Lexical Disambiguation Based on Distributed Representations of Context Frequency

Marshall R. Mayberry, III, and **Risto Miikkulainen**
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712
martym,risto@cs.utexas.edu

Abstract

A model for lexical disambiguation is presented that is based on combining the frequencies of past contexts of ambiguous words. The frequencies are encoded in the word representations and define the words' semantics. A Simple Recurrent Network (SRN) parser combines the context frequencies one word at a time, always producing the most likely interpretation of the current sentence at its output. This disambiguation process is most striking when the interpretation involves semantic flipping, that is, an alternation between two opposing meanings as more words are read in. The sense of `throwing a ball` alternates between `dance` and `baseball` as indicators such as the agent, location, and recipient are input. The SRN parser demonstrates how the context frequencies are dynamically combined to determine the interpretation of such sentences. We hypothesize that several other aspects of ambiguity resolution are based on similar mechanisms, and can be naturally approached from the distributed connectionist viewpoint.

Introduction

In understanding ambiguous sentences, humans seem to employ automatic and immediate lexical disambiguation mechanisms even when they are compelled to alternate between two or more senses of an ambiguous word. Consider the sentence

> `John put the pot in the dishwasher because the police were coming over for tea.`[1]

As a reader processes this sentence, he or she is inclined to interpret the word `pot` first as a cooking utensil, then as marijuana, and lastly again as a cooking utensil, more specifically a teapot. Yet this processing seems to occur at such a low level that the reader will hardly be aware that there was any conflict in his interpretation. There does not seem to be any conscious inferencing required, no moment's cogitation as might be observed if, say, the reader were instead asked to compute the product of two double-digit numbers.

The primary goal of the study reported in this paper was to model such automatic semantic flipping behavior using distributed connectionist networks. The connectionist paradigm offers explanatory and predictive power difficult to achieve using more conventional symbolic methods (Elman, 1991; Lange, 1992; St. John and McClelland, 1990; Seidenberg, 1993). We chose to use the Simple Recurrent Network parser architecture (SRN; Elman 1990, 1991; Miikkulainen 1993), which has in recent years become a standard tool in research into language comprehension in connectionism. Our choice was motivated by two concerns:

1. Although lexical disambiguation has been studied in dedicated connectionist architectures before (Kawamoto, 1988), our primary interest is in understanding the mechanisms of ambiguity resolution as an integrated part of the parsing task itself.
2. Aside from the basic assumption that connectionist models share mechanisms in common with actual human language comprehension facilities, the use of a standard model makes a minimal number of assumptions regarding the internal processes in lexical disambiguation.

A brief overview of recent research into lexical ambiguity will help put this task into perspective (for a more comprehensive review, see Simpson, 1984). Several models have been proposed to account for how ambiguities are resolved during reading. The three most prominent in recent years have been the context-dependent, the single-access, and the multiple-access model.

The context-dependent model (Glucksberg et al., 1986; Schvaneveldt et al., 1976) is based on the assumption that only one meaning of a word is activated at any given time, namely, the one most appropriate to the context in which the word occurs. The primary reason is that the context primes the meaning which is most applicable, while suppressing others.

The single access (or ordered-access) model (Forster and Bednall, 1976; Hogaboam and Perfetti, 1975; Simpson and Burgess, 1985) posits that only one active interpretation of an ambiguous sentence is maintained at any one time. If in the course of processing the sentence information is encountered that does not accord well with the active interpretation, then that interpretation is abandoned and a representation that accounts for the established information as well as for the current ambiguity is sought, most probably through backtracking to the point of ambiguity. The activation level of an interpretation is determined by the relative frequencies of the meanings of the word or words that are the source of the ambiguity. The search process for the appropriate meaning takes place serially, terminating when a fit is made, or retaining the most dominant meaning when no contextually relevant match can be found. In

[1] Adapted from (Lange and Dyer, 1989).

To appear in the *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, 1994.

the strongest statement of the model (Hogaboam and Perfetti, 1975), only the most dominant meaning of an ambiguous word is retrieved first, regardless of whether the context supports a subordinate meaning.

The multiple access model (Onifer and Swinney, 1981; Seidenberg et al., 1982; Tanenhaus et al., 1979), which is the most widely accepted, suggests that several interpretations may be actively maintained when ambiguous information is encountered. At a later time, when additional input allows resolving the ambiguity, only the appropriate interpretation is retained. However, not all of the interpretations may be maintained with equal activation levels. Rather, the strength of a particular activation would be proportional to the likelihood of that interpretation being the correct one. Unlike the single access model, in which a single meaning is sought and selected, the multiple access model claims that all meanings are activated simultaneously regardless of context, but the context later influences selection of the most appropriate one.

As is not unusual when aspects of behavior are supported by several more or less opposing models, refinements are proposed to include elements from several models. For example, recent research by Burgess and Simpson (1988) supports the multiple access model, but suggests that meaning frequencies determine which interpretations reach the recognition threshold first. The role of context is to select which of the meanings remains activated.

Our research confirms Simpson and Burgess's work computationally, but also suggests that semantic frequencies could play an even more fundamental role than previously allowed. This is especially likely at the lowest level of sentence comprehension in which disambiguation seems to occur automatically in humans. We conclude that

1. multiple activation levels are maintained simultaneously during the processing of a sentence, and
2. the various meanings of each word are activated to the degree that corresponds to the frequency with which that word has been associated to the previous words in the sentence in the past.

Lexical ambiguity resolution, then, is a matter of finding the most likely interpretation in the context of the processed sentence.

Below, our experimental method for studying the semantic flipping phenomenom is first explained, followed by a description of the training data and the parser architecture. A detailed discussion of the simulation results and their implications on the study of lexical ambiguity resolution and language comprehension in general concludes the paper.

Modeling semantic flipping

Our approach was to train a simple recurrent parser network to map a sequence of words to the case-role representation (Fillmore, 1968; Cook, 1989) of the sentence, and observe the evolution of the sentence representation during parsing an ambiguous sentence. The parser was trained on sentences generated from two basic sentence templates:

1. `The` *agent* `threw the ball in the` *location* `for the` *recipient.*
2. `The ball was thrown in the` *location* `for the` *recipient* `by the` *agent.*

The fixed words in the template are indicated by `courier` typeface. The *location*, *recipient*, and *agent* stand for slots to be filled by actual content words. Depending on these words, the sentences could be interpreted as statements about baseball (e.g. `The pitcher threw the ball in the ballbark for the fans`), dance (`The emcee threw the ball in the ballroom for the princess`) or something rather ambiguous (`The visitor threw the ball in the court for the victor`). The output of the parser is one of two possible case-role representations:

1. *agent* *act:*`tossed` *patient:*`baseball` *location* *recipient*
2. *agent* *act:*`hosted` *patient:*`dance` *location* *recipient.*

Which of these two representations were used in the output depended on how strongly the words occupying the *location*, *recipient*, and *agent* slots were associated with baseball and dance.

Feature	Set to 1 for words
1. Ball	`ball`, `baseball` and `dance`
2. Verb	`thrown`, `tossed` and `hosted`
3. Other	`the` and `was`
4. Preposition	`in`, `for`, and `by`
5. Location	the five *location* words and `in`
6. Recipient	the five *recipient* words and `for`
7. Agent	the five *agent* words and `by`
8. Sense	Graduated according to word sense

Table 1: **The word representation vectors.** The words in the lexicon were encoded by these eight features. The first seven components were set to either 0 or 1; the right column lists those words that had the value 1. The **Sense** feature was used to indicate the degree of association to which a word had the two senses of `throw` and `ball`.

Training data

There were a total of twenty-six words in the lexicon: five for each of the three slots, two interpretations of `throw` and `ball`, and seven fixed words for the input sentences. Each were given hand-coded representation vectors according to the eight features shown in table 1. The last component, **sense**, is particularly important: it indicates how strongly the word is to be associated with `tossed baseball` (0) and `hosted dance` (1).[2] Thus, if a word was given a sense of 0.25, it would be more strongly associated with baseball than dance. Tables 2 and 3 summarize the sense values assigned to each word.

[2]This representation strategy was chosen mostly because it allows easy encoding and decoding of the word sense. Distributed representations (Miikkulainen, 1993; van Gelder, 1989) could be used as well, but sense would then have to be represented as distances between vectors.

Location	Recipient	Agent	Sense
ballpark	fans	pitcher	0.00
stadium	press	coach	0.25
court	victor	visitor	0.50
clubroom	vips	diplomat	0.75
ballroom	princess	emcee	1.00

Table 2: **Content word senses.**

Fixed words	Sense
tossed, baseball (output only)	0.00
ball,thrown/threw,the,was,in,for,by	0.50
hosted, dance (output only)	1.00

Table 3: **Sense values for the fixed words.**

The sense of the entire sentence was then computed by averaging the sense values of the individual words. Since these values were graduated in fourths, averaging over the three content words (per sentence) would result in twelfths. Thus, each input sentence was repeated twelve times in the training corpus, with the two different case-role representations assigned in proportion of the sense value of the sentence. In this way, context frequency could be simulated.

For example, if the passive sentence template is instantiated with the words `clubroom` (sense: 0.75), `fans` (0.00), and `emcee` (1.00), the following sentences is obtained:

`The ball was thrown in the clubroom for the fans by the emcee.`

Averaging the sense values gives $\frac{7}{12}$, or 0.5833. Accordingly, this sentence was repeated twelve times in the training corpus, with seven of the sentences assigned to `hosted dance` at the output, and five to `tossed baseball`. Thus, in the experience of the parser, $58\frac{1}{3}\%$ of the contexts in which `ball`, `clubroom`, `fans`, and `emcee` appeared were associated with `hosted dance`, and the remaining with `tossed baseball`. Hence, the dance sense would be slightly more dominant in this context, and would be the preferred interpretation.

There are 125 possible combinations of the five words in the three categories. Each combination was used to instantiate the two sentence templates, giving a total of 250 sentences. Since each sentence is repeated 12 times, the training corpus is composed of 3000 sentences. These sentences comprise the contextual history of the ambiguous words `throw` and `ball`. Both active and passive constructions in the sentence templates were used to contrast whatever priming effects the words might have on the general sense of the sentence.

Network architecture

The parser network used in our experiments (figure 1) is a variation of the Simple Recurrent Network architecture (SRN; Elman 1990, 1991; Miikkulainen 1993), trained to map a sequence of input word representations into a static case-role representation of the sentence. The single input assembly consists of eight units, corresponding to the eight components in the word representation. The output layer is a concatenation of five word-representation assemblies, corresponding to the case-role assignment of the sentence.

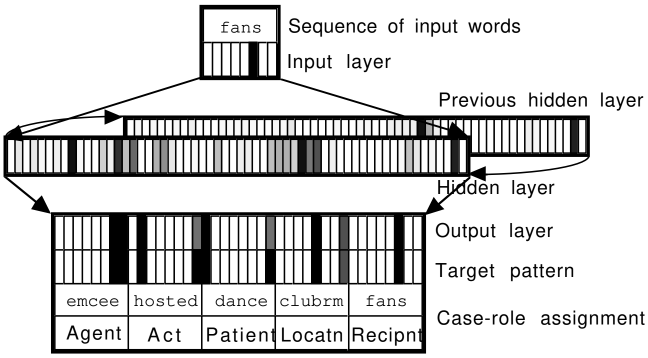

Figure 1: **The simple recurrent parser architecture.** The model consists of a simple recurrent network trained through backpropagation to map a sequence of input word representations into a case-role representation of the sentence.

At each step in the sequence, a word representation is loaded in the input assembly and the activity is propagated through the hidden layer to the output. The activity in the hidden layer (60 units wide) is saved in the previous-hidden-layer assembly, and used together with the word representation as input to the hidden layer in the next step. Throughout the sequence, the complete case-role assignment is used as the training target, and the error is propagated and the weights are changed (through the backpropagation algorithm) at each step.

In effect, the network is trained to shoot for the complete sentence interpretation from the first word on. As a result, it learns to indicate the current sense of the sentence in the **sense** components of the *act* and the *patient* assemblies at its output. If the current interpretation is predominantly `hosted dance`, these components have high values, and if it is `tossed baseball`, they have low values. A completely ambiguous interpretation is indicated by activation 0.5.

The parser was trained with 0.5 learning rate for 100 epochs, then 0.1 for 50 epochs, 0.05 for 5, and finally 0.01 until epoch 200. At this point, the average error per unit after reading a complete sentence was 0.024.

Results

The parser was tested with the same set of sentences used to train it to determine how well it captured the sense for each sentence.[3] The theoretically optimal values for the sense outputs were obtained from the training data based on the distribution of the words in the sentences. The **sense** outputs of the network were then compared to the theoretical values.

[3] Generalization was not tested in this study because tight control over the theoretical frequencies was desired, and because good generalization is common for this type of models and offers no new perspective on the problem being addressed.

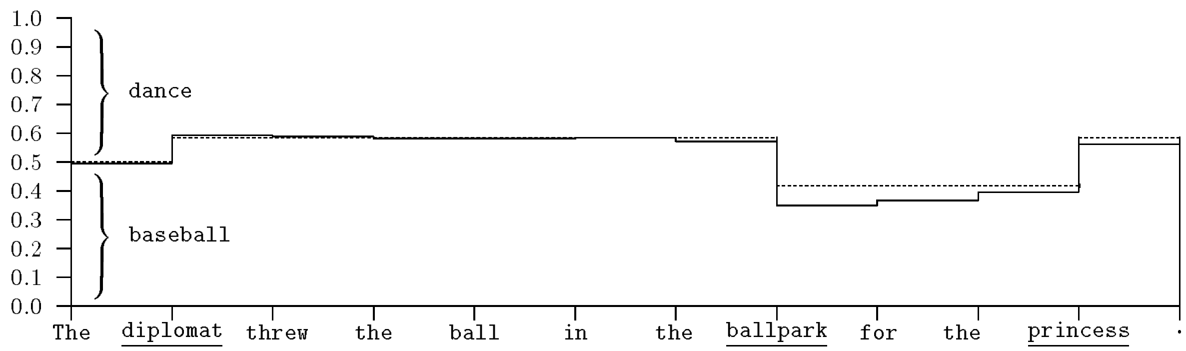

Figure 2: **Evolution of the interpretation of an active construction.** The dotted line represents the theoretical sense level during processing the sentence, and the solid line indicates the average of the two **sense** output units. The content words have been underlined. The average error per unit on this sentence was 0.0180.

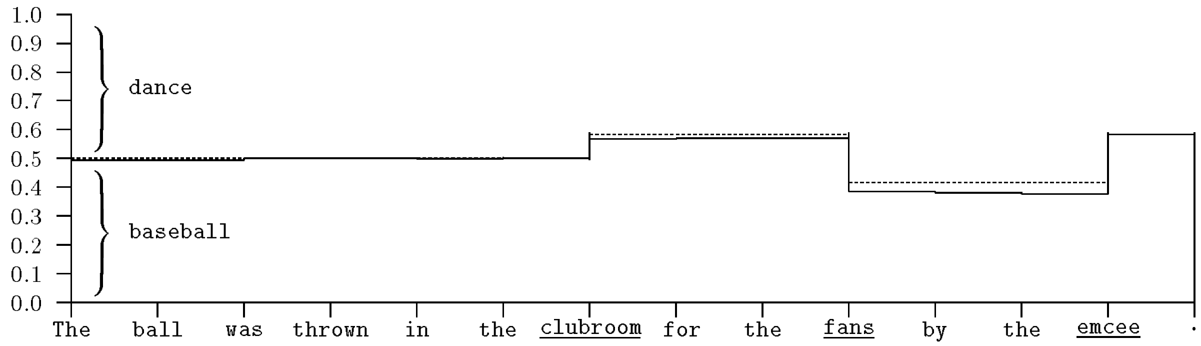

Figure 3: **Interpretation of a passive construction.** The average error per unit on this sentence was 0.0123.

The network had captured the sense frequencies very accurately: The average error across the entire data set was found to be 0.0114 (0.0122 for the active constructions and 0.0107 for the passive). Moreover, all sentences where at least one of the content words was associated to a sense opposite of that of the other content words resulted in semantic flipping behavior. Below, the processing of two sentences, one active and one passive, is analyzed in detail. These examples are particularly interesting because they require revising the semantic interpretation twice during processing.

In reading the active sentence

`The diplomat threw the ball in the ballpark for the princess`

(figure 2), the average of the two sense unit activations is initially very nearly 0.5, indicating no bias one way or the other (i.e. complete ambiguity) because the two senses of `ball` were equiprobable in the training set. After processing the word `diplomat`, the activation level rises to 0.5921 since a slight majority ($58\frac{1}{3}$%) of the sentences in the training set in which `diplomat` occurs have the sense `dance`. In effect, `diplomat` has a priming effect on the interpretation of the rest of the words. The activation remains at roughly this level until the word `ballpark` is encountered. At this point, the semantic bias *flips* to 0.3481 in favor of `baseball`, because in the experience of the parser, a majority ($58\frac{1}{3}$%) of the sentences in which both `diplomat` and `ballpark` appear have the sense of `baseball`. The activation stays below 0.5 until `princess` is read in as the last word, whereupon it flips back to 0.5610, indicating that the sentence is one again interpreted as `diplomat tossed baseball`. The theoretical expectation of those sentences containing the words `diplomat`, `ballpark` and `princess` is 0.5833, which is close to the activation level the parser finally settled upon.

Similarly, in processing the passive sentence

`The ball was thrown in the clubroom for the fans by the emcee`

(figure 3), after a long sequence of neutral fixed words, the network encounters `clubroom` and the interpretation becomes biased toward `dance`, because $58\frac{1}{3}$% of the training sentences with `clubroom` have this sense. Upon reading `fans`, the interpretation flips toward `baseball`, because now the majority of sentences (again $58\frac{1}{3}$%) with both `clubroom` and `fans` have the sense `baseball`. When the last word is read, the bias flips again back to `dance`, because a sentence with `clubroom, fans` and `emcee` has an overall sense average 0.5833.

The biases and flips in the sense values are not particularly dramatic because the frequency differences are fairly small in the training corpus. For a more stark contrast, these allocations could be adjusted; however, it is important to note that even such minor differences will result in reliable semantic revision behavior.

Discussion

The most salient effect observed was that the semantic sense of an input sentence as a whole varied as a function of the semantic senses of its component words. It is this variation that accounts for the flipping behavior that we set out to model. Let us speculate how this result could be interpreted in terms of human language comprehension.

A reader has experienced each word in a variety of contexts. Instead of regarding the word semantics as a collection of discrete and disjoint definitions in the lexicon as is commonly done in artificial intelligence, it is possible to view semantics simply as an encoding of all these past contexts. For most words, these contexts share much in common. For an ambiguous word, however, there are two or more distinctly different contexts, some of them more frequently observed than others.

In this view, a general mechanism emerges by which lexical disambiguation could proceed. As a reader processes a sentence, there is an interaction between the semantics (i.e. past contexts) of each word and the evolving interpretation of the current sentence. Each word primes the interpretation according to the frequency with which the word has been associated to the current context in the past. In the final interpretation, all the past contexts of its constituent words are combined. This view accords well with the multiple-access model of lexical disambiguation. All meanings of an ambiguous word are activated in the sense that they are an inherent part of the representation of the word. Which meaning reaches recognition threshold is affected by its past association with any other words in the sentence.

On the other hand, the final sentence context serves to reinforce the applicable semantics of its words, and provides additional context—and therefore additional semantics—for each word in the sentence. This way the word meanings continually adapt according to how they are used in the language.

It is important to note, however, that the frequency-based mechanism alone is insufficient to explain all of lexical disambiguation. Rather, it suggests how disambiguation might occur at its most basic, subconscious level alluded to in the introduction. This process should be distinguished from what can be called pragmatic disambiguation, which requires higher-level inferencing. Pragmatic disambiguation might be invoked if the semantics of the sentence comes into conflict with the larger context. Consider, for example, a variation of the sentence in the opening paragraph of this paper:

```
John took the pot out of the dishwasher
because the police were coming.
```

Even though it is unclear why John would take marijuana out of the dishwasher in this situation, because of the strong association of `police` with the marijuana meaning of `pot`, marijuana would be the dominant sense in the absence of other cues. However, what actually might be inferred from the above sentence, for example that John is trying to find a better hiding place for the marijuana, or that he wishes to make some tea for his guests, results from higher-level inferencing, or *pragmatic* disambiguation. The context in which the sentence appears would be used to decide the point. If this context was in conflict with the basic sense suggested by the frequency-based mechanism, then the appropriate sense would be decided perhaps again by the frequency-based mechanism, but now with the new context as additional input. However, if there was no conflict, then the basic sense would prevail with the reader unaware that a potential conflict even existed—the disambiguation would occur without a moment's cogitation, as modeled in our experiments.

The context-based semantics idea can be modeled particularly well in the distributed connectionist framework. In such systems, similar concepts have similar representations, and even when the history of previous contexts is not available, it is possible to simulate its effects. On the other hand, it is also possible to devise learning methods for automatically adjusting the representations according to the contexts (Miikkulainen, 1993; Miikkulainen and Dyer, 1991).

The mechanisms of frequency-based inferencing, we believe, are central to semantic representation in language comprehension in general. Despite the rich variety in which language can be used, there are multitudes of features that form the framework for its usage. These regularities are what are captured and represented in the mind—and modified continuously with experience—to make language comprehension possible. Although our research so far has focused on only one aspect of ambiguity resolution (the semantic context frequency), many other aspects of the problem can be viewed in a similar light. Whether it be syntactic, semantic, referential, or any of the many other aspects of ambiguity that is scrutinized, we hypothesize that they are all based on regularity and frequency, and the connectionist paradigm is particularly well-suited for accounting for them computationally.

Conclusion

Our model of semantic disambiguation is based on the following principles:

1. word usage determines word meaning,
2. ambiguity results when a given word is used in multiple ways,
3. past frequencies of the various connotations in relation to other words are encoded in a word's distributed representation, and
4. these frequencies are combined in the sentence parsing process to produce the most likely interpretation of the word sense.

Given these principles, the disambiguation process can be modeled by a standard neural network architecture (the simple recurrent network) as an integral part of the parsing process without special mechanisms. This result has important implications on the study of lexical ambiguity. Whereas the multiple access model of the disambiguation process posits simultaneous activation of all meanings of an ambiguous word, with the most frequent being the most dominant, and an *a posteriori* role for context in selecting the most appropriate meaning, the model presented in this paper proposes that the *meaning evolves from context*. Disambiguation occurs as meanings, or past contexts, of each word are combined. How the meanings could be gradually evolved as words are used in new contexts, and how this idea could be extended to other aspects of ambiguity resolution such as syntax and reference constitute the main directions of future work.

References

Burgess, C., and Simpson, G. B. (1988). Neuropsychology of lexical ambiguity resolution. In Small, S. L., Cottrell, G. W., and Tanenhaus, M. K., editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology & Artificial Intelligence*, 411–430. San Mateo, CA: Morgan Kaufmann.

Cook, W. A. (1989). *Case Grammar Theory*. Washington, DC: Georgetown University Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Fillmore, C. J. (1968). The case for case. In Bach, E., and Harms, R. T., editors, *Universals in Linguistic Theory*, 0–88. New York: Holt, Rinehart and Winston.

Forster, K. I., and Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory and Cognition*, 4:53–61.

Glucksberg, S., Kreutz, R. J., and Rho, S. (1986). Context can constrain lexical access: Implications for interactive models of language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:323–335.

Hogaboam, T. W., and Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 14:265–274.

Kawamoto, A. H. (1988). Distributed representations of ambiguous words and their resolution in a connectionist network. In Small, S. L., Cottrell, G. W., and Tanenhaus, M. K., editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology & Artificial Intelligence*, 195–288. San Mateo, CA: Morgan Kaufmann.

Lange, T. E. (1992). Lexical and pragmatic disambiguation and reinterpretation in connectionist networks. *International Journal of Man-Machine Studies*, 36:191–220.

Lange, T. E., and Dyer, M. G. (1989). High-level inferencing in a connectionist network. *Connection Science*, 1:181–217.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press.

Miikkulainen, R., and Dyer, M. G. (1991). Natural language processing with modular neural networks and distributed lexicon. *Cognitive Science*, 15:343–399.

Onifer, W., and Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory and Cognition*, 9:225–226.

Schvaneveldt, R. W., Meyer, D. E., and Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Child Development*, 48:612–616.

Seidenberg, M. S. (1993). A connectionist modeling approach to word recognition and dyslexia. *Psychological Science*, 4:299–304.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., and Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, 14:489–537.

Simpson, G. B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, 96:316–340.

Simpson, G. B., and Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance*, 11:28–39.

St. John, M. F., and McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46:217–258.

Tanenhaus, M. K., Leiman, J. M., and Seidenberg., M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, 18:427–440.

van Gelder, T. (1989). *Distributed Representation*. PhD thesis, Department of Philosophy, University of Pittsburgh, Pittsburgh, PA.