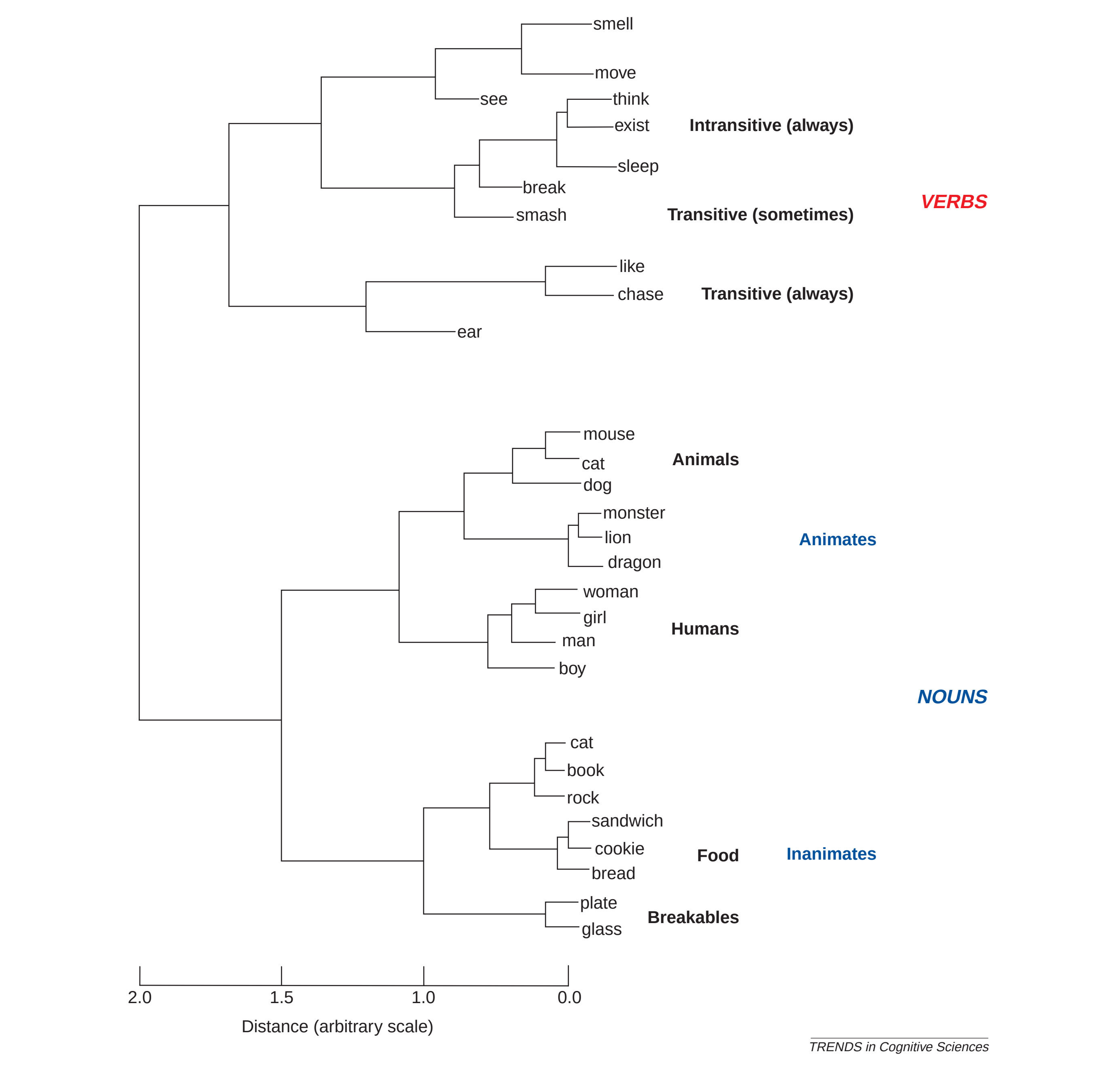

Figure 2. Hierarchical clustering diagram of hidden-unit activation patterns in response to different words. The similarity between words and groups of words is reflected in the tree structure; items that are closer are joined further down the tree (i.e. to the right as shown here).

that is, abstract representations that reflect what is known about words. The actual words that one processes in any given utterance are 'tokens' of that type. Thus, we recognize that in the sentence *The big boy picked on the little boy*, the two instances of *boy* refer to different individuals but that the individuals are instances of the same type.

In the SRN, this distinction might seem to be lost. Each occurrence of *boy* results in a state that is similar to all other occurrences, but not identical. Because the internal states also reflect prior context, the states produced by different tokens will differ slightly. 'Gotcha!' exclaim supporters of the traditional lexicon.

Not so fast. The internal states resulting from different instances of *boy* are indeed different. But, importantly, two additional things are true. First, every *boy* state inhabits a bounded region of state space inhabited only by other members of this lexeme. In Figure 3, for example, only a single instance of *boy* is shown but, in fact, every occurrence of the word will produce a state within the same bounded region. The various states differ slightly because they have occurred in different contexts but are clustered tightly together. Each region contains only tokens of the same type. The type *boy* is not explicitly represented, but this does not matter; the type membership of the token is easily recoverable from the fact that this state region is reserved only for tokens of the same type.

Second, there is a pay-off in the context-sensitivity of the token representations [32]. The variation in the state

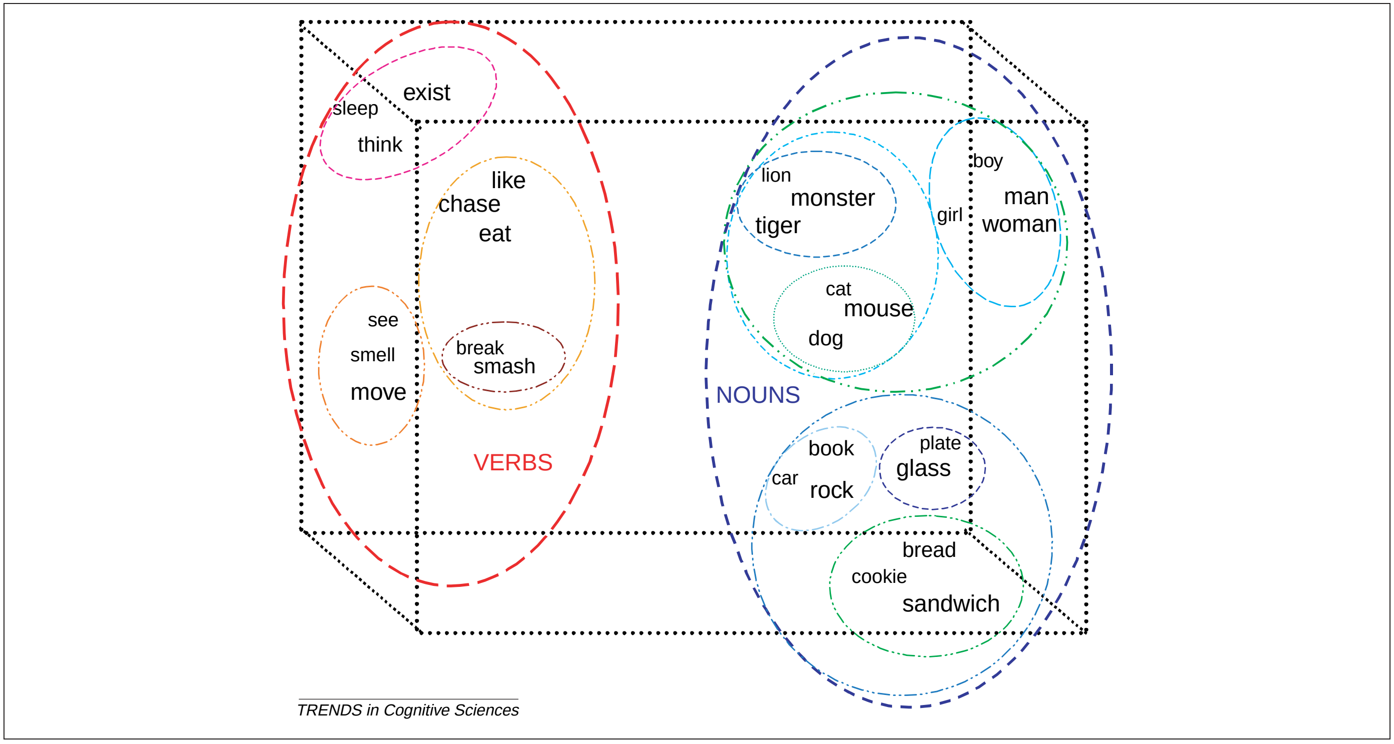

Figure 3. Schematic visualization, in 3D, of the high-dimensional state space described by the SRN's hidden-unit layer. The state space is partitioned into different regions that correspond to grammatical and semantic catgories. Nesting relationship in space (e.g. man within Animate within Noun categories) reflect hierarchical relationships between categories.

produced by the different tokens of a word type is not arbitrary, but reflects the context in a systematic and interpretable manner. Thus, perturbations along different axes in the state space might correspond, for example, to number, grammatical role (subject, object, etc.), inflection, and so forth. This gives us representations that simultaneously tell us: (i) which individual is picked out by the word; (ii) what type it is; and (iii) what properties the individual has by virtue of its context, grammatical role, and so on.

Categories

This same mechanism can be extended to account for the distinction between categories and their members. At the largest granularity of analysis, all nouns inhabit the same region of state space, and verbs another (Figure 3). Thus, the Noun and Verb categories are again implicit but recoverable. Successively smaller grains of analysis yield subordinate categories. This notion of categories as emergent from the location in a high-dimensional state space, in which at any given moment different dimensions might be attended to and others ignored, suggests that different viewing perspectives on that space might yield new categories. This is precisely the sort of categorization that Barsalou has found in subjects who are called upon to organize existing objects in novel ways [34].

Language acquisition

This category structure emerges over time as a result of learning, and requires a certain critical mass [35]. Early word knowledge tends to be fragmented. This has several important consequences. First, the rate of learning new words reflects the marginal cost of learning how the new word differs from those already known. The rate of acquiring new words is therefore initially slow, because there is no category structure to draw upon. Once a critical mass is achieved, though, the rate of acquisition accelerates and undergoes a 'vocabulary burst', as is often observed in children at around 18 months.

Second, once in place, the category structure supports generalization, in the sense that the network needs to see a word only a few times and in only one grammatical context to generalize its usage to novel contexts. This is because the new word, identified as belong to an existing category, inherits all the properties that are accorded that category.

Semantic combination

The meaning of word groupings, from phrases to sentences, clearly depends in important ways on the meaning of the constituent words. In the traditional view, the lexicon provides the meaning of the utterance parts, which are then combined to form the meaning of the whole. However, the relationship between the meaning of the whole and the parts is exceedingly complex. Idioms are an obvious and extreme example of this. But as many have noted, there are many productive constructions in which the meaning of the whole is greater than could be predicted from individual words and their syntactic relationship; for example '*One more beer and I'm leaving*'. Indeed, some have argued that such an outcome is the rule, not the exception.

Similarly, it has been observed that even the meaning of a single word can differ, depending on context [13,36], in ways that are not easily captured by appealing either to homonymy or polysemy. Thus, in (1a) below, *bake* simply means something like 'cause a change of state through heat', but in (1b), it means 'create/cause to come into being through heat'.

(1) a. Ray baked a potato.
b. Ray baked a cake.

The problem is fundamentally that sense and meaning are often, arguably, always, context-dependent. Traditional approaches in which the lexicon contains, at best, an enumeration of possible meanings invoke special machinery to do the integration. In the SRN, this happens for free, because the state produced by a word is always – by virtue of the architecture – context-sensitive. Figure 4 illustrates this by showing how the state of the network varies when the same verb, *run*, is preceded by different subjects.

From word to construction

Finally, we come to the important question: How might the approach here interface with knowledge of grammar? Some theorists have suggested that the lexicon should contain units for entities that are larger than words, such as constructions [1,37]. Others have suggested that constructions themselves emerge from word knowledge [9,10,38]. Does this model have any place for constructions?

Several issues arise in connection with this. One is whether mechanisms such as the SRN possess the requisite computational power to account for (for instance) the compositional structure that appears to underlie sentences. Although much remains to be understood, it has already been established that recurrent neural networks go well beyond the power of Finite State Automata [39–42], despite views to the contrary (e.g. [1,43]). What I want to focus on here is the way in which networks of this sort might provide an entry point into understanding how lexical and constructional knowledge are related.

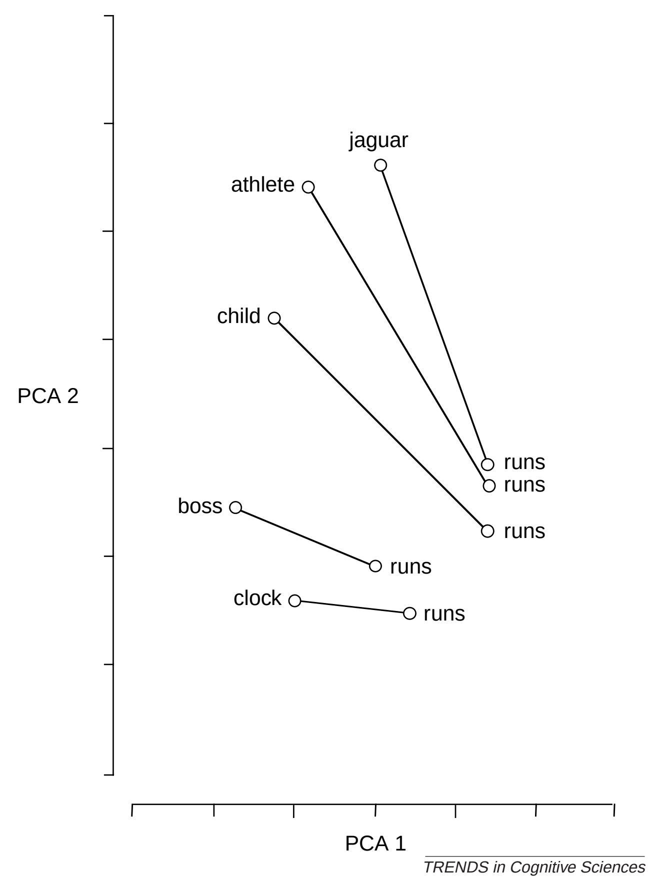

Figure 4. Trajectories through two dimensions of the SRN's state space while processing the verb 'run' preceded by different nouns. Differences in the location of the verb's state reflect systematic effects of context due to the noun. (PCA, principal component analysis.)

Clearly, part of knowing how to use a word involves knowing how it combines with other words. Verb knowledge, for instance, includes knowing which argument structures are possible, what grammatical categories the verb subcategorizes for, and what sorts of semantic restrictions are placed on arguments.

Valence properties do not reduce to a simple matter of 'X predicts that Y will follow'. In so-called filler-gap constructions – for example, '*Which book did you read?*' – the direct object precedes the verb that, in a simple declarative, it would follow. Or in sentences with embedded clauses, such as '*The movies John rents are scary*', the form of the verb *are* agrees numerically not with the closest noun, *John*, but with the more distant main clause subject, *movies*. The solution found by the SRN to such problems involves complex dynamics [39–42].

The solution itself develops over time. Early in learning, the SRN is conservative in the grammatical properties it attributes to words and sticks close to what it has seen. As the network encounters more examples, the same mechanism that gives rise to the discovery of Noun and Verb categories also operates to support generalization regarding more specific behaviors of classes within those categories; for example, verbs that cause motion, or that involve transfer of some entity. Thus, knowledge of constructions is a straightforward extension, by generalization, of knowledge of groups of words that behave similarly.

Interestingly, because the state space is continuous rather than discrete, the network is also able to retain information that can be highly word-specific. Thus, the network accommodates the need to know about both very general and abstract grammatical patterns, as well as the narrower requirements of particular verbs. The latter is important, because we know that comprehenders' knowledge of what makes a good filler of thematic roles can be very verb-specific [7]. In the extreme, a verb might permit only one entity to fill a role, as in *diagonalize* → *matrix* [44]. These role-filler preferences might be even more complex than commonly appreciated, because there is evidence that the filler of one role sometimes determines what is a good filler of another role. Thus, for the same verb, different agents might favor different patients [45]. Such three-way interactions (involving a potentially very large set of participants) are not easily captured in the lexical entry of a verb in the traditional lexicon.

Conclusion

The view advanced here is similar in some respects to several previous proposals. The notion of 'direct perception' is itself not new [46,47]. There is a close affinity with MacDonald and Christiansen's proposals regarding working memory [48], and the ideas presented here are also close in spirit and detail to those articulated earlier by McClelland, St John, and Taraban in their *Sentence Gestalt Model* [18].

The fundamental suggestion of the present proposal is to treat words as stimuli, whose 'meaning' lies in the causal effects they have on mental states. Or, to paraphrase Dave Rumelhart – words do not *have* meaning, they are *cues* to meaning. On the face of it, this might seem to demote the role of any given word in determining the meaning of utterances, but in fact it gives it far greater potential for interacting flexibly with other cues. Understanding the often systematic and sometimes idiosyncratic effects of these cues remains the challenge. It is here that computational models might help to lead us to more precise and formal theories.

Acknowledgements

This work was supported by grant MH6051701A3. Although they do not necessarily share the opinions expressed here, I am grateful for many discussions about this topic with Liz Bates, Mary Hare and Ken McRae. I owe a special debt to Jay McClelland. Our two decade long conversation about mental representations has been a source of much intellectual stimulation and great pleasure.

References

1 Jackendoff, R.S. (2002) *Foundations of Language: Brain, Meaning, Grammar, and Evolution*, Oxford University Press

2 Rumelhart, D.E. (1979) Some problems with the notion that words have literal meanings. In *Metaphor and Thought* (Ortony, A., ed.), pp. 71–82, Cambridge University Press

3 Bresnan, J. (2001) *Lexical–Functional Syntax*, Blackwell

4 Goldberg, A.E. (2003) Constructions: a new theoretical approach to language. *Trends Cogn. Sci.* 7, 219–224

5 Sag, I. *et al.* (2002) *Syntactic Theory: A Formal Introduction, Center for the Study of Language and Information*, Stanford University

6 MacDonald, M.C. (1997) Lexical representations and sentence processing: an introduction. *Lang. Cogn. Processes.* 12, 121–136

7 McRae, K. *et al.* (1998) Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *J. Mem. Lang.* 38, 283–312

8 Taraban, R. and McClelland, J.L. (1988) Constituent attachment and thematic role assignment in sentence processing: influences of content-based expectations. *J. Mem. Lang.* 27, 597–632

9 Bates, E. and Goodman, J.C. (1997) On the inseparability of grammar and the lexicon: evidence from acquisition, aphasia, and real-time processing. *Lang. Cogn. Processes.* 12, 507–584

10 Tomasello, M. (2002) The emergence of grammar in early child language. In *The Evolution of Language out of Prelanguage* (Givon, T. and Malle, B., eds), pp. 309–328, John Benjamins

11 Cameron-Faulkner, T. *et al.* (2003) A construction based analysis of child directed speech. *Cogn. Sci.* 27, 843–874

12 Christiansen, M.H. *et al.* (1999) Special issue Connectionist models of human language processing: progress and prospects. *Cogn. Sci.* 23, 415–415

13 Pustejovsky, J. (1996) *The Generative Lexicon*, MIT Press

14 Chomsky, N. (1986) *Knowledge of Language: Its Nature, Origin, and Use*, Praeger

15 Jackendoff, R.S. (1972) *Semantic Interpretation in Generative Grammar*, MIT Press

16 Rappaport Hovav, M. and Levin, B. (1998) Building verb meanings. In *The Projection of Arguments: Lexical and Compositional Factors* (Butt, M. and Geuder, W., eds), pp. 97–134, Center for the Study of Language and Information, Stanford University

17 Elman, J.L. (1990) Finding structure in time. *Cogn. Sci.* 14, 179–211

18 McClelland, J.L. *et al.* (1989) Sentence comprehension: a parallel distributed processing approach. *Lang. Cogn. Processes.* 4, 287–336

19 Tabor, W. and Tanenhaus, M.K. (2001) Dynamical systems for sentence processing. In *Connectionist Psycholinguistics* (Christiansen, M.H. and Chater, N., eds), pp. 177–211, Ablex Publishing

20 Chomsky, N. (1957) *Syntactic Structures*, Mouton, The Hague

21 Federmeier, K.D. and Kutas, M. (2001) Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 202–224

22 Kutas, M. and Hillyard, S.A. (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163

23 Brent, M.R. and Cartwright, T.A. (1996) Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93–125

24 Mintz, T.H. *et al.* (2002) The distributional structure of grammatical categories in speech to young children. *Cogn. Sci.* 26, 393–424

25 Redington, M. *et al.* (1998) Distributional information: a powerful cue for acquiring syntactic categories. *Cogn. Sci.* 22, 425–469

26 Gomez, R.L. and Gerken, L.A. (1999) Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition* 70, 109–135

27 Gomez, R.L. and Gerken, L.A. (2000) Infant artificial language learning and language acquisition. *Trends Cogn. Sci.* 4, 178–186

28 Gomez, R.L. (2002) Variability and detection of invariant structure. *Psychol. Sci.* 13, 431–436

29 Saffran, J.R. (2001) Words in a sea of sounds: the output of infant statistical learning. *Cognition* 81, 149–169

30 Saffran, J.R. (2001) The use of predictive dependencies in language learning. *J. Mem. Lang.* 44, 493

31 Christiansen, M.H. and Chater, N. (1999) Toward a connectionist model of recursion in human linguistic performance. *Cogn. Sci.* 23, 157–205

32 Elman, J.L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7, 195–224

33 Elman, J.L. (1995) Language as a dynamical system. In *Mind as Motion: Dynamical Perspectives on Behavior and Cognition* (Port, R. and van Gelder, T., eds), pp. 195–225, MIT Press

34 Barsalou, L.W. (1983) Adhoc ategories. *Mem. Cogn. 11*, 211–227

35 Elman, J.L. (1998) *Generalization, Simple Recurrent Networks, and the Emergence of Structure*, Erlbaum

36 Langacker, R.W. (1987) *Foundations of Cognitive Grammar*, Stanford University Press

37 Goldberg, A.E. (1995) *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press

38 Tomasello, M. (1992) *First Verbs: A Case Study of Early Grammatical Development*, Cambridge University Press

39 Boden, M. and Wiles, J. (2000) Context-free and context-sensitive dynamics in recurrent neural networks. *Connect. Sci.: J. Neural Comput. Artif. Intell. Cogn. Res.* 12, 197–210

40 Rodriguez, P. (2001) Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Comput.* 13, 2093–2118

41 Rodriguez, P. *et al.* (1999) A recurrent neural network that learns to count. *Connect. Sci.* 11, 5–40

42 Siegelmann, H.T. and Sontag, E.D. (1995) On the computational power of neural nets. *J. Comput. Syst. Sci.* 50, 132–150

43 Steedman, M. (1999) Connectionist sentence processing in perspective. *Cogn. Sci.* 23, 615–634

44 McCawley, J.D. (1968) The role of semantics in a grammar. In *Universals in Linguistic Theory* (Bach, E. and Harms, R.T., eds), pp. 124–169, Holt, Rinehart & Winston

45 Kamide, Y. *et al.* (2003) The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156

46 Gibson, E.J. (1969) *Principles of Perceptual Learning and Development*, Appleton Century Croft

47 Brookes, R.E. (1991) Intelligence without representation. *Artif. Intell.* 47, 139–159

48 MacDonald, M.C. and Christiansen, M.H. (2002) Reassessing working memory. Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54