Representing the bilingual's two lexicons.

Michael S.C. Thomas
Department of Experimental Psychology,
University of Oxford,
South Parks Road, Oxford OX1 3UD.
UK.
michael.thomas@psy.ox.ac.uk

Kim Plunkett
Department of Experimental Psychology,
University of Oxford,
South Parks Road, Oxford OX1 3UD.
UK.
plunkett@psy.ox.ac.uk

Abstract

A review of empirical work suggests that the lexical representations of a bilingual's two languages are independent (Smith, 1991), but may also be sensitive to between language similarity patterns (e.g. Cristoffanini, Kirsner, and Milech, 1986). Some researchers hold that infant bilinguals do not initially differentiate between their two languages (e.g. Redlinger & Park, 1980). Yet by the age of two they appear to have acquired separate linguistic systems for each language (Lanza, 1992). This paper explores the hypothesis that the separation of lexical representations in bilinguals is a functional rather than an architectural one. It suggests that the separation may be driven by differences in the structure of the input to a common architectural system. Connectionist simulations are presented modelling the representation of two sets of lexical information. These simulations explore the conditions required to create functionally independent lexical representations in a *single* neural network. It is shown that a single network may acquire a second language after learning a first (avoiding the traditional problem of catastrophic interference in these networks). Further it is shown that in a single network, the functional independence of representations is dependent on inter-language similarity patterns. The latter finding is difficult to account for in a model that postulates architecturally separate lexical representations.

Introduction

Studies involving children learning two languages indicate that even as young as two years old they are aware that there are two distinct languages present in the linguistically mixed environment to which they are exposed. These children acquire two separate language systems, and can be observed to switch between the use of their two languages in a coherent fashion, depending on the linguistic context negotiated with their parents (Lanza, 1992). Adult bilinguals show a high degree of skill in using either of their integrated language systems, appearing to be able to set aside one of their systems while operating in the other. The impression in both of these cases is of functionally separate language systems.

Research in the language processing of adult bilinguals has investigated how the bilingual's two language systems may be represented in his or her cognitive system. There are several theories as to the relation of the two systems, but the majority view is that there are separate lexical representations for each language, but combined semantic representations (see Smith, 1991, for a review). The evidence for this view comes mainly from repetition and semantic priming effects. Repetition priming effects are obtained from tasks such as lexical decision, word fragment completion, and perceptual identification. Semantic priming effects come mainly from recall performance and classification tasks. Conclusions are based on the assumption that if one task serves as a prime for a second, then it is accessing the same underlying representation. Although findings are mixed, suggesting a sensitivity to experimental conditions and task design (Durgunoglu and Roediger, 1987), it appears that tasks which access semantic representations allow priming between the bilingual's two languages (e.g. Caramazza and Brones, 1980; Kolers and Gonzalez, 1980; MacLeod, 1976), but those that access lexical information alone allow only priming within each language (e.g. Scarborough, Gerard, and Cortese, 1984; Watkins and Peynircioglu, 1983). For example, in the case of a French-English bilingual, if in some task, *chien* were followed by *dog,* the response time to *dog* would be reduced if the task involved, say, semantic classification, but not if it involved, say, lexical decision.

There is also evidence to suggest that word similarity plays a role in bilingual lexical processing. For example, if a word exists in both languages (such as *pain* in French and English) or is morphologically similar in each language, between language priming effects have been found (Cristoffanini, Kirsner, and Milech, 1986; Gerard and Scarborough, 1989; Kerkman, 1984). Between language interference at the lexical level has been found for words that are legal in both languages but not for those that have characteristics unique to each language (Grainger and Beauvillain, 1987). This evidence implies that the lexical representations of each language may not be as distinct as previously thought.

In this paper, we will explore the hypothesis that bilingual lexical representation is best accounted for using a model that stores both languages in a single network. Recent work within the connectionist framework has shown that a functional separation in psycholinguistic processing need not be taken to imply separation at the level of mechanism (Rumelhart and McClelland, 1986; Plunkett and Marchman, 1993). Here, we examine the possibility that, for bilinguals, the description of their overall language system as having two lexicons may merely be a functional one, and that the evidence to date need not necessarily imply the existence of two physically separate structures. We will show that:

1. Two sets of lexical information can be stored in the same network, even when training on the second set follows that on the first (modelling the case of second language acquisition).
2. Second language learning in a single network device can be achieved without catastrophic interference from the second language.
3. The network is sensitive to the similarity of words in the two languages. Words that are dissimilar show more functional independence than those that are similar.

Learning two lexicons in a single network.

The independence of the representations underlying lexical processing implies that there is no interference between these representations. Learning independent representations is easily achieved in a neural network by training on simultaneously presented orthogonal languages (i.e. those which have no features in common). However in the case of second language learning, where exposure to the second language occurs only after the first has been learned, neural networks experience the problem of Catastrophic Interference (CI). If training on one set of patterns ceases and training on another inconsistent set commences, information about the first set may be overwritten. This disruption occurs since the same connection weights are required to do a different job in learning the mappings for each pattern set. With simultaneous training, the network has the opportunity to find a set of weights that can do both jobs. But with sequential training, the connections responsible for learning the first set are changed so that they can learn the second, and this may damage the network's performance on the first set. In terms of lexical representation, CI would translate into a second language learner overwriting their first language with their second, which clearly does not happen.

Standard solutions to CI have involved orthogonalizing the representations for each input/output mapping, so that they no longer use the same connections. Each new mapping can be learned separately without disrupting any that have been learned before (see Sharkey & Sharkey, 1994, for a review of these techniques). However, such solutions cannot be learned using a standard backpropagation network, and more seriously, the ability of these networks to generalize between patterns is lost. It is important to retain the generalization between patterns if we are to capture the empirical data on lexical processing. We require another solution to CI that preserves generalization *within* lexicons but not between them.

A Dual Route Model.

Within lexicon generalization would be easy to achieve in a dual route model. A separate network would be devoted to the representation of each lexicon. Generalization would occur within each route, but since the routes were physically separated, no generalization would take place between the lexicons. A simple dual route model, however, fails to capture important characteristics of the data, namely that interlanguage similarity is significant in establishing the independence of the representations. Furthermore, from the perspective of bilingual acquisition, no account is provided of how the child discovers that there are two languages in its linguistic environment and hence determines the need for a dual-route representation.

Many researchers hold that infant bilinguals do not initially differentiate the two languages (e.g. Redlinger and Park, 1980; Vihman, 1985). The one-system view of bilingual development supposes that the bilingual child must undergo a process of language differentiation through which two separate linguistic systems are gradually formed. If a child is to assign two routes to its system at a point when it detects that there are two languages present, it would need have some firm basis on which to make such a judgement. The child should not, for instance, assign two routes to its system merely because its parents have different accents, or slightly different vocabularies. Yet it must, if they are using different languages. Identifying the presence of two (or for that matter three) languages is not a trivial matter, especially if they are closely related. The decision to construct multiple lexical representations would, therefore, appear to be contingent upon a careful analysis of the characteristics of the ambient linguistic environment.

The one-system view of bilingual development could, on the other hand, be taken to suggest that the child starts out with a single mechanism underlying its language learning, but that this mechanism develops representations which come to exhibit functional independence.

A Single Route Model.

A single route model of bilingual and lexical representation would suppose that a single mechanism underlies lexical processing in both children and adults. In the child, a single route model need make no assumptions about differentiation of the two languages at the onset of learning. Initially, the languages are treated in an identical fashion and are only differentiated by learning the pattern of characteristics unique to each language. This suggests that the separation of representations may be driven by differences in the structure of the input to a common architectural system. In the adult, the bilingual lexicons may have achieved a status of *functional modularity* within the single mechanism though residual patterns of interference between the two languages may be observed for items that are similar in both linguistic systems.

The issue as to whether single or dual mechanisms are involved in bilingual language processing is not merely a theoretical nicety. Each account has different implications for our understanding of how languages are learned and for the patterns of errors and mastery observed *en route* to the mature adult system. Even in the mature adult state, residual traces of the moulding forces of development can still be observed and used to reconstruct that process.

Simulations.

Two simulations are presented in this paper. The first simulation demonstrates that a single network can store information about two separate lexicons even when training on the second lexicon follows that on the first. In effect, this simulation offers a solution to the problem of CI between distinct training sets in a network, while maintaining the desirable property of generalization within a training set. The solution requires no domain specific modifications to the backpropagation algorithm. The second simulation addresses the problem of simultaneous acquisition of two partially overlapping lexical systems. It demonstrates how the functional independence of lexical items depends on their inter-language similarity. It is argued that these results offer a plausible simulation of patterns of lexical priming in adult bilinguals and language differentiation in the language learner.

Architecture.

The simulations used 3-layer feedforward networks, trained using the backpropagation algorithm. The networks were set the task of autoencoding two sets of language-like information. In the autoencoding task, a network usually has fewer hidden units than input or output units. This hidden unit 'bottleneck' forces the network to create a more efficient representation of the sets of input patterns by removing unnecessary redundancies. If the network is required to learn two input sets, it must therefore discover the features that characterize both those input sets.

Training sets.

Two sets of items were constructed for each simulation, corresponding to the lexical representations of simple words in different languages. The input sets were constructed using orthogonal representations for the encoding of individual lexical constituents. This form of representation provided a stringent method for controlling word similarity in each pair of languages. The orthogonal word constituents can be thought of as corresponding to either letters or phonemes.

Each language was defined in terms of its own set of orthographic/phonological rules. The constituents were classified as consonants or vowels. Only certain combinations of consonants and vowels were permissible in each language. These rules were then used to generate tokens in the language. The words were 3 letters/phonemes long with a choice of 10 letters/phonemes in each position. Since we used an orthogonal representation for the letters/phonemes, the network required 30 input and output units to code the lexical information.

Language Specific Units.

Bilinguals can selectively access information about either of their two languages. If two languages are to be stored in a single network, information specifying language membership must therefore be associated with each item. We can think of each item as being tagged for membership in a language on the basis of language specific features available to the language learner. For example, in spoken French there is a tendency to nasalize phonemes, and Chinese is a tonal language.

The information which distinguishes the two languages is presented to the network on an extra set of Language Specific Input Units. Since the task is autoencoding, there is a corresponding set of output units. Two orthogonal vectors are used to represent this information. For example, if there were two Language Specific Units, a vector of (1, 0) might index one language, and a vector of (0, 1) the other. Figure 1 shows the network architecture.

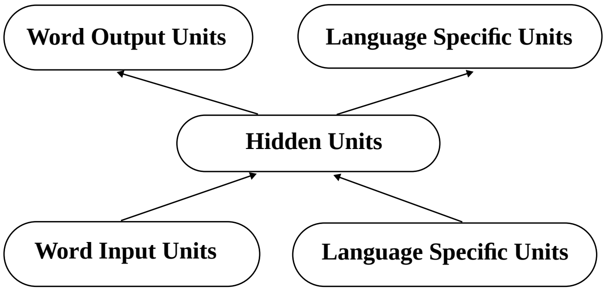


Figure 1: Network Architecture.

Simulation 1.

Modelling second language acquisition in a single network demands that we avoid disruption of the performance on the first input set by training on the second. This disruption is maximized when no information learned during training on the first set is useful in learning the second set. There is no advantage for the network to retain information about the first input set, so the connections are changed maximally to learn the second set. Simulation 1 explored the effect of varying the amount of language tagging information available to the network. This manipulation provided the opportunity to evaluate the relative success of second language acquisition under conditions where inter-language differences could be easily quantified.

To model the hardest case for second language acquisition, two languages were constructed which were based on the same orthographic/phonological rules but which used different letters/phonemes. Each language comprised 32 words. The rules and alphabets used are shown in Figure 2. The network had the architecture shown in Figure 1, with 20 hidden units used to enforce the representational bottleneck.

The network was trained to autoassociate the words in L1. When the error asymptoted, training on L1 ceased, and training on L2 commenced. The mean squared error on the Word Output Units (see Figure 1) was measured for each language

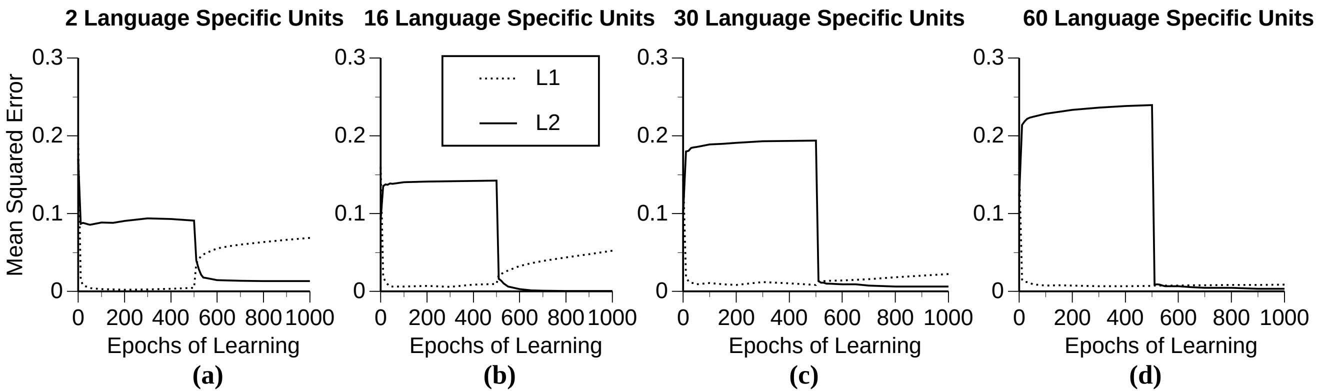


Figure 3: Graphs a) to d) show the elimination of Catastrophic Interference as the number of Language Specific Units is increased

throughout this process. Simulations were repeated at four levels of language specific tagging, namely where the number of units tagging each language was *1*, *8*, *15*, and *30*.

L1:	**L2:**
Alphabet:	**Alphabet:**
Vowels: o, i.	Vowels: a, e.
Consonants: f, p, g.	Consonants: b, t, c.
Rules: CVV, CVC, VCV, VVC	

Figure 2: Languages used in Simulation 1.

Results.
The resulting learning profiles are shown in Figure 3a) to d). Each figure plots the mean squared error on the Word Output Units for all items in a training set, with the error for each language plotted separately. Figure 3a) shows the error for each language when the number of language separating units is set to two. L1 improves in performance until it is virtually error free. When training on the first language ceases and the second language is introduced (after 500 epochs of learning), the error on L1 increases while that on L2 gradually disappears. In other words we observe catastrophic interference from L2 on L1. However, as the number of language specific units is increased, as shown in Figure 3b) to Figure 3d), there is a decreasing level of catastrophic interference from L2 on L1 during the second phase of training. By the time the number of language specific units reaches 30 per language, the catastrophic interference has disappeared.

Discussion.
These results show that a single network can learn two lexicons when training on the second follows that on the first, *provided there is sufficient language specific information to separate the languages.*

Increasing the amount of language specific information eliminates catastrophic interference because it allows the backpropagation algorithm greater scope to develop orthogonal internal representations for each language. The language specific information that tags language membership biases the network's interpretation of the input sets so that the representations it forms for each are quite different. Since these representations are different, the weights from which they emerge tend to be different: training on L2 no longer tends to change weights that contain information about L1. If one imagines the network defining a representational space, then the bias that the language specific information provides allows the network to partition this space and place each language in a different partition.

Although the representational resources to avoid catastrophic interference are costly, there is little evidence to suggest that the neural substrate for language acquisition is a limiting factor. For example, there is evidence that children have up to 50% more brain cell connections than adults (Collins & Kuczaj, 1991, p.50).

The simulation outlined here depicts the most extreme case of second language acquisition, where learning the second language occurs without any further exposure to the first language. Second language acquisition occurs more often in the context of continued first language usage, serving as additional language exposure rather than as a replacement. This simulation shows that catastrophic interference is avoidable in the most extreme case. The more usual case would be easier for the network to solve, since input sets would be trained with an element of simultaneity.

The simulation does not, however, offer us the opportunity to examine the role of inter-language similarity in bilingual language processing, since by design there is no similarity between the languages at all. This was the goal of the next simulation.

Simulation 2.

This simulation demonstrates two points. Firstly, similarity is important in establishing the independence of lexical representations in a single network. Secondly, as well as simulating data difficult to account for in the simple dual route approach, a single network simulates human data previously taken as strong evidence of separate lexical representations.

Similarity Effects.

Two languages were constructed which shared the same letters/phonemes and 3 out of 4 orthographic/phonological rules. The rules are shown in Figure 4. Words in each language could be categorized into three classes in the following way:

(1) Those that exist in both languages.

(2) Those that are legal in both languages but exist in only one language.

(3) Those that exist and are legal in only one language.

Examples taken from French and English fulfilling these criteria are (1) *pain* & *pain*, (2) *trop* & *time*, and (3) *soeur* & *cough* respectively. The word classes were constructed to reflect a similarity gradient between the languages.

Each language comprised 78 words. The network was trained on both languages simultaneously. To reflect the greater similarity of the languages, only eight units were used to code the language specific information. In order to store the greater number of patterns, 45 hidden units were used[1].

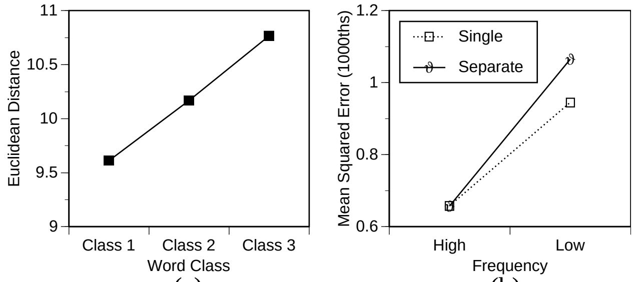
Alphabet: Vowels: a, e, i, o, u.	Consonants: s, t, b, g, p.
L1 Rules:	**L2 Rules:**
CVV, CVC, VCV, VVC.	CVV, CVC, VCC, VVC.

High frequency: 4 duplications in the input set.
Low frequency: 1 token in the input set.

Figure 4: Languages used in Simulation 2.

Results.

Figure 5a) summarizes how the network represents words from the three classes outlined above. The hidden unit activation obtained when a word is presented to the network may be thought of as a point in representational space. For each class, the hidden unit activations for all the words in that class were averaged together to define its ‘centre of gravity’ in representational space. This graph shows the distance between the centres of gravity for the same classes in each language. It shows that words that exist in both languages (Class 1) are closest together in this space. Words that exist and are legal in only one language (Class 3) are furthest apart. Words that exist in one language but are legal in the other (Class 2) represent an intermediate case.



(a) (b)

Figure 5: a) Distance between Word Classes; b) Class 1: Error by frequency

Discussion.

Studies that have investigated bilingual lexical representation (and the importance of similarity in processing) have used priming paradigms. In the context of the network models employed here, priming may be thought of as activation persisting in a network. The more similar the activation is between two words, the more likely it is that activation persisting from processing one word will facilitate processing of the next word, and therefore serve as a prime for the next word. The distance between two points in representational space is a measure of the similarity of two patterns of activation. Figure 5a) can therefore be interpreted as a measure of the degree to which words within a class will prime words in the same class in the other language.

Recall that in experiments using the lexical decision task, priming between languages was found for words that existed in both languages, i.e. Class 1 words (e.g. Gerard and Scarborough, 1989). Cristoffanini et al (1986) found between language priming to a slightly lesser degree for morphologically similar but not identical words, corresponding to Class 2 words. Scarborough et al (1984) found no priming for orthographically distinct translation equivalents, corresponding roughly to Class 3 words.

This simulation shows that the similarity gradient in the input between the languages translates into the functional independence of the representations.

Frequency Effects.

Words in Class 1 may have a different meaning in each language (for example *pain* means bread in French). In such cases, it is likely that the same lexical item will have a different frequency in each language. Using such words, Gerard and Scarborough (1989) showed that Spanish-English bilinguals responded in a lexical decision task according to the *within language* frequency. They interpreted this evidence as favouring the view that the lexical representations for each were independent.

To examine this issue in the model, words were defined as having a high or low frequency. High frequency words were presented to the network four times as often during training. For the words existing in both languages, half were high frequency in L1 and low frequency in L2, the other half high frequency in L2 and low frequency in L1. In networks modelling lexical representation, it has been shown that the error score which results when a word is presented to a network may in some circumstances be interpreted as equivalent to a subject’s reaction time in the lexical decision task (Seidenberg and McClelland, 1989). In this part of the simulation, we examined the error score for words in Class 1 in both languages. An additional simulation was performed in an attempt to control for absolute levels of error score separating high frequency and low frequency words. This control

[1] Other simulations indicated that increasing the number of hidden units actually served to partition the representational space less adequately. Greater numbers of hidden units led to the languages being less separated in this space, and thus less independent.

involved training each language on a separate network with 30 input and output units and 22 hidden units.

Results & Discussion.

Figure 5b) shows the mean squared error when the network was tested on the Class 1 words, split by frequency. Since both languages show the same pattern of results, we depict the error scores for both languages averaged together. The results from the control simulation are shown on the same graph. This figure shows that performance on these words varies with within language frequency. Lower error scores are observed for high frequency words, even though these words have the same form and are stored in the same network. There is only a small difference between the error scores for the single and separate network solutions. We interpret the lower error scores for high frequency words as indicating faster response times to high frequency words than low frequency words. Hence within language frequency effects can be observed even when both languages are represented in the same device.

Conclusions.

In this paper, we have offered a solution to the logical problem of how a child can simultaneously acquire two languages without the need for innate assumptions concerning the cognitive architecture required to represent these languages, yet as adults show behavior suggesting separate routes for the processing of lexical information in each language. Evidence for independent lexical representations may be taken as a functional description of the behavior of a single mechanism. Furthermore, second language acquisition can be achieved in a single network whilst avoiding the potential problem of catastrophic interference. Importantly, the single network model can account for data that a simple dual route model cannot account for, namely the role of similarity in lexical processing, and can also simulate sensitivity to within language frequency—a finding that has previously been taken as strong evidence for separate lexical representations in bilinguals.

Acknowledgements

The work in the paper is supported in part by a BBSRC grant. We are grateful to Sharon McHale, Denis Mareschal, Neil Forester, and Chrys Meula.

References

Caramazza, A. & Brones, I. (1980). Semantic classification by bilinguals. *Canadian Journal of Psychology, 34,* 77-81.

Collins, W. A. & Kuczaj, S. A. (1991). *Developmental Psychology: Childhood and Adolescence.* Macmillan: New York.

Cristoffanini, P., Kirsner, K., & Milech, D. (1986) Bilingual lexical representation: The status of cognates. *Quarterly Journal of Experimental Psychology, 38A*, 367-394.

Durgunoglu, A. Y. & Roediger, H. L. (1987). Test differences in accessing bilingual memory. *Journal of Memory and Language, 26,* 377-391.

Grainger, J. & Beauvillain, C. (1987). Language blocking and lexical access in bilinguals. *Quarterly Journal of Experimental Psychology, 39A,* 295-319.

Kerkman, J. P. M. (1984). Word recognition in two languages. In A. Homassen, L. Noordman, & P. Eling (Eds.), *The Reading Process.* Lisse: Swets Zeitlinger.

Kolers, P. A. & Gonzalez, E. (1980). Memory for words, synonyms and translations. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 53-65.

Lanza, E. (1992). Can bilingual two-year-olds code-switch? *Journal of Child Language, 19,* 633-658.

MacLeod, C. M. (1976). Bilingual episodic memory: Acquisition and forgetting. *Journal of Verbal Learning and Verbal Behavior, 15,* 347-364.

Plunkett, K.R. & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition, 48,* 1-49.

Rumelhart, D.E. & McClelland, J.L. (1986). On learning the past tense of English verbs, in McClelland, J.L., Rumelhart, D.E., and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.2,* (Ch.21). MIT Press.

Redlinger, W. & Park, T. Z. (1980). Language mixing in young bilingual children. *Journal of Child Language, 7,* 337-352.

Scarborough, D. L., Gerard, L., & Cortese C. (1984). Independence of lexical access in bilingual word recognition. *Journal of Verbal Learning and Verbal Behavior, 23,* 84-99.

Seidenberg M. S. & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96.* 523-568.

Sharkey, N.E. & Sharkey, A.J.C. (1994). Understanding Catastrophic Interference in Neural Nets. (Technical Report). Department of Computer Science, University of Sheffield, U.K.

Smith, M. (1991) On the recruitment of semantic information for word fragment completion: evidence from bilingual priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 234-244.

Vihman, M. (1985). Language differentiation by the bilingual infant. *Journal of Child Language, 12,* 297–324.

Watkins, M. J. & Peynircioglu, Z. F. (1983). On the nature of word recall: Evidence of linguistic specificity. *Journal of Verbal Learning and Verbal Behavior, 22,* 385-394.