Modelling Retroactive Context Effects in Spoken Word Recognition with a Simple Recurrent Network

Alain Content
Laboratoire de Psychologie Expérimentale
Université Libre de Bruxelles
Avenue Buyl 117
B-1050 Bruxelles
acontent@ulb.ac.be

Pascal Sternon
Laboratoire de Psychologie Expérimentale
Université Libre de Bruxelles
Avenue Buyl 117
B-1050 Bruxelles
psternon@ulb.ac.be

Abstract

We present a new variant of a simple recurrent network to model auditory word recognition in continuous speech and address the issue of lexical segmentation. Simulations based on small word sets show that the system provides a near-optimal solution to the opposite constraints of speed, which requires that lexical processing be immediate, and reliability, which imposes that identification decisions be postponed until unambiguous information is available. Contrary to an often-heard statement, the simulations show that the existence of embedded words is not incompatible with the notion of continuous on-line lexical processing.

Lexical Segmentation and Retroactive Context

Upon hearing the string /sɛrpãdãsã/ as in the two French sentences below,

(1) ʒevyœ̃sɛʀpãdãsãlatɛʔot
J'ai vu un serpent dansant la tête haute
"I saw a snake dancing head held high"
(2) ʒevyœ̃sɛʀpãdãsãlatɛt
J'ai vu un cerf pendant sans la tête
"I saw a deer hanging with no head"

how and when does the listener discover the appropriate lexical parsing to distinguish between the "snake" and the "deer" interpretation?

The speech signal unfolds in time, and the boundaries between linguistic units are not explicitly marked. These characteristics raise problems for theories that postulate that the mapping process, the comparison between sensory input and lexical representations, proceeds on-line and continuously. The immediacy principle yields one important computational benefit, which is the speed of word identification. Indeed, access to lexical information becomes possible as soon as the sensory input is sufficient to uniquely specify one lexical candidate (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987). However, one essential difficulty for such theories consists in stating the nature of the constraints that guide the mapping process, in order to avoid the burden of having to entertain the exhaustive series of lexical hypotheses compatible with any part of the speech stream (Shillcock, 1990; Harrington & Johnson, 1987).

In some theories, among which the Cohort model (Marslen-Wilson & Welsh, 1978) is probably the most well-known, the only lexical candidates considered are the ones that are aligned with the point in the signal corresponding to the current word onset. An obvious requirement for such a hypothesis to work is that the onset of a new word can be reliably identified in the speech stream. Perhaps the simplest strategy that may be envisaged to determine word boundaries consists in predicting the next word's onset based on the identity (and predicted offset) of the current word. Such a strategy fits well with the immediacy requirements, since it would allow to locate a word's beginning very early on, even before its realisation.

However, because natural lexicons comprise *embedded* words, i.e., words made of parts of other words (like cerf and serpent), such a predictive segmentation strategy will not succeed in all cases. In fact, a count based on a 20,000-word American English database indicated that less that 40% of words are made of unique phonemic strings (Luce, 1986), and experimental data have shown that a substantial proportion of words are not uniquely identified before their acoustic offset (Grosjean, 1985; Bard, Shillcock & Altmann, 1988). Furthermore, as our initial example demonstrates, the appropriate parsing may depend on information that comes in much later in the sentence.

It is well known that there are phonetic, phonotactic and prosodic cues which may help determine word boundaries. Although French is not a stress-language, prosodic cues to word segmentation may nevertheless be available. Intuitively, for a native speaker of French at least, the presence of stress on the fifth syllable of (1) or the fourth of (2) seems sufficient to eliminate any ambiguity. Rietveld (1980) has observed that systematic variations in duration, intensity and fundamental frequency are associated with word boundaries in French. Similarly, for English, Cutler and Norris (1988) have suggested the use of a metrical segmentation strategy, whereby a strong syllable is processed as a possible word onset.

It is however far from clear that these kinds of bottom-up heuristics provide a complete solution to the problem of lexical segmentation. Recent experimental findings by Tabossi (1993), using a cross-modal semantic priming technique in Italian, suggest that embedded words as well as straddling words may be activated even in contexts in which clear segmentation cues are available. Such findings point to an alternative principle, which consists in postponing

decisions until sufficient information is available. Experiments based on the phoneme restoration effect (Samuel, 1990) and lexical influence on phonemic categorisation (Connine, Blasko & Hall, 1991) suggest that the processing system is indeed capable of revising its earlier decisions, at least for a limited time period. Of course, delaying commitment may eliminate any potential effect of local ambiguity and may thus ensure reliability, but at the expense of speed.

The TRACE model (McClelland & Elman, 1986) provides one instance of an optimal compromise between the two opposite requirements of speed and reliability. The lexical segmentation problem is solved in TRACE by combining three sources of constraint : the sequential reception of sensory information that accumulates progressively into the "memory trace" constituted by the set of feature and phoneme detectors for successive time slices; the mapping from the memory trace to lexical representations, based on a parallel, exhaustive and gradual comparison mechanism; and the direct competition between lexical hypotheses, implemented through the inhibitory lateral connections between lexical units. The sequential upcoming of the input ensures that compatible lexical hypotheses will be activated as soon as possible, but on the other hand, the existence of a memory trace covering a substantial interval of time and the gradual nature of the mapping mechanism permits to modify the strength of lexical hypotheses to account for the entire portion of signal available in the trace at each moment. In fact, detailed simulations have shown that, most of the time, TRACE is capable of finding the lexical candidates that best fit a given string of phonemes (Frauenfelder & Peeters, 1990, 1993).

The architecture of the TRACE model has however been criticised because of its representation of temporal information. The spatial time-window metaphor, based on reduplication of detectors for successive time slices, appears both unsatisfactory and unplausible. It leads to untractable numbers of connections with large corpora (Norris, 1990), and cannot easily handle temporal variability in the signal. One potential solution stems from the use of recurrent networks, in which temporal information is not represented explicitly, but can be encoded dynamically thanks to the connectivity of the system. One topology of recurrent network, which has been proposed by Elman (1990), and is currently known as a Simple Recurrent Network (SRN), consists in providing a copy of the hidden unit activation state vector at one cycle as input to the network at the next time step. Because the connections can be modified to reflect training experience, the nature of the temporal information encoded will depend on the particular task imposed on the network.

One attempt to use a SRN to model spoken word recognition has been reported by Norris (1990). Norris used a small corpus of 50 forms, and trained a SRN on a continuous sequence of words, presented segment after segment. The output consisted in a bank of 50 word detectors. Norris showed that the network captured some of the basic properties of the Cohort model. Cohort members, i.e. lexical candidates aligned with the onset of the upcoming word, were activated initially and dropped out progressively upon reception of diverging information. Thus, a single candidate remained active as soon as the uniqueness point was attained. Despite these observations, Norris concluded that the SRN could not provide a viable model of human word recognition, because it appeared unable to accommodate retroactive context effects. For instance, the network could not correctly discriminate sequences such as CATLOG and CATALOG in which CAT corresponds to a word in the first case but not in the second. Because the activation corresponding to the word CAT will never become reliably higher than the activation of its carrier CATALOG, the SRN cannot distinguish between these two situations. Basically, the thrust of the criticism is the point already raised by late-isolation findings in the gating paradigm (Grosjean, 1985; Bard *et al.*, 1988): Models that assume onset alignment and immediate mapping cannot work when the lexicon includes words which do not become unique before their end point. Non-unique strings (embedded words) will not be identified, and the processing of their immediate successor will be obstructed by the lack of clear boundary decision.

The aim of the present study was to develop and evaluate a new variant of the SRN which we thought could help solve the problem of retroactive context effects. Our main objectives were to examine the feasibility of the approach and to explore in details the basic (behavioural) properties of the system, with regards to current psycholinguistic models and available empirical findings. The essential idea was to modify the task imposed to the network so that its state would not only reflect its best choice for the *current* word, but also the best fitting *previous* word.

Network Architecture and Training Method

Figure 1 presents the network's topology. The input consisted of seven units representing simplified acoustic/phonetic features, as in the TRACE II model. Each segment was represented by a vector of seven continuous dimensions (Consonantal, Vocalic, Diffuse, Acute, Voiced, Power, Burst Amplitude; see McClelland & Elman, 1986, p. 15). No information about prosody or durational cues was provided in the present simulations. Words were presented as an uninterrupted sequence of segments, without pauses or silences. The output consisted of 20 word detector units. Localised output representations were used in order to alleviate the interpretation of output patterns.

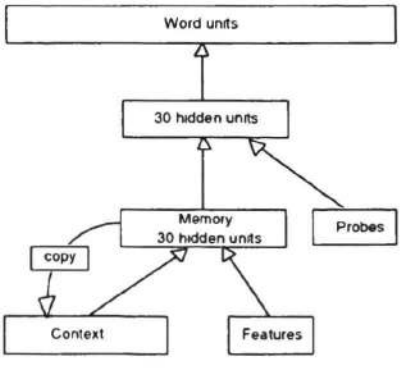

Figure 1. Architecture of the Network

A further set of external input units, the Probes, served to determine the nature of the task imposed to the Net. Two values were used in this study. Under the C probe, the desired response was the unit corresponding to the word currently being presented. Under probe P, the expected response was the preceding word (Cf. St John & McClelland, 1990, for a similar approach to sentence processing). During learning, one probe was selected randomly at each cycle, so that the two tasks were completely interwoven.

From a cognitive standpoint, the Probes are not meant to represent real external stimulation. Rather, they can be seen as an abstract means of implementing the two opposite constraints of immediacy and delayed commitment on the way the network learns to encode, organise and preserve information in its distributed memory trace. The C task acts as a pressure to produce the desired response as soon as possible, since the target output was given throughout the presentation of the word. The P task requires that the desired output be maintained during the following word.

In terms of processing mechanisms, the first set of hidden units elaborates a compact distributed representation of the input sequence, and the second bank of hidden units serves to combine the distributed trace with the probes to extract lexical hypotheses. Although in the present implementation, this extraction mechanism cannot deal with the C and P tasks simultaneously, it could be assumed that the alternation is much faster than the time scale of the external input, or that duplicate sub-networks using fixed probe values extract information about the current and previous word in parallel, so that in practice, both sets of lexical hypotheses would be continuously available.

Results from four simulations will be reported, each based on a small corpus of 20 forms, from 2 to 5-segment-long. LEX1 included ten words (e.a. *stak, star, stap, rad, rak, rab*) and their reversals (*kats, rats, pats, dar, kar, bar*). Uniqueness points varied from position 1 to position 4. LEX2 was identical to LEX1 except that the final segment of four words was stripped out to create left-embedding conditions (e.g. *sta* derived from*stak*, *ra* from *rab*). LEX3 and LEX4 were constructed in a similar way and included respectively four center- and four final-embedded items. The four runs used the same training regimen and parameters (lrate=.05, no momentum, plain back-propagation), but different initial weight values. Each epoch consisted of a continuous sequence of 400 words presented segment by segment without explicit word boundary markings. Sequences were generated randomly with the constraint that 10% of all pairs were kept out to test generalisation. Each word was thus seen with 18 predecessors and 18 followers. All the words had equal probabilities of occurrence during training. No attempt was made to manipulate or capture higher-order distributional regularities, such as sentence or constituency relations.

Results

To analyse the network's behaviour, all possible combinations of three words were generated. Each triplet was preceded by two cycles of "noise" (.5 feature vector) and context units were reset to their initial state (.5) at the onset of each sequence. The main analyses concern responses to the C probe during the presentation of the second word, and responses to the P probe during the third word. Responses to the first word were not analysed in detail, because they are influenced by the initial state of the network and do not reflect its general behaviour.

According to the Cohort model, a word can be identified at its Uniqueness Point (UP), i.e. as soon as it remains the sole lexical candidate compatible with the input. Figure 2 displays the activation of the target and the most active non-target candidate at each time step for LEX1. The left part of the graph corresponds to the C task and activations have been averaged across words aligned on their UP. The target's activation gradually increases, with an abrupt acceleration at the UP, which also corresponds to the first cycle at which targets can be discriminated from other candidates. Similar curves came out for the other runs. Like the SRN used by Norris (1990), the network appears to implement a gradual activation version of the Cohort theory. However, even though the UP appears as an important factor in accounting for the variations of activation, the size of the standard deviation at the UP and the further increase after the UP indicate that other factors come into play.

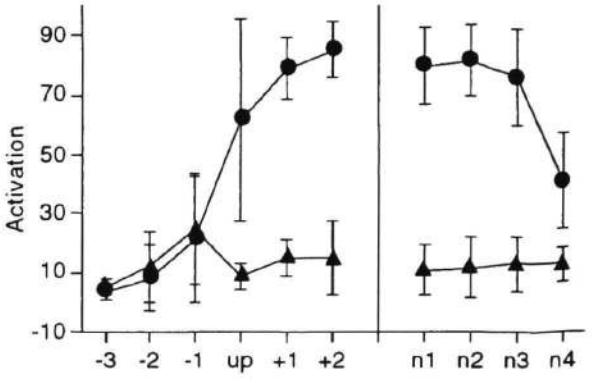

Figure 2. Activation of target and most active non-target unit (LEX1). Left panel is for C probe, right panel for P probe. Bars show standard deviations.

The right part of Figure 2 shows the results of the P task, and it appears that the target unit remains active during the presentation of the following word, while it gradually decreases towards the end. The most visible effect of the dual task architecture is that the activation corresponding to the target is maintained at a high level during several cycles (generally, from the UP of the current word until the end of the successor). This temporal stability might provide a stronger and more reliable signal for other processing components.

To further clarify the relation between the network's behaviour and Cohort theory, we computed the isolation point for the second word of each test triplet. It corresponds to the first time step during the presentation of the word, at which the activation of the target unit exceeds that of all other output units by a difference of .5. Table 1 presents the percentage of isolated words for training and test sequences. For three simulation runs, on-line isolation rates are close to

90%. The lower performance for LEX2 on the C task is as expected: recall that LEX2 includes four left-embedded words, which cannot be isolated on-line.

	Current word		Previous word	
Corpus	Training	Test	Training	Test
LEX1	87.2	75.0	87.2	80.1
LEX2	77.8	75.0	95.3	91.5
LEX3	95.3	97.5	95.6	91.4
LEX4	95.8	95.0	93.5	89.6

Table 1. Isolation rates on training and test sequences for the C and P identification task.

As regards previous word isolation, performance is again around 90%. Most of the time, P isolation was obtained on the first segment of the following word, confirming that the previous word was correctly recalled immediately after its offset. Finally, it may be worth mentioning that the level of performance on the 10% of word sequences that did not appear during training is quite high. In fact, in most cases, isolation rates and isolation point values on the test sequences are hardly discriminable from performance on training materials. Other tests using pseudo-word contexts led to analogous high generalisation rates. The high level of transfer leads to the reassuring conclusion, contrary to Norris' (1990) assertion, that the network need not see all the sentences of the language to identify the words. It also implies that the processing is more sophisticated than rote memorising of sequences of events.

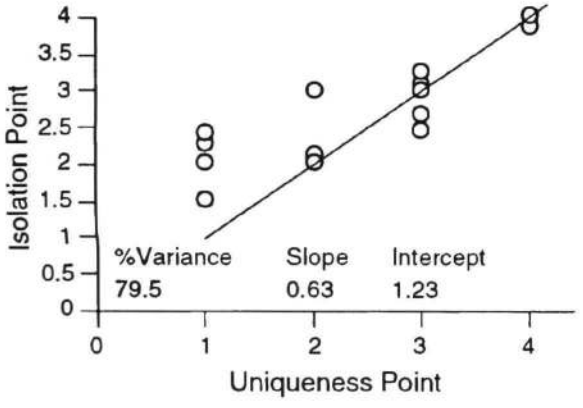

Figure 3. Observed relation between isolation points and uniqueness points. The line corresponds to the Cohort prediction (data from LEX1).

As discussed previously, the optimal behaviour, which corresponds to the Cohort prediction, would be that words are isolated on-line at soon as they become unique, so that the IP is identical to the UP. The network's behaviour comes close to these expectations (Figure 3), particularly for late positions. At early positions within words, on-line isolation tends to be delayed relative to the theoretical prediction of the Cohort model. Similar regression lines were obtained for all four simulations, with slope values smaller than 1 and positive intercepts. Interestingly, the fact that the network underestimates the UP effect is analogous to human observations (Radeau, Mousty & Bertelson, 1989; Radeau & Morais, 1990).

Further analyses were performed on the C task to determine the nature of the factors that influence lexical activation. Two factors were considered, cohort size, the number of cohort members before the UP, and stimulus quantity, the number of segments of the input string counted from current word onset. This choice was motivated by the following considerations. Previous analyses suggested that stimulus quantity could have an effect, in that the deviation from Cohort behaviour was more marked for early UP words. On the other hand, it has been shown that under conditions in which a single unit response is required, the SRN's output will approximate the conditional probability distribution of the response set (Servan-Schreiber, Cleeremans & McClelland, 1991). Under this analysis activation would be expected to vary strictly as an inverse function of cohort size.

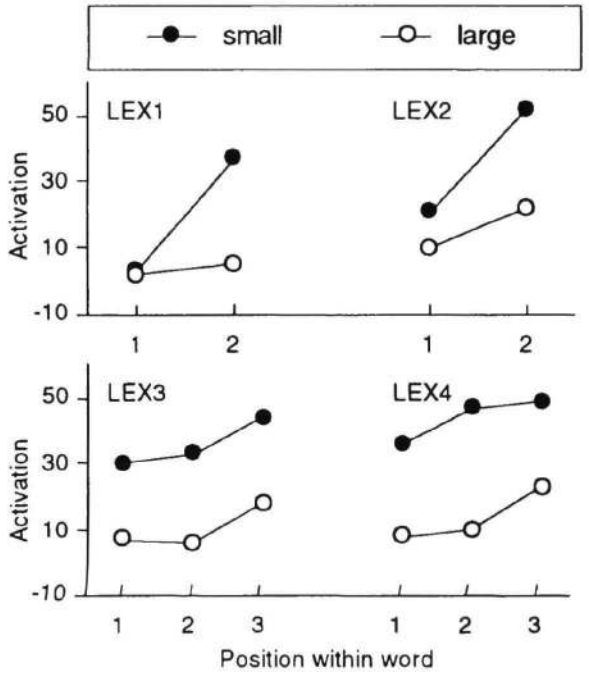

Figure 4. Mean activation of the current word as a function of cohort size before UP and of position within word (see text for explanation).

In each corpus, a contrast could be devised between subsets of words with late UPs (positions 3 or 4), whose cohort includes either few members (i.e., 1) or many (3 in LEX1 and LEX2, 2 in LEX3 and LEX4). Moreover, in all the cases examined, cohort size was stable up to the UP, so that position effects could be assessed independently of cohort size.

As shown in Figure 4, both factors appear to influence the activation of words before their UP. A large effect of cohort size is observed in all cases, and the probability distribution hypothesis does not fall too far from the observed data. Thus, even though the information flow does not incorporate any kind of direct competition between responses, contrary

to models such as TRACE in which there is lateral inhibition, the strength of lexical hypotheses may depend on the number of compatible candidates. Mixed patterns came out for the effect of position: results from LEX1 and LEX2 show large activation increases between the two initial positions, but only when cohort size is small. On the contrary, in LEX3 and LEX4, activation increases only marginally before the UP, suggesting that cohort size is the major force determining lexical activation. Further simulation work is needed to establish whether this is a general feature of the present model.

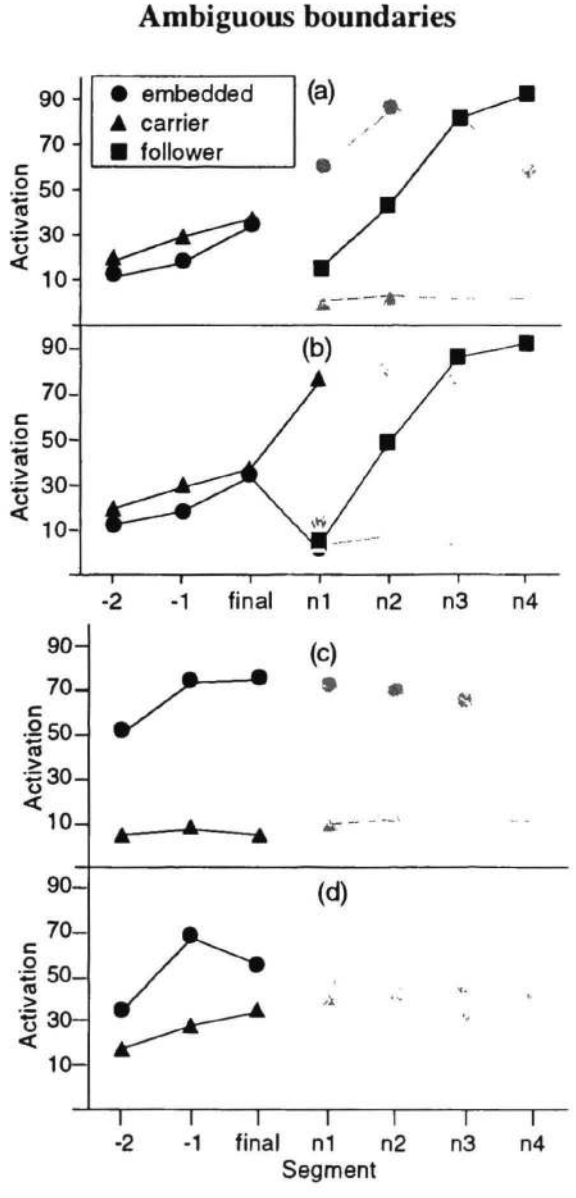

Figure 5. Activation of embedded words and their carriers during the presentation of the embedded words (-2, -1, final) and followers (n1 to n4). Panels a-b for left-embedding (/kat/-/katS/). Panels c-d for right-embedding (/at/-/kat/). Black lines are for C task, grey lines for P task.

Given the nature of the word sets used, the only case in which the immediate analysis does not permit to reach a decision is LEX2, which includes left-embedded words. The following analyses aimed at establishing whether the network can indeed resolve such ambiguities. The upper panel of Figure 5 shows the mean activation of embedded words and their carriers. The nature of the following word determines two possible conditions. Figure 5a corresponds to the case in which the follower is incompatible with the carrier (katpl^s, CAT PLUS). In this case, the activation of the embedded word (CAT) increases and the activation of the carrier (CATCH) drops on the first segment of the next word (position n1), resulting in late isolation of the embedded word.

Figure 5b corresponds to the lexical garden-path situation in which the follower is indeed compatible with the carrier (katSip, CAT SHIP). In this case, the carrier word is erroneously isolated on the first segment of the follower, but the network recovers from this inappropriate interpretation at the next cycle. Thus, the appropriate pattern of activation is attained at position n2. Similar response patterns have been observed in the gating task (Grosjean, 1985). The cost of immediacy is the transient activation of the carrier word, and a one-time-slice delay in the isolation of the embedded word. But a possible way to recover from immediate commitment is to rely on the accumulation of evidence over several time slices. Interestingly, as shown by the activation curves for the following word in Fig. 5a and 5b, the presence of a boundary ambiguity has only minimal effects on the on-line activation of the follower.

The lower panels displays the symmetric condition, from the LEX4 data, in which right-embedded words are preceded by a word which is either compatible with the carrier (Figure 5c, ex.: stakat, STACK AT) or not (Figure 5d, ex.: pl^sat PLUS AT). Although there is no real ambiguity here (since the initial part of the predecessor — sta or pl^— can never be a word), the straddling carrier word (CAT) is clearly activated. Similar findings came out with center-embedded words, as a function of both the predecessor and the successor's compatibility. Thus, for instance, CATCH would be highly activated on the S segment in the sequence stakatSip (STACK AT SHIP). The response of the network is reminiscent of findings reported by Tabossi (1993).

Conclusions

In the present study, we opted for small scale simulations, which in return permitted to conduct detailed analyses of performance. Further simulations are in progress to examine how performance would scale up under various extensions (corpus size, number and nature of probes, information grain of the input signal).

Besides training time, one limiting factor is the nature of output representations. The use of localised codes offers the advantage of making performance analysis more tractable. But it constitutes an important drawback to the processing of more realistic data sets. Obviously, a distributed scheme would need to be devised to handle larger corpora, and richer input representations would also be required. In particular, it seems interesting to explore how the system

would cope with prosodic/metrical input information. Lexical segmentation ambiguities were kept to a minimum in our word sets, and complementary cues would probably help avoiding massive ambiguity.

Despite the limitations of the present study, the system presents a number of interesting properties. It belongs with other direct, continuous and on-line activation models. As in the TRACE model, onset alignment dominates even though no explicit alignment constraint has been implemented. However, the mapping mechanisms differ in two main respects. While TRACE uses a fixed architecture based on spatial reduplication to represent the time dimension, and cannot cope with varying rates of speech, the SRN dynamically encodes temporal information as a compressed distributed representation. Secondly, while in TRACE, lexical activation depends on both sensory information and lateral inhibition, the SRN involves no direct competition at the level of response units. Lexical units activation is entirely determined by bindings between the probes and the distributed trace forged by the recurrent loop of the network.

In a recently published discussion paper, the following question was put forth: "If the processor operates sequentially and receives the acoustic-phonetic information relative to an upcoming word while still engaged in the 'selection' of the preceding one, how can it identify that information on-line as word initial and use it to build the cohort that is so crucial to recognising the new word?" (Tabossi, 1993, p. 289). The SRN with multiple probes appears to provide an original solution to this problem. The multiple task used makes it possible for the system to rapidly extract hypotheses about the current word, which might be utilised by other components of the language processor, as soon as they get sufficient strength. When the information about lexical hypotheses is not clear, it suffices to wait a while, but the uncertainty about the current word offset point does not prevent the network from rapidly isolating the following word.

Acknowledgments

Thanks to Jeff Elman for providing the Tlearn simulator, to Uli Frauenfelder for stimulating discussions and to Axel Cleeremans for comments on a previous draft. The present research benefited from funds from the French Community of Belgium (ARC "Language Processing in Different Modalities: Comparative Approaches").

References

Connine, C.M., Blasko, D.G. & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: temporal and linguistic constraints. *Journal of Memory and Language*, 30, 234-250.

Cutler, A. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.

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

McClelland, J.L. & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

Frauenfelder, U.H. & Peeters, G. (1990). On lexical segmentation in TRACE : an exercise in simulation. In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing* (pp. 50-86). Cambridge, Mass.: M.I.T. Press.

Frauenfelder, U.H. & Peeters, G. (1993). *Simulating the time-course of word recognition: an analysis of lexical competition in TRACE*. Unpublished manuscript.

Grosjean, F. (1985).The recognition of words after their acoustic offset: evidence and implications. *Perception and Psychophysics*, 38, 299-310.

Harrington, J. & Johnson, A. (1987). The effects of equivalence classes on parsing phonemes in to words in continuous speech recognition. *Computer Speech and Language*, 17, 134-140.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. Perception and Psychophysics, 39, 155-158.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25, 71-102.

Marslen-Wilson,W.D. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10, 29-63.

Norris, D. (1990). A Dynamic-net model of human speech recognition. In Altmann, G.T.M. (Ed.) *Cognitive Models of Speech Processing* (pp. 87-104). Cambridge, Mass.: M.I.T. Press.

Radeau, M., Mousty, P. & Bertelson, P. (1989). The effect of the uniqueness point in spoken-word recognition. *Psychological Research*, 51, 123-128.

Radeau, M. & Morais, J. (1990). The uniqueness point effect in the shadowing of spoken words. *Speech Communication*, 9, 155-164.

Rietveld, A.C.M. (1980). Word boundaries in the French language. *Language and Speech*, 23, 289-296.

Samuel, A. G. (1990) *Perception delayed is perception refined : retroactive context effects in speech perception*. Unpublished manuscript.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J.L. (1991). Graded-state machines: the representation of temporal contingencies in simple recurrent networks. *Machine Learning*, 7, 161-193.

Shillcock, R. (1990). Lexical hypotheses in continuous speech. In Altmann, G.T.M. (Ed.) *Cognitive Models of Speech Processing* (pp. 24-49). Cambridge, Mass.: M.I.T. Press.

St John, M. & McClelland, J. L. (1990). Leaning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-257.

Tabossi, P. (1993). Connections, competitions and cohorts : comments on the chapters by Marslen-Wilson; Norris; and Bard and Shillcock. In G.T.M. Altmann and R. Shillcock (Eds.), *Cognitive models of speech processing. The second Sperlonga meeting* (pp. 277-294). Hillsdale, NJ: Erlbaum.