The Lexical Distance Model and Word Priming

Noel E. Sharkey
Centre for Connection Science
University of Exeter

The Lexical Distance (LD) model, presented here, functions as the front end of a connectionist Natural Language Understanding system (e.g. Sharkey, 1989a and b). The lexicon consists of a vector of microfeatures which are divided among 3 classes: orthographic, semantic and situational. Treating lexical space as an energy landscape, the entry for each word is learned as a minimum of the energy function E (see Kawamoto, in press for a similar treatment). Initial access to the lexicon is via the graphemic microfeatures. When these are activated by the visual presentation of an word, the lexical net is destabilised and the system begins gradient descent in the energy function until it relaxes in an attractor basin which represents the meaning of the input word. The model characterises context effects in word recognition experiments by deriving time predictions based on the movement of the system from its initial state to the target state. Two classes of context are discussed along with their interactions with word frequency and stimulus degradation. The research demonstrates how these effects fall quite naturally out of the processing specifications of the LD model without need for *ad hoc* parameters.

Contextual effects on word recognition may be divided into two classes: those that occur as a result of processes within the lexicon (*lexical effects*) and those that occur as a result of processes occurring after proposition construction (*textual effects*). The class of effect is determined by the priming stimulus used. Lexical effects are found when single-word primes such as DOCTOR precede targets such as NURSE. Alternatively, textual effects occur only when the priming comes from complete propositions. For example, Sharkey and Mitchell (1985) found that sentences such as, 'Colonel Jones realised that he was late as he rushed into the station.' could be used to prime words such as BENCH. The resulting effects are textual in the sense that they rely on the construction (or activation) of related propositions. In the 'Colonel Jones example', BENCH is primed by propositions containing the reader's knowledge about stations i.e. people waiting for trains sit on benches.

The distinction between textual and lexical effects may be maintained empirically as follows: The lexical effects are instantaneous (Neely, 1976) and can be disrupted by one intervening item (e.g. Meyer, Schvaneveldt & Ruddy, 1972; Gough, Alford, & Holly-Wilcox, 1981; Foss, 1982; A.J.C. Sharkey, 1989). The textual effects have a slow onset i.e. they appear only after an unfilled delay (Kintsch & Mross, 1985) or a filled delay (Till, Mross, & Kintsch, 1988; A.J.C. Sharkey, 1989). In addition, textual priming has been shown to sustain over a number of unrelated items (Foss, 1982; Sharkey & Mitchell, 1985; A.J.C. Sharkey, 1989), and is deactivated only when textual cues indicate that a new knowledge domain is in focus (Sharkey & Mitchell, 1985).

In the new model presented here, lexical effects (from single-word primes) are entirely bottom-up in the sense that they occur within the lexicon without influence from other modules. However, in word priming which is textual by an assembly of text propositions, there is minimal a top-down component which is supported in a number of studies (e.g. Glucksberg, Kreutz, & Rho, 1986; Tabossi, 1988; Blutner & Sommer, 1988; Keenan, Golding, Potts, Jennings, & Aman, in press). Although, the empirical arguments currently present a muddy picture, the LD model was built partly out of engineering considerations and so shows *one* efficient way in which to model context effects. It is not argued here that it is

the only way to model them. Nonetheless, by building an explicit alternative computational model with precise process predictions it is hoped that some of the issues can be resolved empirically. The model has recently been empirically compared to the Kintsch (1988) model and been shown to fit the data better (Sharkey and Sharkey, 1989).

THE LEXICAL DISTANCE MODEL

Traditionally the lexicon has been considered to be a store of information about words e.g. information about their meanings, syntactic class, orthography etc. In the current model, the lexicon consists of a set of units such that each unit corresponds to a microfeature. This leads to quite a different class of model than previous models of word recognition (c.f. Sharkey, 1989b, for a detailed comparison). In earlier models, a concept, where mentioned, is represented either as a single network node or similarly, as the contents of some addressed location in memory. The radical change here is that the concept associated with a word does not occupy a single location in memory. Instead, it is distributed across several different memory locations. Each concept is composed of a number of microfeatures which represent elements of its meaning (see Sutcliffe, 1988). Such meaning microfeatures may be thought of as propositional predicates (e.g. a stereotypical set of microfeatures for man might be: is male, is tall, is strong, can't cook, likes women etc.). In the present model meaning is not only represented by semantic microfeatures such as is-male, has-wings, etc. It is also represented by situational microfeatures which provide information about the activities or events that a word is involved in, or locations in which it may be found. The lexical net is illustrated in Figure 1. Moreover, each microfeature may appear in several concepts. Thus DOCTOR shares many situational microfeatures with NURSE - they are both persons and they have overlapping job roles. This distributed representation makes it difficult to maintain the old addressing metaphor because each lexical entry would occupy a number different addresses. Of course there is more to the lexicon than meaning microfeatures. There may also be phonemic, and syntactic microfeatures etc. (e.g. Kawamoto, in press). Meaning microfeatures are the main concern of the current model, but it is the graphemic microfeatures that are used to gain access to the lexicon.

SOME PROPERTIES OF THE MODEL

(i) Each microfeature in the lexicon may be thought of as having an activation value. Thus a lexical entry may be characterised as a vector of microfeature activations. And, more importantly, each vector of microfeature activations may be identified as a point in an n -dimensional energy landscape (called lexical space here), where n is the number of microfeatures in the lexicon.

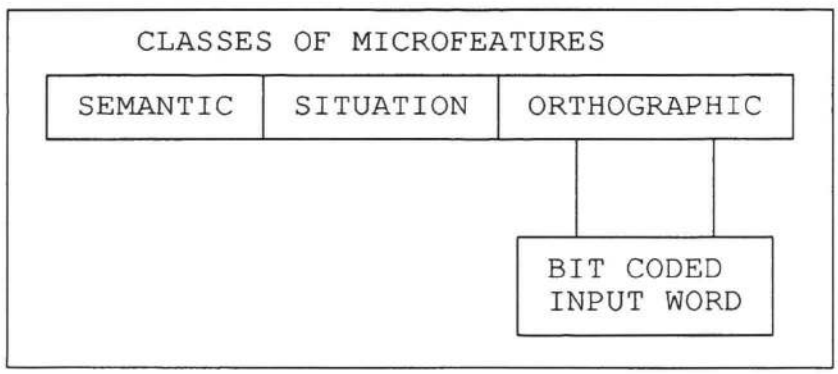

Figure 1. A diagramatic representation of the lexical net in the LD model showing three categories of microfeatures. A bit coded word activates the graphemic microfeatures in order to retrieve its appropriate meaning.

More formally, each collection of microfeatures representing a word's meaning and its visual characteristics (its lexical entry) is installed as a minimum of an energy function E (Hopfield 1982) given by:

$$E = 1/2 \sum_{i=j} s_i w_{ij} s_j + \sum_j s_j \theta$$

where s_i is the activation level of the i^{th} unit, w_{ij} is the weight between the i^{th} and j^{th} units.

This provides a new formalism with which to discuss the lexicon. Rather than considering a lexical entry as having a location or address in memory, it may now be considered as a point in an n-dimensional energy landscape created by the E function as shown in Figure 2. Each microfeature assembly is represented as a low point or basin in this landscape. Thus lexical access is characterised as a point moving through energy space to relax on appropriate assemblages of microfeatures. (Note that the new metaphor could be aligned with the old by saying that the unique point in n-dimensional space is a *location* for a lexical entry and that the vector of microfeature activations is the address of that entry).

(ii) Hopfield's (1982) *gradient descent* method is used to retrieve the meaning microfeatures which best fit the graphemic input constraints. Such a scheme is easy to implement on a parallel machine because an important property of Hopfield's formalism is that a given unit can locally compute the difference in energy a change in its state from 1 to 0 or vice versa will make. This is done simply by summing the total activity that a unit receives from all other active units in the network.

The change in energy for a unit is given by $\Delta E_{ij} = \sum_i s_i w_{ij}$.

The gradient descent rule is then simple. If the energy change results in a positive number, the unit adopts a +1 state, and if it results in a negative number the unit adopts a 0 state. Eventually the system will settle in a minimum of the energy function (one of the attractor basins as shown in Figure 2 i.e. a state which prevents the system from moving downwards in energy regardless of a change of state in any of the units. When the system relaxes in one of these stable states it is said to have retrieved that state i.e. the set of microfeatures corresponding to word meaning. One problem with gradient descent is that the system will only move in a downward direction. So when a new word in input, the system will be stuck in the minimum corresponding to the previous word's meaning. To overcome this problem, the pulse mechanism was employed here (c.f. Sharkey, Sutcliffe, Wobcke, 1986). Essentially, this means that when a new unit, not in the current microfeature set, comes on, it will pulse (switch off) all units to which it is negatively connected. This has the effect that the state of the lexical net jumps to a high point in the energy landscape which will be closer to the appropriate microfeature minimum than to any other minima in the net; only the microfeatures shared by the old and new pattern will stay active.

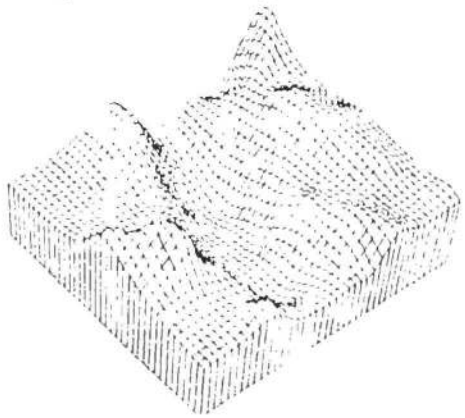

Figure 2: A three dimensional idealisation of an Energy landscape. The wells or minima are points representing lexical entries.

(iii). The relationship between the outside world and the lexicon in this model is via word units. These may be thought of as being like the outputs from McClelland and Rumelhart's (1981) interactive activation model. That is, in the simple simulation reported below, the output from the visual features of a word is represented as a single unit. The association between this unit and the graphemic microfeatures is learned using the delta rule. A property of this learning is that the weights for more frequently presented stimuli are stronger than the weights for less frequently occurring stimuli. (c.f. Sharkey, 1989 for a discussion of frequency and learning.

(iv). The activation values passed between the visual features and the graphemic microfeatures are incremental and continuous. The activity on a graphemic microfeature affects the probability of it adopting the +1 state during an update in the lexicon.

A SIMULATION OF CONTEXT EFFECTS

To model the experimental findings, a simulation was conducted for pairs of related words such as DOCTOR/NURSE, KNIFE/FORK, BREAD/BUTTER, DOG/BONE, and FOOT/SHOE. This simulation was exploratory and so ten microfeatures were arbitrarily assigned to each word in the lexicon. Four of these represented graphemic microfeatures, three represented semantic microfeatures and the other three represented situational microfeatures. The three situational microfeatures for each word were shared with its associate e.g. DOCTOR shared three situational microfeatures with NURSE.

Installing the meanings. The lexical microfeature sets, corresponding to lexical entries, were installed in the lexical net using an autoassociative version of the delta rule in combination with the pulse mechanism (Sharkey, Sutcliffe, & Wobcke, 1986). Briefly, the lexical entries of words were collated as patterns of microfeature activations to be learned by the system. The patterns were then presented to the system one at a time. Each pattern was used to activate a set of input-units. The states of these units were then propagated across a set of weights to produce values on a set of output-units. These output values were compared with the values of the corresponding input-units to produce an error vector **d**, where δ_i = (input-unit$_i$ - output-unit$_i$). The change in weights between the input-units is given by $\Delta \mathbf{W} = \eta \mathbf{di}$, where $\Delta \mathbf{W}$ is the weight change matrix, h is the learning rate parameter, and **i** is the vector of input-unit values.

Once the learning had been completed the system was started in one of two initial states; either (i) a stable state resulting from the presentation of a prime word, or (ii) an arbitrary state resulting from the presence of a neutral (e.g. a row of Xs). A prime word activates a set of microfeature units and sets the system on a downward descent in the energy function until a stable minimum has been reached. This minimum will be the lexical entry for the prime word. In contrast, when the target is preceded by a neutral instead of a prime, the resulting starting state will be arbitrary (and it may not be a minimum of E). Now, when a target word is presented, some new graphemic units are activated, and the system begins to move from the current state to a state which best fits the input.

Timing predictions. The metric used to derive the time predictions is the *distance* which must be traversed to get from an initial state to a target state in the lexicon, where distance, d, is defined as the length between two points in n-dimensional lexical space. To make this clearer, imagine two vectors of microfeature activations in the lexical space $\mathbf{L}^n$. Let these vectors represent the starting state of the system **s** and the required or target state **r**. Then the distance between the two points **s** and **r** is given by $\|\mathbf{s} - \mathbf{r}\|^{1/2}$, where length $\|\mathbf{v}\| = (\mathbf{v} \cdot \mathbf{v})^{1/2}$. A major assumption of the model is that the greater the distance from an initial state to a target state, the longer will be the recognition time for a target word.

Context effects. The model makes the correct time predictions for context effects because a target that shares a number of microfeatures with a prime (semantically and/or situationally related), will be closer to the prime than a word which shares no features (semantically and situationally unrelated). Thus, by definition, the state resulting from presentation of a related

prime will be closer, in lexical space, to the target state than the state resulting from an unrelated prime. Figure 3 plots distance against energy for two pairs of words DOCTOR/NURSE and DOCTOR/FORK. For simplicity binary activation has been used here and so the graph shows the energy of the system as it moves from the initial DOCTOR state to the NURSE and FORK states. Note that FORK is much further from DOCTOR than NURSE is and that there is a much steeper ascent and descent to reach FORK. In the DOCTOR/NURSE graph, the first circle indicates the state of the system with only the shared microfeatures on, and second circle indicates the state after all of the graphemic microfeatures have come on and the pulse mechanism has been run. If the second circle is compared to the circle on the DOCTOR/FORK graph, it can be seen that NURSE is considerably closer than FORK to the initial state (DOCTOR). Therefore, the NURSE state will take less time to reach.

Lexical priming. This simulation presents a very simple and entirely bottom-up model of lexical context effects in which lexical priming is a measure of network distance from an initial to a decision state. The main factor in time to respond (e.g. Lexical Decision) is the relationship between the target and the initial state of the system. It is assumed that, overall, the target states are further away from the arbitrary Neutral state than from the Related prime states; but the target states may, on average, be closer to the Neutral states than to the Unrelated prime states.

Textual priming. It was shown that textual effects can be produced by exactly the same processes as the lexical effects i.e. through shared situational features. However, the different onset/offset properties of the two types of priming result from processes which operate externally to the lexicon. There is not space to delve into these processes here (but see Sharkey, 1989b and in press). Briefly, in reading text, the primary aim is to construct meaning propositions. In the Sharkey (1989b) model, once a proposition has been constructed, it activates a knowledge-net which results in a stable state of situationally related propositions. This stable state is maintained until cues from the text indicate otherwise (Sharkey & Mitchell, 1985). In order to explain the sustain of textual priming on word decisions in the current model, the stable state in the knowledge-net holds the situational microfeatures active in the lexicon. Moreover, the time taken to construct a proposition explains why the onset of textual priming is slower than lexical priming. Note that this minimal top-down view is different from previous top-down models. It is not claimed that particular lexical items are *expected*. Nor are particular words or visual features being anticipated. On the contrary, it is only the shared abstract contextual properties of

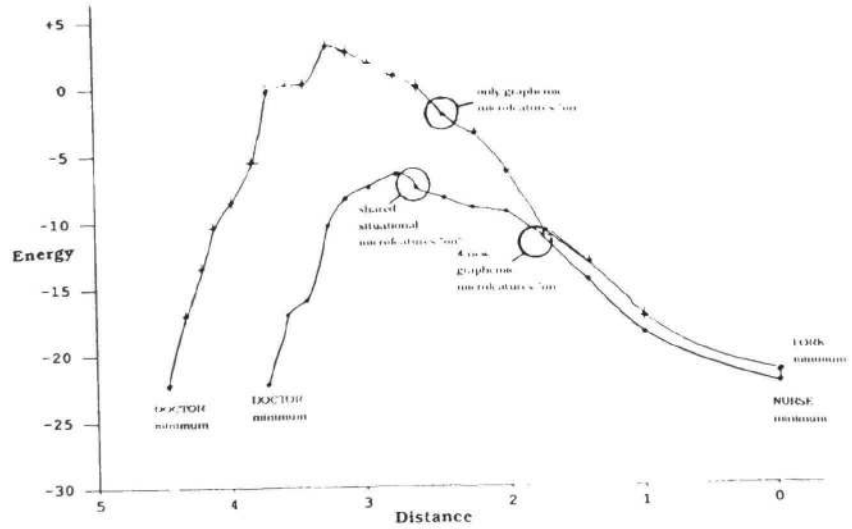

Figure 3: A graph of the movement in the lexicon from DOCTOR to NURSE and from DOCTOR to FORK. This is plotted as Distance against Energy. See the text for a discussion.

words (the situational microfeatures) which are held active. Thus the system is, in a sense, predisposed to receive certain contextual classes of words. This saves on the computational complexity of earlier models and serves the function of disambiguating word meaning[1]. However, before the utility of this model can be assessed, it is important to examine whether it accurately predicts the interaction of context effects with other variables such as word frequency and stimulus quality.

Rate of microfeature activation. In order to understand the predictions arising from the model on the combined effects of word frequency, context, and stimulus quality, the way in which the rate of graphemic microfeature activation affects movement from the initial to the target state is first examined. At time t_0, just before presentation of the target, and therefore before any new graphemic units have been activated, the distance from the start state **s** to the required or target state **r** is $\delta = \|\mathbf{s} - \mathbf{r}\|^{1/2}$. At time t_1, the presentation of the target will activate the word unit vector **v**, and at t_2 this activation will be broadcast across the weights **W** using the limiting function LIMIT[**Wv**] = **f**, where **f** is the vector of new graphemic microfeature activations. The the graphemic microfeatures are activated at t_3 by adding **f** to the initial lexical vector **s**. So the new state of the system, before update at t_4, will be **s** + **f**, and the distance will now be $\delta = \|(\mathbf{s} + \mathbf{f}) - \mathbf{r}\|^{1/2}$. Thus the magnitude[2] of the new microfeature activations, **f**, will affect the distance moved between the start state and the target states; the greater **f** the smaller the difference between **s** and **r**. The process described in these four time cycles continues to iterate until a stable state is attained.

Stimulus quality and frequency effects. Predictions concerning frequency effects rely on a property of learning to associate the visual features with the graphemic microfeatures. In delta rule learning, the weights for more frequently presented stimuli are larger than the weights for less frequently occurring stimuli. Now, as shown above, the distance moved by the system towards the target state depends on the magnitude of **f** which in turn depends on two factors: the structure of the weights **W** and the magnitude of word unit vector **v**. Since frequency is encoded in **W**, when the stimulus quality in **v** is held constant, low frequency targets take longer to maximise activation on the graphemic microfeatures than do high frequency targets. In other words, the stronger the connection between a set of graphemes and their lexical representation, the greater will be the rate of microfeature activation. Thus high frequency targets will be responded to faster than low frequency targets.

The effect of degrading the quality of a stimulus word is simulated in the model by varying the magnitude of **v** while holding frequency constant. It should be clear from the above analysis that the smaller the magnitude of **v** (stimulus quality), the smaller the magnitude of **f**, and consequently the longer it will take to maximise activation on the graphemic microfeatures. Thus the model predicts that degraded stimuli will take longer to recognise.

Combined effects of context, frequency and stimulus quality. The model predicts that both stimulus quality and frequency will interact with context. This is because the closer the initial state **s** is to the target state **r**, the less effect the magnitude of **f** will have on the movement of the system from **s** to **r**. An initial state close to a target state will reach the target state before the microfeature activations have been maximised. Because of the nature of update, the lexicon will in effect "clean up" degraded stimulus. These predictions have been supported empirically. Becker and Killion (1977) and Becker (1979) found context by frequency interactions, and interactions of context and stimulus quality have been demonstrated by Meyer, Schvaneveldt, and Ruddy (1975), and Becker and Killion (1977). Moreover, since frequency and stimulus quality effects are brought about by changes in the magnitude of **f**, our model predicts an additive effect of stimulus quality and frequency. However, demonstrations have had mixed results. Some research has shown frequency and stimulus

---

[1]By derivation from the Kawamoto (in press) model, the current model can also explain the lexical ambiguity effects (c.f. Sharkey, in press).

[2]For mathematical simplicity, it is assumed here that after initial activation, **f** does not change direction.

quality to be interactive (Stanner, Jastrzembski, and Westbrook, 1975) though the majority have found them to be additive (Norris, 1984; Becker & Killion, 1977).

It should be noted that the distance metric has parallels with the older Location Shifting model (e.g. Meyer, Schvaneveldt & Ruddy, 1972; Posner & Snyder, 1975). Both models accurately predict that associative priming effects can be disrupted by the presentation of an unrelated item between the prime and the target (e.g. Meyer, Schvaneveldt & Ruddy, 1972; Gough, Alford, & Holly-Wilcox, 1981; Foss, 1982; A.J.C. Sharkey, 1988). In both models, an unrelated intervening item would move the state/location of the system to a new state/location. And it is this new state/location which would be the initial state/location before the presentation of the target. Therefore, priming of the target would be disrupted. However, because we use the computational power of a distributed representation, it is not possible, as in Posner & Snyder (1975), to speak of *the* location for a concept; it may share meanings with other concepts in more than one location. Instead, location is discussed more abstractly in terms of n-dimensional energy space and Euclidean vector distance. In addition, the LD model makes prediction about textual priming which would not be possible from the older model.

CONCLUSIONS

It has been shown how a simple connectionist model can generate accurate predictions for both lexical and textual context effects and their interactions with frequency and stimulus quality. There are no hidden processes in this model and all of the associations are learned. The model provides an entirely bottom-up account of lexical effects which does not rely on fast spreading activation to contextually related concepts; recognition threshold adjustments; plausibility checks; or shortlist search. Indeed, it does not require any special purpose mechanisms to handle context effects; the effects fall naturally out of the normal operation of the lexicon during access. If the prime word is contextually related to the target word, in the restricted definition of sharing microfeatures, the lexicon will have less distance to travel in order to stabilise on the best fitting lexical microfeatures. Moreover, textual effects fall out of the same lexical processes as the lexical effects. The two types of priming differ only in processes that occur externally to the lexicon. For textual priming, the activation of a propositional knowledge-net holds active the situational microfeatures in the lexicon and thus provides a sustain of priming. Since the knowledge net is most responsive to propositional input, the time taken to construct propositions accounts for the slow onset of textual priming.

REFERENCES

Becker, C.A.(1979) Semantic context and word frequency effects in visual word recognition. Journal of Experimental Psychology: Human Perception and Performance, **5**, 252-259.

Becker, C.A. & Killion, T.M. (1977) Interaction of visual and cognitive effects in word recognition. Journal of Experimental Psychology: Human Perception and Performance, **3**, 389-401.

Blutner, R., & Sommer, R. (1988) Sentence processing and lexical access: The influence of the focus-identifying task. Journal of Memory and Language, 27, 359-367.

Foss, D.J. (1982) A discourse on Semantic Priming, Cognitive Psychology, **14**, 590-607.

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

Gough, P.B., Alford, J.A., Jr., & Holley-Wilcox, P. (1981) Words and Contexts. In J.L. Tzeng & H. Singer (Eds.). Perception of print: Reading research in experimental psychology. Hillsdale, NJ: Erlbaum.

Hopfield, J.J. (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences, U.S.A., 79, 2554-2558.

Kawamoto, A.H. (in press) Distributed representations of ambiguous words and their resolution in a connectionist network. In S.L. Small, G.W. Cottrell and M.K. Tanenhaus (Eds) Lexical ambiguity resolution in the comprehension of human language.

Keenan, J. M., Golding, J.M., Potts, G.R., Jennings, T.M. & Aman, C.J. (in press) Methodological Issues in Evaluating the Occurrence of Inferences. In A. Grasser and G.H. Bower (Eds.) Learning and Motivation, Vol. 24. Academic Press.

Kintsch, W. (1988) The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 95, 163-182.

Kintsch. W., & Mross, E.F. (1985) Context effects in word identification. Journal of Memory and Language, **24**, 336-349.

McClelland, J.L. & Rumelhart, D.E. (1981) An interactive model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, **88** , 375-407.

Meyer, D.E., Schvaneveldt, R.W., & Ruddy, M.G. (1972) Activation of lexical memory. Paper presented to the psychonomic society, St Louis, Mo.

Meyer, D.E., Schvaneveldt, R.W., & Ruddy, M.G. (1975) Loci of contextual effects on word recognition. In P.M.A. Rabbitt & S. Dornic (Eds) Attention and Performance V. New York: Academic Press.

Neely, J.H. (1976) Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. Memory and Cognition, **4**, 648-654.

Norris, D.G. (1984) The effects of frequency, repetition and stimulus quality in visual word recognition. Quarterly Journal of Experimental Psychology, **36A**, 507-518.

Posner, M.I., & Snyder,C.R. (1975) Attention and cognitive control. In R.L. Solso (Ed) Information processing and cognition: The Loyola Symposium. Hillsdale, N.J., Lawrence Erlbaum Associates.

Sharkey, A.J.C. (1989) Contextual mechanisms of text comprehension. Unpublished Ph.D. Dissertation, University of Essex.

Sharkey, A.J.C. & Sharkey, N.E. (1989) Lexical processing and the mechanism of context effects in text comprehension. The proceedings of the 11th Annual Conference of the Cognitive Science Society.

Sharkey, N.E. (1989 a) A PDP learning approach to natural language understander. In I. Alexander (Ed) Neural Computing Architectures. London: Kogan Page.

Sharkey, N.E.(1989 b) Connectionist Memory Modules for Text Comprehension, Research Report #170, Dept. Computer Science, University of Exeter.

Sharkey, N.E.(in press) A Connectionist Model of Text Comprehension. In D. Balota, G.B. Flores d'Arcais and K. Rayner (Eds.) Comprehension Processes in Reading.

Sharkey, N.E and Mitchell D.C. (1985) Word Recognition in a Functional Context: the Use of Scripts in reading. Journal of Memory and Language. **24** 253-270.

Sharkey, N.E., Sutcliffe, R.F.E. & Wobcke, W.R. (1986) Mixing Binary and Continuous Connection Schemes for Knowledge Access. Proceedings of the American Association for Artificial Intelligence.

Stanners, R.F., Jastrzembski, J.E. & Westbrook, A. (1975) Frequency and visual quality in a word-nonword classification task. Journal of Verbal Learning and Verbal Behavior, **14**, 259-264.

Tabossi, P. (1988) Accessing lexical ambiguity in different types of sentential contexts. Journal of Memory and Language, **27**, 324-340.

Till, R.E., Mross, E.F., & Kintsch, W. (1988) Time course of priming for associate and inference words in a discourse context. Memory and Cognition, **16** (4) 283-298.

Acknowledgements: I would like to thank Amanda Sharkey for comments on earlier versions of this paper, and the Leverhulme Trust (A/87/153) and ESRC (CO820015) for supporting this research.