J. Child Lang. **26** (1999), 545–575. Printed in the United Kingdom

The learning of first and second person pronouns in English: network models and analysis*

YURIKO OSHIMA-TAKANE, YOSHIO TAKANE AND THOMAS R. SHULTZ

Department of Psychology, McGill University

ABSTRACT

Although most English-speaking children master the correct use of first and second person pronouns by three years, some children show persistent reversal errors in which they refer to themselves as *you* and to others as *me*. Recently, such differences have been attributed to the relative availability of overheard speech during the learning process. The present study tested this proposal with feed-forward neural networks learning these pronouns. Network learning speed and analysis of their knowledge representations confirmed the importance of exposure to shifting reference provided by overheard speech. Errorless pronoun learning was linked to the amount of overheard speech, interactions with a greater number of speakers, and prior knowledge of the basic-level kind PERSON.

INTRODUCTION

Learning the semantic rules for first and second person pronouns poses problems for young children because the referent of these pronouns shifts with speech roles, and because a model for correct use of these pronouns is not provided in speech addressed to the child (Oshima-Takane, 1985, 1988). When a mother talks to her child, *me* refers to herself, and *you* to the child. However, when the child talks to the mother, *me* refers to the child, and *you* to the mother. Yet most children master the correct use of these pronouns in English by the age of three (Clark, 1978; Oshima-Takane, 1985). Previous psycholinguistic research has documented that there are individual differences in pronoun acquisition in English-speaking children (Clark, 1978; Oshima-Takane, 1985, 1992; Chiat, 1986). A majority of normally developing children master the correct usage of these pronouns with few errors, whereas some produce persistent reversal errors (i.e. producing *you* instead of *me* or *me* instead of *you*) in the course of acquisition.

Although several different accounts have been proposed for children's

[*] This research was supported by the Natural Sciences and Engineering Research Council of Canada. We thank David Buckingham, Sylvain Sirois, Sheldon Tetewsky, Marina Takane and David Nicolas for helpful comments. Address for correspondence: Yuriko Oshima-Takane, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, Quebec, H3A 1B1.

initial hypothesis about the meaning of the personal pronouns, none of the previous theories has accounted for the variations in pronoun acquisition (Shipley & Shipley, 1969; Clark, 1978; Charney, 1980; Schiff-Myers, 1983; Chiat, 1986). These theories have focused either on children's pronominal errors or on the lack of pronominal errors. Oshima-Takane's account (1985) is the first to explain why some children make persistent errors, whereas a majority do not. Based on her theoretical analysis of pronoun learning, Oshima-Takane (1985) hypothesized that children learn the correct semantic rules for personal pronouns by observing the shifting reference of pronouns used in speech not addressed to them (non-addressed speech), whereas they learn the incorrect semantic rules by observing pronouns used in each speech addressed to them (addressed speech). She argued that, in addressed speech, children simply observe that second person pronouns refer to themselves and that first person pronouns refer to the person who is speaking to them. Thus, children are more likely to entertain incorrect semantic rules that second person pronouns refer to themselves and first person pronouns refer to the person talking to them. Consequently, they would show persistent reversal errors when they use these pronouns. In non-addressed speech, on the other hand, children often observe that second person pronouns refer to a person other than themselves and that first and second person pronouns shift systematically. In this case children are more likely to induce the correct rules that first person pronouns refer to the person who is using them and that second person pronouns to the person addressed. They would consequently produce correct pronouns without errors.

In support of her theoretical analysis, Oshima-Takane (1985, 1988) conducted a training experiment with children at 1;7 who were about to learn personal pronouns. The results indicated that children benefit in pronoun production from non-addressed speech. In fact, only those children who had opportunities to hear pronouns in non-addressed speech could produce pronouns without errors. A subsequent observational study done by Oshima-Takane, Goodz & Derevensky (1996) provided naturalistic evidence consistent with her experimental finding. The study demonstrated that secondborn children produced correct pronouns earlier than firstborns, even though these children did not differ on other language measures such as mean length of utterance and vocabulary size. They argue that secondborn children acquire the correct usage of pronouns earlier than firstborn children because they have relatively more opportunities to hear pronouns used in non-addressed speech, that is, in conversation between their parent and older sibling. Other observational as well as experimental studies done by Oshima-Takane and her collaborators also provided converging evidence in support of her theory (Oshima-Takane & Benaroya, 1989; Oshima-Takane & Oram, 1991; Oshima-Takane, 1992; Oshima-Takane, Cole & Yaremko, 1993; Cole, Oshima-Takane & Yaremko, 1994).

Recent computer simulations by Shultz, Buckingham & Oshima-Takane (1994) simulated these psychological findings using the cascade-correlation (CC) learning algorithm (Fahlman & Lebiere, 1990).[1] They found that networks trained by non-addressee parent-speaking training patterns were quick to learn the correct rules, whereas networks trained by addressee parent-speaking training patterns learned incorrect rules and showed persistent reversal errors. In addition, the more the networks were exposed to the shifting reference of first and second person pronouns in non-addressed speech, the faster the mastery of the correct rules. However, unlike human children, none of the networks in their study could learn to produce the correct pronouns without requiring explicit error-correcting feedback, even though they were trained by non-addressed speech patterns.

A primary motivation for the present network simulations is to understand the mechanisms by which children learn to produce these pronouns correctly without explicit corrections. In the previous computer simulations, speech role information (i.e. who is the speaker and who is the addressee) and the referent information (i.e. who is the referent of the pronoun) were implemented as the minimum prior knowledge for learning the first and second person pronouns. The present simulation study investigated whether the implementation of other prior knowledge, as well as the opportunity to overhear more speakers, would facilitate learning without explicit corrections.

Previous theoretical analysis of pronoun learning has emphasized the importance of the basic-level kind category PERSON in learning the meaning of personal pronouns (Oshima-Takane, 1985; Macnamara, 1986; Macnamara & Reyes, 1994). For instance, Macnamara (1982, 1986) argued that unless the child understands the basic-level kind PERSON, the child would not be able to understand that a personal pronoun refers to a person, not just a person's face, nose, or visible surface. Furthermore, the notion of the kind PERSON would help the child to pick out a person from all other animate and inanimate objects as the referent of the pronoun and to make a correct generalization to any member of the kind PERSON. Unless the child knows that he/she is also a member of the kind PERSON, the child would not be able to realize that he/she could also use first person pronouns in reference to him/herself. Recent evidence suggests that children can make global kind distinctions (e.g. animal/artifact, vehicle/animal) and basic-

[1] The CC algorithm is one of the class of so-called generative learning algorithms that build their own topology by recruiting new hidden units as needed. Therefore, it affords a more principled approach to network construction than static learning algorithms such as back-propagation that require a full specification of the network. Another important feature of CC is that it uses second order error minimization techniques in computing weight changes and learns only one level at a time. Thus, it is typically 10 to 50 times faster at learning than is back-propagation. See Section 2 for a more detailed description.

level kind distinctions (e.g. bottle/ball, cup/book) by 12 months of age (Mandler, Bauer & McDonough, 1991; Xu & Carey, 1996; Sorrentino, 1999). Then, it seems reasonable to assume that children can recognize an individual including themselves as a member of the kind PERSON by the time they learn personal pronouns.

In the previous computer simulations (Shultz, Buckingham & Oshima-Takane, 1994), three persons appeared in the training patterns (mother, father and child) and their identities were coded in an arbitrary, binary fashion distributed over two input units. The child was coded as 11, the mother as 10, and the father as 01. They were coded as distinct individuals but not explicitly as a member of the same kind. In subsequent simulations (see Shultz, Schmidt, Buckingham & Mareschal, 1995 for a discussion of the preliminary results) the distinction between 'self' and 'other' was explicitly coded by adding a unit for each individual to see if this additional information would facilitate the correct production of first and second person pronouns. But again individuals were not coded as a member of the same kind due to limitations of the distributed binary coding used.[2] The results indicated that the addition of 'self/other' information did not enable immediately correct generalization to child-speaking patterns for networks trained with non-addressed speech patterns. Generalization tests immediately after non-addressee training revealed that none of the networks produced the first person pronoun *me* in reference to the child (self), although some did produce the second person pronoun *you* in reference to mother and father (other) correctly. Overall, networks exposed to non-addressee speech patterns required only about 12 epochs of child-speaking patterns to master the correct rules, about one-tenth of the additional training required by networks trained initially on addressee speech patterns. The lack of information that the child (self) as well as mother and father (other) belongs to the same kind might have been the reason why none of the networks trained by non-addressee speech patterns produced correct first person pronouns without errors. In the present simulations, we use analogue coding to represent the individual members and the kind to which they belong by assigning a number to the members on the same unit.

Experiencing many examples involving various persons may be another important factor in learning correct semantic rules of the first and second person pronouns. In the previous simulations, networks learned parent-speaking patterns in which only the mother and the father were using the pronouns (Phase I) before they learned child-speaking patterns (Phase II).

[2] With distributed binary coding used in the previous simulations (Shultz, Buckingham & Oshima-Takane, 1994), explicit coding of PERSON is not informative for the networks because individuals appearing in the training patterns were all persons. Non-persons could not be added to the training patterns because they do not take speaker and addressee roles.

Children learning personal pronouns normally also hear persons other than parents using pronouns (e.g. older siblings, grandparents, babysitters, etc.) even though their parents' utterances are the major source of input they hear. Overhearing multiple speakers may help children realize that a pronoun not only refers to a specific person who has interacted with them but also refers to any person depending on the speech role. In the present simulation study we included conditions in which two other persons were added to Phase I training patterns in order to investigate whether exposure to more speakers would facilitate the learning of the correct rules.

In this paper we report four computer simulations using the CC learning algorithm. In simulation 1, prior knowledge of the basic-level kind PERSON was added to the pure addressee and non-addressee conditions. There was a maximum of three persons appearing in the training patterns as in the previous simulations (Shultz, Buckingham & Oshima-Takane, 1994). In the present simulation we investigated whether networks subjected to pure non-addressee training could produce correct child-speaking patterns without explicit corrections but with this prior knowledge. In simulation 2, two additional persons were included in the training. By comparing the network's production of child-speaking patterns across simulations 1 and 2, we tested whether the addition of two other persons in the training patterns would facilitate correct production of the child-speaking patterns. In simulation 3 two mixtures of addressee and non-addressee materials were included in addition to the pure addressee and non-addressee conditions in order to more realistically simulate the child's natural language learning environment. By comparing across simulations we assess the extent to which non-addressed speech is needed for the learning of the semantic rules for the personal pronouns. In addition, network analysis was performed to understand what the networks learn at various points in the acquisition process. In particular, we examined the network's knowledge representations and generalization capability to determine whether the networks have arrived at highly abstract generalizations or if the networks have simply memorized correct responses to the training situations. Simulation 4 is a remedial study of one of the pure addressee trained networks which have learned incorrect semantic rules in simulation 3. We examine the effects of additional remedial training, with different degrees of non-addressee materials, in unlearning the incorrect rules. Network analysis is also performed to understand how these networks unlearn incorrect rules and eventually learn the correct ones.

SIMULATIONS

The acquisition of first and second person pronouns can be regarded as a special kind of nonlinear function learning because the referent of the pronoun shifts with speech roles. The semantic rules involve an interaction between referent and speech roles.[3] We investigated how feed-forward

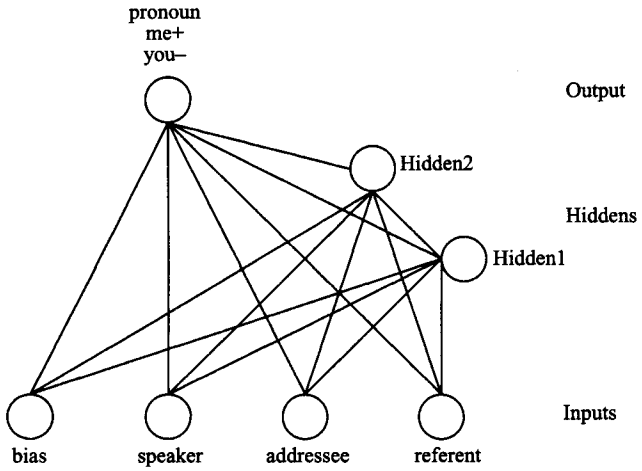

Fig. 1. Pronoun network after recruitment of two hidden units.

networks approximated the nonlinear function underlying pronoun learning. We used the CC algorithm because it is particularly good at capturing interaction effects among input variables without being told which interactions are important. Furthermore, this algorithm dynamically grows networks to approximate increasingly more complicated functions, thus allowing for the growth in representational power that is often assumed to underlie human development (Shultz *et al.* 1995).

In CC learning, no *a priori* net topology has to be specified. Each network starts without hidden units, and hidden units are added to improve its performance until a satisfactory degree of performance is reached. Hidden units are added one at a time so that all pre-existing units are connected to new ones. The topological changes in the network may define distinct developmental stages in learning. Cross connections that bypass hidden layers are used and often simplify the solutions by capturing linear effects in the simplest way. Hidden units with sigmoid activation functions produce nonlinear, interactions effects in the mapping of inputs to outputs. When a new hidden unit is recruited, incoming weights to the new unit are determined by increasing the correlation of the unit's activation with network error, and are fixed throughout the rest of the learning process. This avoids

[3] In the present paper we did not employ the traditional semantic feature analysis of personal pronouns. Instead, we adopted Kaplan's (1978) formulation of semantic rules for first and second person pronouns which Oshima-Takane's (1985) pronoun learning model was based on. That is, (1) a first person pronoun, in each utterance, refers to the person who uses it, and (2) a second person pronoun, in each utterance, refers to the person who is addressed when it is used.

the necessity of back-propagating error across different levels of the network, and leads to faster and more stable convergence. The weights associated with output connections are, however, re-estimated after a new hidden unit is recruited. The CC algorithm has been successfully applied to a number of problems in cognitive development, including the balance scale (Shultz, Mareschal & Schmidt, 1994), seriation (Mareschal & Shultz, 1993), the integration of velocity, time and distance cues (Buckingham & Shultz, 1994), conservation (Shultz, 1998) and prediction of effect sizes from the magnitudes of causal potencies and effect resistances (Shultz *et al.* 1995). Mathematical and computational details of the CC algorithm can be found elsewhere (Fahlman & Lebiere, 1990; Shultz, Mareschal & Schmidt, 1994; Shultz *et al.* 1995; Mareschal & Shultz, 1996).

The initial CC network used in our computer simulation had three input units representing speaker, addressee and referent, and one output unit representing the pronoun. All of the input units were connected to the output unit. In addition, there was a bias unit, which was always on, connected to the output unit. The bias is a constant similar to the constant term in a regression analysis. It can also be interpreted as the negative threshold of receiving units. If the value goes beyond the threshold, the neuron is supposed to 'fire'. A sample pronoun network after recruitment of two hidden units is presented in Figure 1.

We used analogue coding for all simulations in order to implement prior knowledge about the kind PERSON on inputs in an implicit way. The child was coded as 0, the mother was coded as $+2$, and the father as -2. Two other persons were coded $+1$ and -1, respectively. Numbers other than these training points indicate other persons who did not appear in the training patterns. The number assigned to each person was on nominal scale (Stevens, 1951). It simply identifies the members with respect to the property in question, although networks do not know what each number indicates (i.e. each person in the present study) and what the property in question is (i.e. the kind PERSON). The networks were expected to learn the nominality of the numbers used for the input variables but learn to ignore orders, sizes of differences and ratios (Takane, 1998). The learning of the nominality of numbers can be achieved by a nonlinear transformation applied to the summed contributions, which forms the activation at the output unit of the CC networks.

This type of analogue coding has several advantages over the distributed binary coding used in our earlier pronoun simulations (Shultz, Buckingham & Oshima-Takane, 1994). First, persons appearing in the training patterns are treated as if they are the same kind. Yet the child is treated differently from other persons without explicitly indicating to the networks that the child is the self. In other words, analogue coding allows us to represent members and the class to which members belong by assigning a number to

the members on the same unit. Thus, it should facilitate generalization to untrained members of the same kind without explicit teaching. With the previous distributed binary coding, on the other hand, the networks treat each person as a distinct individual with no relation to other individuals in the training patterns. As a result, they may learn rules for distinguishing *me* from *you* separately for each individual. Second, with analogue coding, any number of persons could be added to the training patterns without adding input units. Third, the target semantic rules can easily be translated into the target functions that the networks are approximating. Finally, the generalization capability of the networks can be examined using graphing techniques (Takane, Oshima-Takane & Shultz, 1994).

The output unit was coded +0·5 for *me* and −0·5 for *you*. We used a score-threshold value of 0·1 in all of the present simulations. CC networks stop learning when all of the outputs are within score-threshold of their targets on all of the training patterns. Thus, the score-threshold parameter reflects the allowable differences between actual and target output activations. With the score-threshold parameter set to the value of 0·1, the activation level of the output unit needs to be above +0·4 to be interpreted as *me*, and below −0·4 to be interpreted as *you*.

Simulation 1 : pure conditions with three persons

This study investigated whether networks in the pure non-addressee condition could produce correct child-speaking patterns without explicit corrections by having prior knowledge about the basic-level kind PERSON. Only three persons, child, mother and father, were involved in this study. Table 1 summarizes 12 possible ways *me* and *you* occur when there are only

TABLE 1. *Training patterns*

Condition	Input Speaker	Addressee	Referent	Output Pronoun
Phase 1: Parent-speaking patterns				
Pure addressee	father	child	father	me
	father	child	child	you
	mother	child	mother	me
	mother	child	child	you
Pure non-addressee	father	mother	father	me
	father	mother	mother	you
	mother	father	mother	me
	mother	father	father	you
Phase 2: Child-speaking patterns				
	child	father	child	me
	child	father	father	you
	child	mother	child	me
	child	mother	mother	you

three persons. The first four are called pure addressee patterns, in which the addressee is always the child. The next four are called pure non-addressee patterns, in which the child is neither the speaker nor the addressee (i.e. the conversation is between mother and father). The remaining four patterns occur when the child is the speaker, producing the pronouns, and they are called child-speaking patterns.

There are two training phases. In Phase I training, networks learn parent-speaking patterns (four patterns). In Phase II training, child-speaking patterns (four patterns) are added to Phase I training patterns. Phase I training can be seen as the period during which children hear pronouns used by others but have not yet produced any pronouns, and Phase II training as the period during which they not only hear the pronouns but also start producing them. In this simulation there were two conditions: addressee and non-addressee. In phase I training networks in the addressee condition were trained with the four addressee patterns and those in the non-addressee condition were trained with the four non-addressee patterns. There were 20 runs in each condition, with each run starting from a different random set of connection weights.

Learning time was measured in terms of the number of epochs required to learn all of the training patterns. An epoch is a sweep through all of the training patterns. The mean epochs to learn in Phase I were 60·6 (S.D. = 4·1) in the addressee condition and 62·3 (S.D. = 4·7) in the non-addressee condition. There was no significant difference between the two. All networks recruited one hidden unit during Phase I training. The mean epochs to learn Phase II training patterns were 106·8 (S.D. = 9·8) in the addressee condition and 50·6 (S.D. = 32·0) in the non-addressee condition. The networks in the non-addressee condition took significantly fewer epochs to learn than those in the addressee condition, $t(23)$[4] $= 7{\cdot}51$, $p < 0{\cdot}001$.[5] However, unlike human children, none of them could produce the correct child-speaking patterns without some Phase II training, where the networks were explicitly taught the correct child-speaking patterns. For Phase II, one additional hidden unit was recruited by all the networks in the addressee condition and by 9 networks (45 %) in the non-addressee condition. The performance of the networks in the pure non-addressee condition is similar to that of those in the previous simulations with three persons (Shultz, Buckingham & Oshima-Takane, 1994), although learning takes longer here because of a much smaller score-threshold (0·1 vs. 0·4). It appears that having prior knowledge about

[4] A separate variance estimate was used to calculate the t-value because there was a significant difference in variances between the two conditions. The number of degrees of freedom was adjusted accordingly.

[5] Although the 1 % significance level was employed for all statistical tests in the present paper, the probability was reported for each test when $p < 0{\cdot}001$ for the interest of the readers.

the kind PERSON is not enough for pure non-addressee networks to produce the correct child-speaking patterns without error-correcting feedback.

Simulation 2 : pure conditions with five persons

This study was conducted to test whether additional persons appearing in Phase I non-addressee patterns would improve learning and generalization to the child-speaking patterns. For this purpose, two other persons were included in Phase I training patterns besides those used in simulation 1. With five persons (child and four other persons) there are 40 possible ways in which *me* and *you* occur. Eight patterns were child-speaking patterns. The remaining 32 patterns were other-speaking patterns in which the speaker was someone other than the child. Eight out of 32 patterns were pure addressee patterns and the remaining 24 patterns were pure non-addressee patterns. In Phase I training, networks in each condition learned a total of 24 other-speaking patterns. Networks in the addressee condition were given eight addressee patterns three times in each epoch, whereas networks in the non-addressee condition were given 24 non-addressee patterns once per epoch. In Phase II training, the eight child-speaking patterns were added to Phase I training patterns.

The mean epochs to learn Phase I training patterns were 91·0 (S.D. = 8·4) in the addressee condition and 271·4 (S.D. = 17·0) in the non-addressee condition, which shows a significant difference between the two, $t(38) = 27{\cdot}7$, $p < 0{\cdot}001$. One hidden unit was recruited by all networks in both conditions during Phase I training. All networks in the addressee condition needed Phase II training to produce correct child-speaking patterns. The mean epochs to master child-speaking patterns was 169·3 (S.D. = 26·2). One additional hidden unit was recruited by 19 networks and two additional hidden units by one network during Phase II training. On the other hand, none of the networks in the non-addressee condition needed any Phase II training to produce correct child-speaking patterns.[6]

[6] In order to rule out a possibility that difference in input complexity rather than the shifting references may explain this significant condition effect, we conducted an additional simulation where networks in the pure addressee condition received the same number of distinct patterns (24) as those in the pure non-addressee condition by adding 8 more persons to the pure addressee patterns. The results indicated that these pure addressee networks needed mean epochs of 396·5 and 298·8 for learning the Phase I and Phase II training patterns, respectively. Further, consistent with our hypothesis, network analysis revealed that these networks learned the incorrect *me–you* reversal function during Phase I training. These results confirmed that networks trained by non-addressee patterns with 5 persons in simulation 2 learned the correct function, not because the input patterns are three times more varied than those for the pure addressee networks, but because the input contains systematic shifting references. Another piece of evidence is

The non-addressee networks in the 5-person condition showed perfect generalization to the child-speaking patterns without Phase II training. On the other hand, non-addressee networks in the 3-person condition needed some Phase II training to master the child-speaking patterns, although they showed better generalization than addressee networks. The results clearly indicate that addition of two other persons in the Phase I training patterns facilitates correct production of the child-speaking patterns without Phase II training.

Simulation 3 : pure and mixed conditions with five persons

The child's natural language learning environment involves some mixture of addressee and non-addressee materials. Simulation 3 was conducted to simulate the child's natural language learning environment by including two conditions with different mixtures of addressee and non-addressee materials. Although the pure addressee and the pure non-addressee conditions are not a realistic simulation of the child's language environments, they were included to determine to what extent non-addressed speech is needed to learn the nonlinear function underlying the semantic rules of the pronouns.

As in simulation 2, five persons (the child and four other persons) were involved and 40 different training patterns were used. There were four conditions with the frequency multiplies of addressee:non-addressee of 10:0, 9:1, 5:5 and 0:10. The 10:0 and the 0:10 conditions were essentially the same as the pure addressee and the pure non-addressee conditions, respectively, in simulation 2. Two mixed conditions, 9:1 and 5:5, were included to model more realistic language learning environments. We assume that the 9 addressee vs. 1 non-addressee mixed condition is similar to the linguistic environment of firstborn children, whereas the 5:5 mixed condition is similar to the linguistic environment of secondborn children (Shultz, Buckingham & Oshima-Takane, 1994). In Phase I training, networks in each condition learned other-speaking patterns (a total of 240 patterns). In Phase II training, child-speaking patterns (eight patterns) were added to Phase I training patterns. Other-speaking patterns used in Phase I training depended on the conditions. We hypothesized that the more the non-addressee materials appearing in Phase I training patterns, the faster the learning and the better the generalization.

---

from the significant condition effect observed in the three-person study (simulation 1). There both addressee and non-addressee conditions received the same number of different patterns, and the non-addressee condition was clearly superior to the addressee condition. These two pieces of evidence support our hypothesis that observation of shifting reference in overheard speech is crucial for inducing the correct semantic rules. The results cannot be explained by the alternative of input complexity.

The mean epochs required for learning Phase I and Phase II training patterns are given by condition in Table 2. Planned comparisons revealed

TABLE 2. *The mean epochs required for learning Phase I and II training patterns by condition*

	Condition (addressee:non-addressee)			
	10:0 ($n = 20$)	9:1 ($n = 20$)	5:5 ($n = 20$)	0:10 ($n = 20$)
Phase I				
Mean	92·7	372·1	219·7	247·9
S.D.	7·4	42·9	21·7	20·3
Phase II				
Mean	147·4	64·8	0	0
S.D.	13·8	27·9	0	0

that it took more epochs to learn Phase I training patterns in the 9:1 mixed condition than in the 5:5 mixed and the 0:10 pure non-addressee conditions combined, $F(1, 76) = 280{\cdot}12$, $p < 0{\cdot}001$, and fewer epochs to learn in the 10:0 pure addressee condition than in the other three conditions combined, $F(1, 76) = 1699{\cdot}8$,[5] $p < 0{\cdot}001$. This result is consistent with the psychological finding that firstborn children are delayed in pronoun production compared to secondborn children (Oshima-Takane *et al.*, 1996). All 0:10 pure non-addressee and 5:5 mixed networks recruited one hidden unit during Phase I training and none of them needed Phase II training to produce correct child-speaking patterns. On the other hand, most 9:1 mixed networks (90%) recruited two hidden units during Phase I training. They needed some Phase II training, although none of them recruited an additional hidden unit. All 10:0 pure addressee nets recruited one hidden unit during Phase I training and needed Phase II training in which they recruited an additional hidden unit. The 10:0 pure addressee networks needed more epochs than the 9:1 mixed networks to learn Phase II training patterns, $t(28)^4 = 11{\cdot}86$, $p < 0{\cdot}001$.

How much non-addressee material is needed in Phase I to produce correct child-speaking patterns without Phase II training? It is clear from the present results that more than 10% of the training patterns must be non-addressee to produce correct child-speaking patterns without Phase II training. However, there may not be a need for 50% of the training patterns to be non-addressee materials as in the 5:5 mixed condition. To answer this question, we conducted an additional simulation with two new mixed conditions with the frequency multiples of addressee:non-addressee of 8:2 and 7:3. Interestingly, all networks in both conditions produced correct child-speaking patterns without Phase II training. The mean epochs to learn

Phase I training patterns were 232·7 (S.D. = 17·2) in the 8:2 mixed condition and 217·7 (S.D. = 18·8) in the 7:3 condition. There was no significant difference between the two. One hidden unit was recruited by each network in both conditions. Planned comparisons showed that there was no difference in the mean epochs to learn Phase I training patterns between the 8:2 and the 7:3 mixed conditions combined and the 5:5 mixed and the 0:10 pure non-addressee conditions combined. However, networks in the 9:1 mixed condition needed more epochs than those in the other three mixed conditions and the 0:10 pure non-addressee condition combined, $F(1, 114) = 409{\cdot}0$,[7] $p < 0{\cdot}001$. The networks in the 10:0 pure addressee condition needed fewer epochs to learn Phase I training patterns than those in all other conditions combined, $F(1,114) = 1844{\cdot}1$,[7] $p < 0.001$. These results are consistent with the above finding that the 10:0 pure addressee networks needed the fewest epochs to learn Phase I training patterns, whereas the 9:1 mixed networks needed the most. The results indicate that if at least 20% of the training patterns were non-addressee materials, the networks could produce correct child-speaking patterns without Phase II training. An interesting question is whether the knowledge representation and generalization capability of the networks when only 20% of the training patterns are non-addressee materials differ from when they constitute more than 20%. This issue will be examined by analysing networks' representation of function approximations as learning progresses.

NETWORK ANALYSIS

We analysed networks' representation of function approximation and generalization capability by examining how the function approximation was accomplished over time. In particular, we examined whether non-addressee materials are crucial in approximating the target function underlying the correct use of first and second person pronouns.

Target function

The target function is the correct function connecting inputs to outputs that the network has to learn. In the case of *me-you* pronoun learning, this function outputs the value representing *me* (+0·5) when the speaker is the referent and the value representing *you* (−0·5) when the addressee is the referent. We define

$$y = (A-R)/(A-S)-0{\cdot}5,$$

where A is the addressee, R is the referent of the pronoun to be produced,

[7] The F-value was obtained from the log transformations to stabilize variance. The original mean values were reported in the text, however, because a monotonic transformation does not change what is originally measured by the dependent variables and conclusions can be made on the original measures (Ferguson & Takane, 1989, p. 267). The F-value became smaller but the text result did not change with the transformations.

S is the speaker, and y is the output. This correct function with training points under 5-person conditions is depicted in Figure 2. The graphic representation of the correct *me* surface is presented on the top and that of the *you* surface at the bottom. In the case of *me*, the left side horizontal axis of the graph represents the speaker = referent dimension (Sp = Rf) and the right side horizontal axis the addressee dimension (Ad). In the case of *you*, the former represents only the speaker dimension (Sp) and the latter the addressee = referent dimension (Ad = Rf). A number on these dimensions indicates who is the speaker, ranging from $-3{\cdot}5$ to $+3{\cdot}5$, and who is the addressee, ranging from $-3{\cdot}5$ to $+3{\cdot}5$. Note that the number assigned to each person was on a nominal scale and it simply identified a discrete entity (i.e. person). The vertical axis represents pronouns in terms of the output activation. When the speaker and the referent agree and when the addressee and the referent disagree, then the correct pronoun to be produced is *me* and the output activation is $+0{\cdot}5$. When the addressee and the referent agree, on the other hand, the correct pronoun to be produced is *you* and the output

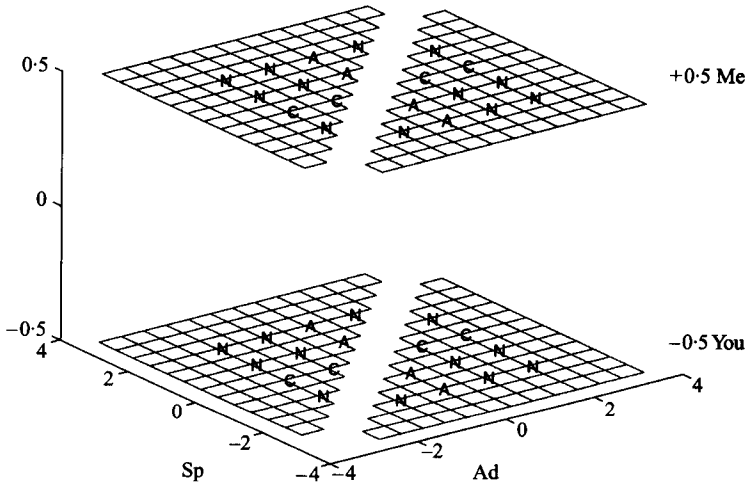

Fig. 2. Graphic representation of the target (correct) function with training points under 5-person conditions. The *me* surface is presented on the top ($+0{\cdot}5$) and the *you* surface at the bottom ($-0{\cdot}5$). A number on the left side horizontal axis of the graph indicates who is the speaker, ranging from $-3{\cdot}5$ to $+3{\cdot}5$ and a number on the right side horizontal axis indicates who is addressee, ranging from $-3{\cdot}5$ to $+3{\cdot}5$. The child is coded as 0, the mother as 2, and the father as -2. Two other people are coded as 1 and -1. The letter A on each surface indicates addressee patterns, the letter N non-addressee patterns, and the letter C the child-speaking patterns. All other points on the grids of the *me* or *you* surface represent speaker-addressee combinations that do not appear in any of the training patterns. For the *me* surface, the referent is the speaker; for the *you* surface, the referent is the addressee. The diagonal points where the speaker and the addressee agree are excluded. Sp = speaker and Ad = addressee.

activation is $-0{\cdot}5$. For instance, when the child is speaking to the mother and referring to himself/herself, the correct pronoun to be produced is *me*. This is depicted on the *me* surface (top) with the zero on the speaker = referent dimension (i.e. the child is the speaker and the referent) and the $+2$ on the addressee dimension (i.e. the mother is the addressee). Similarly, when the mother is talking to the child and referring to the child, the correct pronoun to be produced is *you*. This is depicted on the *you* surface (bottom) with the zero on the addressee = referent dimension (i.e. the child is the addressee and the referent) and the $+2$ on the speaker dimension (i.e. the mother is the speaker). The letter A on each surface represents the addressee patterns, the letter N indicates non-addressee patterns, and the letter C indicates the child-speaking patterns in the 5-person conditions. All other points on the grids of the *me* or *you* surface represent speaker-addressee combinations that do not appear in the training patterns. Figure 3 presents the graphic representation of an incorrect function (bottom) in contrast with that of the correct function (top). The incorrect *me* and *you* surfaces portray reversed errors in which *me* is produced whenever the referent is a person other than the child (i.e. the output activation is $+0{\cdot}5$ regardless of speech roles), whereas *you* is produced whenever the referent is the child (i.e. the output activation is $-0{\cdot}5$).

We investigated approximations of the target function in networks by graphing performance on training and test patterns. Generalization tests included both interpolation within the range of training values and extrapolation beyond the range of training values. An example of interpolation is speaker $= +1{\cdot}5$, addressee $= -1{\cdot}5$, and referent $= \pm 1{\cdot}5$. An example of extrapolation is speaker $= 0$, addressee $= -2{\cdot}5$ and referent $= -2{\cdot}5$. If the network's approximation is close to the target function, we can conclude that the network's generalization capability is quite good.

Function approximations : developmental data

The CC algorithm constructs a network and estimates connection weights based on a sample of training patterns. The sample of training patterns used for Phase I training depends on the conditions. For each input pattern, a unit in a trained network sends contributions to units it is connected to. A contribution is defined as the product of the activation of the sending unit and the connection weight between the sending unit and the receiving unit. The receiving unit forms its activation by summing up the contributions from other units and applying the sigmoid transformation to the summed contribution. An activation is computed at each unit and for each input pattern in the training sample. An activation at the output unit is the network prediction for the output. In the training phase, connection weights are determined so that the network prediction closely approximates the output corresponding to the input pattern. In order to understand how function approximation is done, we examined network performance at various points

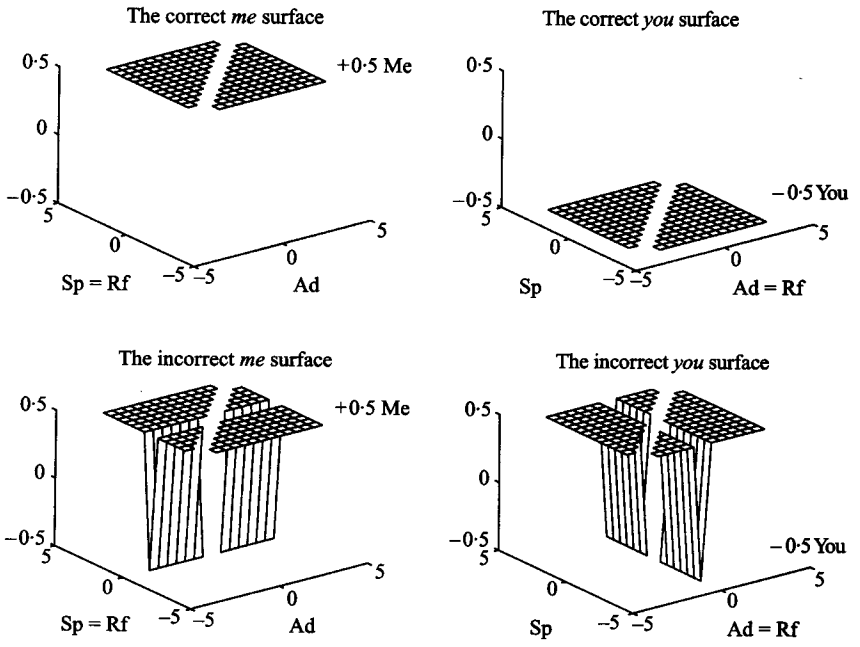

Fig. 3. Graphic representation of the target (correct) and incorrect functions. The correct *me* and *you* surfaces are presented on the top and the incorrect *me* and *you* surfaces at the bottom. Sp = speaker, Ad = addressee and Rf = referent. Sp = Rf indicates that the speaker and the referent agree and Ad = Rf indicates that the addressee and the referent agree.

in learning in each condition. Distinct developmental stages in learning are defined by the topological changes that occur in the network when a hidden unit is added.

Figures 4–6 depict function approximations at different developmental stages, and their changes from one stage to the next for each training phase by one of the networks in each condition. Networks' function approximations in each developmental stage are obtained by deriving network predictions just before a new hidden unit is recruited. Only one figure (Figure 4) is presented for the 8:2, 7:3, 5:5, and 0:10 networks because the topological changes in these networks were essentially the same across these four conditions.

The network's representation of the *me* and *you* surfaces at Stage 1 in Figure 4 shows that the network in the three mixed (8:2, 7:3 and 5:5) and the 0:10 pure non-addressee conditions could not discriminate *me* from *you* at Stage 1, because all points on both surfaces took the value of 0, that is, neither *me* or *you*. After adding the first hidden unit, h_1, it learned to approximate the correct function remarkably well (Stage 2). The network now correctly discriminated the trained non-addressee patterns as well as the

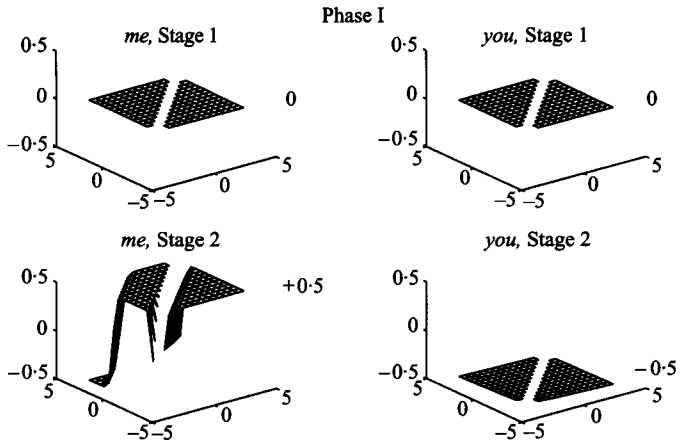

Fig. 4. Development changes in function approximations by one network in the 0:10 pure non-addressee condition. The figures for the 8:2, 7:3, and 5:5 mixed conditions were similar to that for the 0:10 pure non-addressee condition. Both *me* and *you* surfaces have a value of 0 at Stage 1. At Stage 2, most points on the *me* surface have a value of +0·5 and all the points on the *you* surface have a value of −0·5.

untrained addressee and child-speaking patters. That is why the networks in the three mixed and the 0:10 pure non-addressee conditions did not need Phase II training to produce the correct child-speaking patterns. The network's generalization to untrained other-speaking patterns was also very impressive, although extrapolation for the points very far away from the non-addressee patterns was a bit difficult as indicated by the points on the left side corner of the *me* surface taking the value of −0·5 (*you*).

Figures 5–6 indicate that the 10:0 pure addressee and 9:1 mixed networks could not discriminate *me* from *you* at Stage 1. At Stage 2–1, the 10:0 pure addressee network (Figure 5) learned to approximate an incorrect function, producing reversal errors. That is, the network produced *you* (−0·5) when the child is speaking and referring to himself/herself (0 points on the Sp = Rf dimension of the *me* surface), and produced *me* (+0·5) when the child is speaking and referring to others (0 points on the Sp dimension and all points other than 0 on the Ad = Rf dimension of the *you* surface) except the points on the edges. It needed Phase II training to master the correct child-speaking patterns but even after the child-speaking patterns were added to the Phase I training patterns (Stage 2–2), the 10:0 pure addressee network kept producing reversal errors. After the second hidden unit, h_2, was added (Stage 3), the network produced *me* and *you* correctly at the training points (addressee and child-speaking patterns). However, the *you* and *me* surfaces after Phase II training were not as flat as those of the correct function. In particular, the *me* surface at Stage 3 indicates that extrapolation for the points

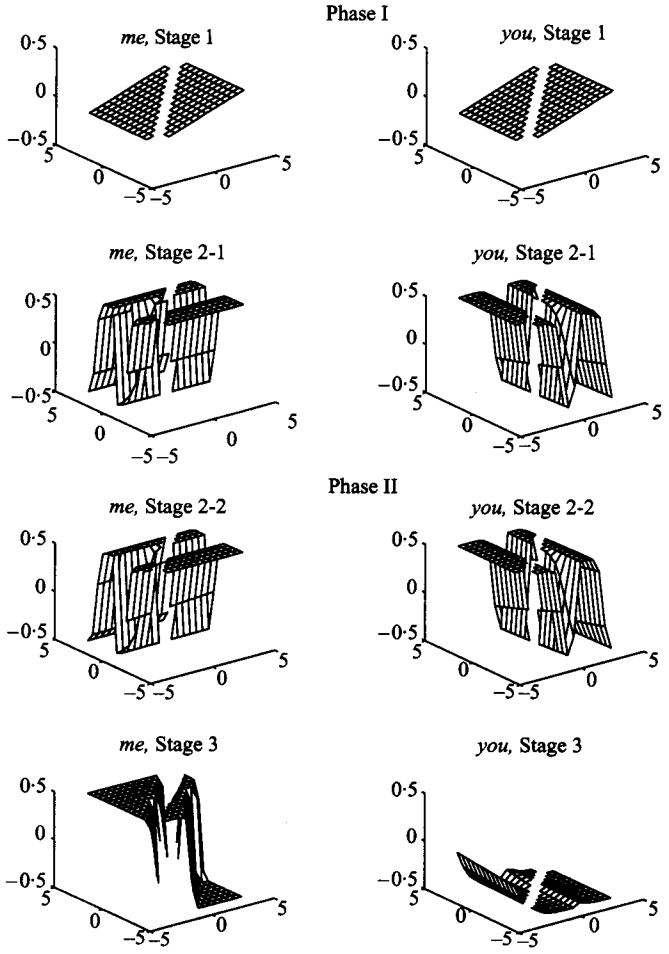

Fig. 5. Developmental changes in function approximations by one network in the 10:0 pure addressee condition. The change in the training set occurred between Stage 2-1 and Stage 2-2.

on the untrained other-speaking patterns far away from the training points (addressee and child-speaking patterns) was very difficult as indicated by the points on the right side corner of the *me* surface taking the value of −0·5 (*you*). This indicates that the generalization capability of the 10:0 pure addressee net was much more limited than the 0:10 pure non-addressee network and the three mixed networks (8:2, 7:3 and 5:5) even after Phase II training.

Similarly, by adding the first hidden unit, h_1 (Stage 2), the 9:1 mixed network (Figure 6) learned to approximate the incorrect function, producing

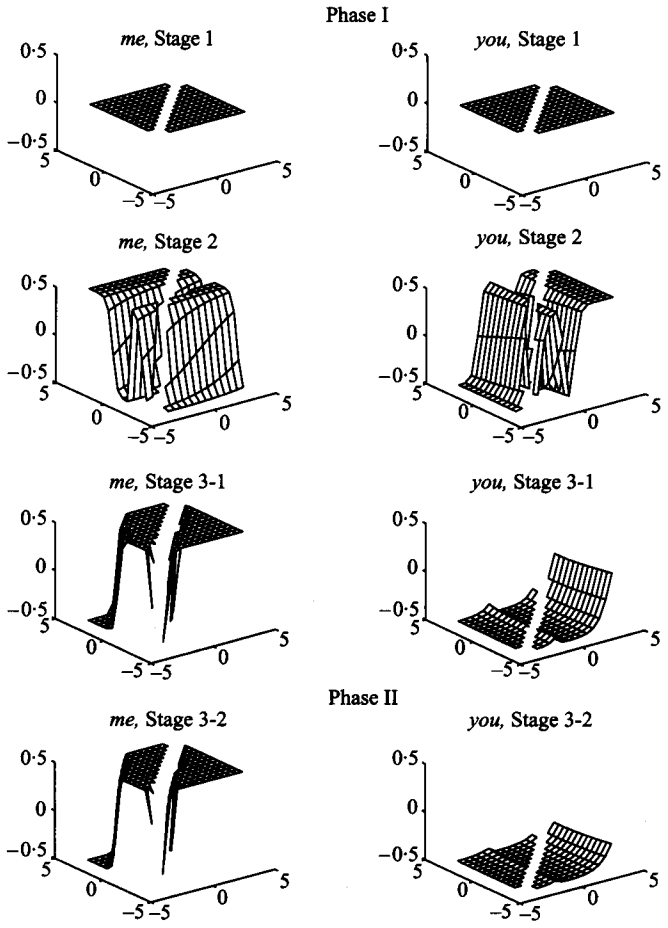

Fig. 6. Developmental changes in function approximations by one network in the 9:1 mixed condition. The change in the training set occurred between Stage 3-1 and Stage 3-2.

reversal errors. However, unlike the 10:0 pure addressee network, this network appeared to unlearn this incorrect function during Phase I training by adding the second hidden unit, h_2 (Stage 3-1). This is why the 9:1 mixed networks needed significantly more epochs than other mixed networks in Phase I training. Although it needed Phase II training, no additional hidden unit was recruited (Stage 3-2). Comparison of the network's approximation of the *me* and *you* surfaces between Stages 3-1 and 3-2 indicates that the network needed Phase II training for adjusting the weights to produce *you* correctly. However, unlike the 0:10 pure non-addressee network and the other three mixed networks, the *you* surface was not as flat as that of the

correct function. This indicates that generalization capability of the 9:1 mixed network was not as good as that of the 0:10 pure non-addressee network and the other three mixed networks.

SIMULATION 4: A REMEDIAL STUDY OF THE PURE ADDRESSEE NETWORKS

Some normal children show persistent reversal errors just like the 10:0 pure addressee networks. For instance, Oshima-Takane (1992) reports a case of a firstborn boy whose pronoun errors persisted for about 10 months. At an earlier stage the boy made consistent errors both in comprehension and production indicating that he learned the incorrect, reversal rules. Oshima-Takane suggested two reasons why his pronoun errors persisted for such a long time. First, once children learn the incorrect reversal rules, it is very difficult to correct them, because they completely misunderstand others' corrections. Suppose that the boy says, 'You want cookie' (meaning 'I want cookie') and his mother says, 'No, you should say, "I want cookie".' The boy's interpretation of what his mother would like him to say would be that Mommy wants cookie and not that he wants cookie. Thus, he would say 'No' or simply ignore his mother's comment. Even though corrections may give children the idea that there is something wrong with their usage of pronouns, these corrections do not seem to tell them the correct semantic rules. Second, it is rather difficult for parents to keep correcting their children's errors all the time. The boy's mother and his babysitter tried to correct the boy's errors, but they also responded to him as if he used correct pronouns most of the time by using the context to figure out what he meant. Consequently, children may not feel any need to change their use, simply because the pronouns work. Oshima-Takane (1992) argues that non-addressed speech plays an important role in unlearning incorrect semantic rules because it provides children with an opportunity to observe how the referent of first and second person pronouns shifts systematically.

In order to test Oshima-Takane's hypothesis, we conducted a remedial study by examining whether the addition of the non-addressee materials to the 10:0 pure addressee materials after Phase I training would help the 10:0 pure addressee networks to unlearn the incorrect function and eventually learn the correct function. One network trained by the 10:0 pure addressee condition in Phase I training was trained by three different mixed conditions in Phase II with the frequency multiples of addressee:non-addressee of 9:1, 5:5, and 1:9. In Phase III training all networks were trained with the child-speaking patterns added to the Phase II training patterns. There were 20 runs for each condition.

The mean epochs required for learning phase II and Phase III training patterns by condition are given in Table 3. All networks in the 9:1 and the

TABLE 3. *The mean epochs required for learning for Phase II other-speaking mixed patterns and Phase III child-speaking patterns*

	Condition (addressee:non-addressee)		
	9:1 ($n = 20$)	5:5 ($n = 20$)	1:9 ($n = 20$)
Phase II			
Mean	306·2	306·4	202·9
S.D.	65·3	79·7	168·9
Phase III			
Mean	87·2	82·3	20·5
S.D.	36·8	27·9	17·9

5:5 mixed conditions needed Phase III training to produce correct child-speaking patterns, whereas six out of 20 networks in the 1:9 mixed condition did not need Phase III training. Planned comparisons revealed that there were no significant differences between the 9:1 and the 5:5 conditions in the mean epochs to learn Phase II and Phase III training patterns, although it took fewer epochs to learn both training patterns in the 1:9 condition than in the other two conditions combined, $F(1, 57) = 36{\cdot}2$,[8] $p < 0{\cdot}001$ for Phase II training and $F(1, 57) = 71{\cdot}2$, $p < 0{\cdot}001$ for Phase III training. Fifteen

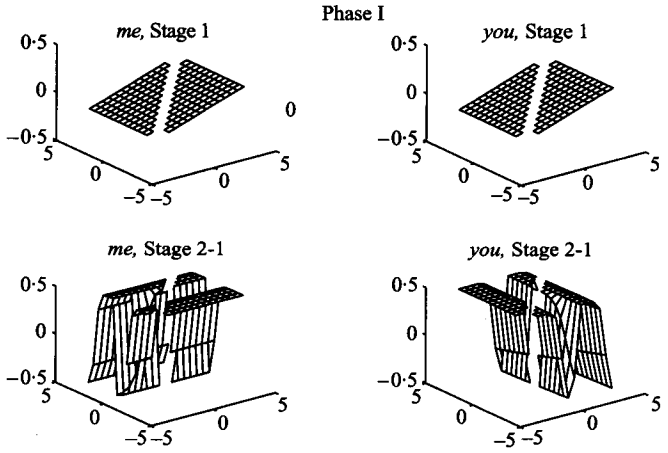

Fig. 7. Function approximation by one 10:0 pure addressee network during Phase I training.

[8] The F-value was obtained from log transformations to stabilize variance. The F-value became larger and the probability became smaller (the F-value without transformation was 10·97, $p = 0{\cdot}002$), but the test result did not change with the transformations ($p < 0{\cdot}01$).

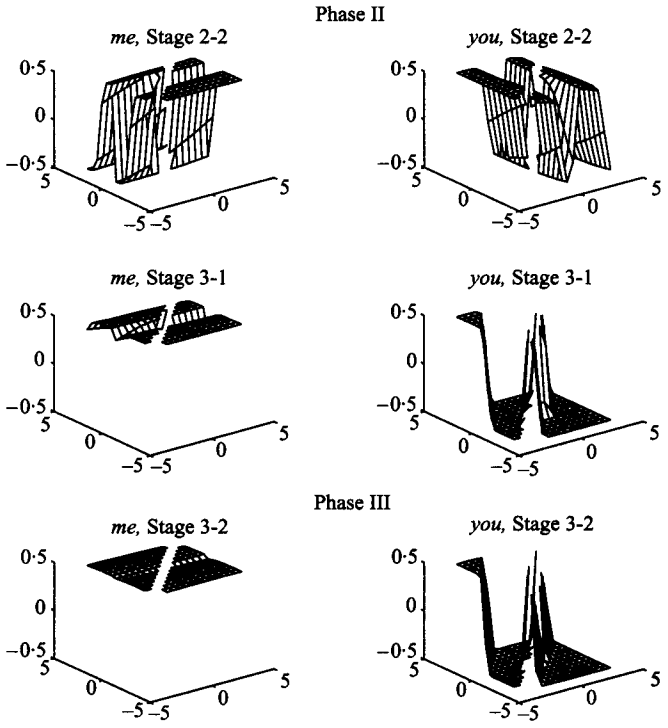

Fig. 8. Developmental changes in function approximations by one 10:0 pure addressee network in the 9:1 mixed condition. The changes in the training set occurred between phases.

networks in the 9:1 condition and 14 networks in the 5:5 condition recruited one hidden unit and the remaining networks recruited two during Phase II training. None of the networks in the 9:1 and the 5:5 mixed conditions recruited additional hidden units during Phase III training. On the other hand, all the networks in the 1:9 condition recruited one hidden unit during Phase II training and only one recruited an additional hidden unit during Phase III training. The results indicate that the more the non-addressee materials appearing in Phase II training patterns, the faster the learning of the correct function.

In order to understand how networks in each condition unlearn the incorrect function and how they eventually learn the correct function, we analysed one network in each condition using the same graphing technique used in section 3. Figure 7 presents function approximations of the 10:0 pure addressee network during Phase I training. All networks in the remedial study started with this 10:0 pure addressee network's representation of the incorrect function at Stage 2-1.

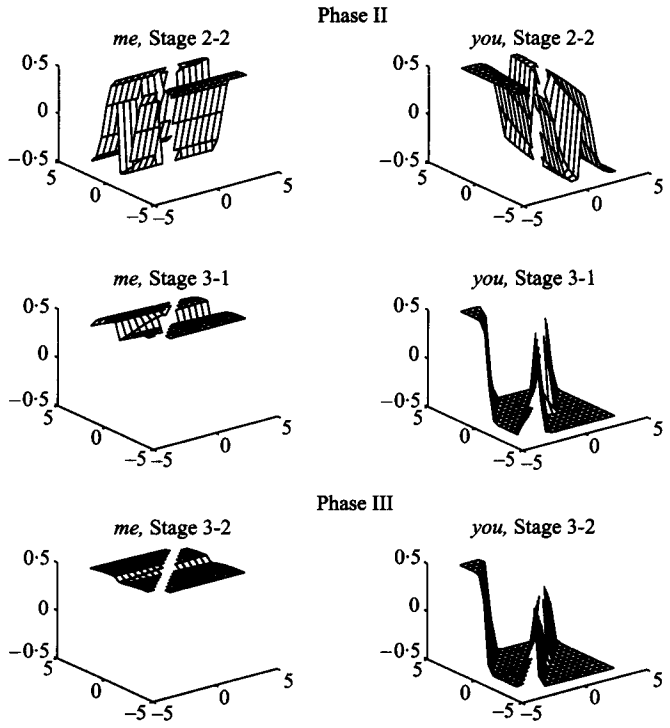

Fig. 9. Developmental changes in function approximations by one 10:0 pure addressee network in the 5:5 mixed condition. The changes in the training set occurred between phases.

Developmental changes from one stage to the next during Phase II and III training by one network in each of the 9:1, 5:5 and 1:9 conditions are presented in Figures 8–10, respectively. Comparisons of the representations of *me* and *you* surfaces by the three mixed networks at Stage 2-2 (Figures 8–10) indicate that the more non-addressee materials appearing in Phase II training patterns, the flatter the *me* and *you* surfaces, suggesting that the non-addressee materials facilitate unlearning of the incorrect function. Although the 9:1 (Figure 8) and the 5:5 (Figure 9) networks still kept producing reversal errors at Stage 2-2, the 1:9 network (Figure 10) produced few reversal errors. However, it could not discriminate *me* from *you* because none of the output activations for *me* or *you* were within the score-threshold; they were all between $-0{\cdot}4$ and $+0{\cdot}4$. Such indefinite responses could be the network equivalent of a child's not being sure about which pronoun to produce. After adding the second hidden unit, h_2 (Stage 3), the networks in all three conditions learned to discriminate *me* from *you*. Although most networks needed some Phase III training, no additional hidden unit was

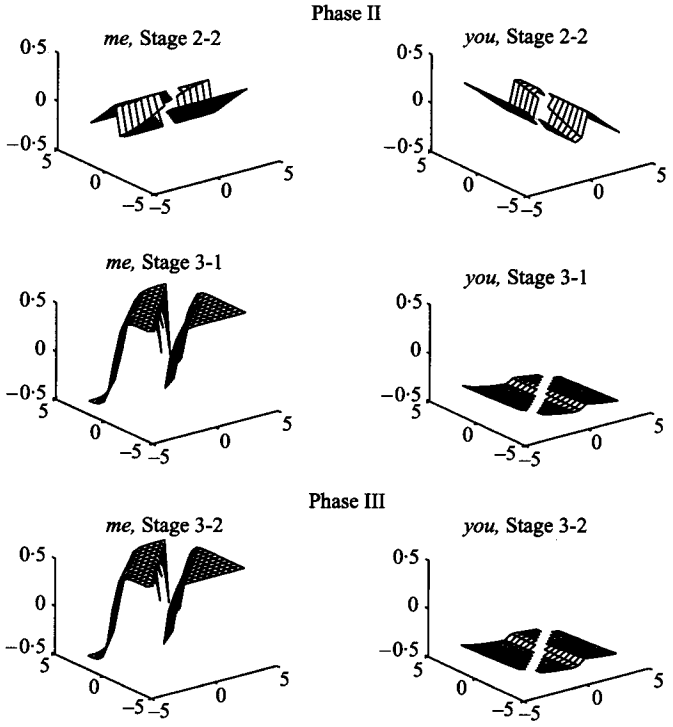

Fig. 10. Developmental changes in function approximations by one 10:0 pure addressee network in the 1:9 mixed condition. The changes in the training set occurred between phases.

recruited. Furthermore, the *me* and *you* surfaces of these networks after Phase III training are similar to those of the 9:1 mixed network rather than those of the 10:0 pure addressee network after Phase II training in simulation 3. The result that the networks in the three mixed conditions (9:1, 5:5, 1:9) in the present study needed fewer epochs to learn child-speaking patterns than those in the 10:0 pure addressee condition in simulation 3 indicates that non-addressee materials are effective for unlearning the incorrect function.

In the present simulation, six out of 20 networks in the 1:9 condition (90% non-addressee materials) could produce correct child-speaking patterns without errors. In simulation 3, all the networks in the 8:2 condition (20% non-addressee materials) could produce correct child-speaking patterns without errors. This suggests that once networks learn an incorrect function solely from addressed speech, they need substantial non-addressee material to unlearn it.

DISCUSSION

The most impressive finding is that the CC networks could produce the correct pronouns without errors if they hear pronouns used by a variety of speakers in non-addressed speech. Although pure non-addressee networks in the 3-person condition showed better generalization than addressee networks, they needed some Phase II training to master the child-speaking patterns. On the other hand, pure non-addressee networks in the 5-person condition showed perfect generalization to child-speaking patterns without any Phase II training. Furthermore, the pure and the mixed conditions with 5 persons indicate that networks in the 8:2 condition could show perfect generalization to child-speaking patterns without any Phase II training, just like the 0:10 pure non-addressee networks. On the other hand, networks in the 9:1 condition needed some Phase II training to master the child-speaking patterns, just like the 10:0 pure addressee networks. Network analysis revealed that without non-addressed speech, networks would learn an incorrect function and make reversal errors. The 9:1 mixed networks showed better generalization than the 10:0 pure addressee networks after Phase I training, but they needed Phase II training to master child-speaking patterns. Furthermore, generalization to untrained, other person-speaking patterns was not as good as in 0:10 pure non-addressee networks and the other three mixed networks (8:2, 7:3, 5:5).

The results are consistent with the hypothesis that non-addressed speech is a necessary ingredient for the learning of first and second person pronouns. Previous psychological studies done by Oshima-Takane and her collaborators (Oshima-Takane & Oram, 1991; Oshima-Takane *et al.*, 1996) have shown that secondborn children acquire first and second person pronouns faster than firstborn children and they make few production errors. Furthermore, they typically do not show systematic errors in comprehension (e.g. *you* refers to the child even when the child is not addressed) before they produce the pronouns. First-born children typically show systematic errors in comprehension at earlier stages, indicating that they initially learn incorrect, reversal rules. Some firstborns correct these errors by the time they begin to produce pronouns and do not show errors in production. Others make a few but inconsistent reversal errors in production. Thus, performances of the networks in mixed conditions other than the 9:1 match the typical developmental stages of secondborn children. On the other hand, performances of the networks in the 9:1 condition match the typical developmental stages of firstborn children.

Developmental stages in the 10:0 pure addressee networks are similar to those of the firstborn boy reported in Oshima-Takane's (1992) study who made persistent reversal errors. At an early stage, this boy made consistent errors in both comprehension and production, indicating that he learned the

incorrect, reversal rules. Just like the 10:0 pure addressee networks, his pronoun errors persisted for a long period of time despite the fact that his mother and babysitter often tried to correct his errors. The present remedial study showed that opportunities to hear shifting reference of pronouns in non-addressed speech facilitate the unlearning of the incorrect function. However, if the incorrect reversal function is learned solely from addressed speech, substantially more such opportunities are needed. Unlike the present simulations, children do not receive corrections consistently when they produce incorrect pronouns (Oshima-Takane, 1992). In addition, children who learned the incorrect, reversal rules misunderstand the error-correcting feedback completely. Therefore, opportunities to observe pronouns in non-addressed speech must play a more important role in discovering the relationship between pronouns and speech roles in actual language learning than in the simulations. Oshima-Takane (1992) speculated that the boy making persistent errors might have first noticed that there was something wrong with his pronoun usage through the parents' corrections or through misunderstandings of his utterances by others. Then perhaps he began observing other people's usage and inspecting the speech roles of the person that a pronoun designated.

A pilot simulation with distributed binary coding indicates that non-addressee networks lacking knowledge of the kind PERSON under a 5-person condition show no improvement over those under a 3-person condition in their generalization to child-speaking patterns. Unlike analogue coding in the present study, direct comparisons between 5-person and 3-person conditions could not be made with distributed binary coding, because additional individuals could not be included in the training patterns without changing the network topology (i.e. two input units are necessary for coding each person under the 3-person condition, whereas three input units are necessary for the 5-person condition). Nonetheless, close examination of the networks' performances in the pilot simulation suggests that binary-coded networks cannot generalize to child-speaking patterns without error-correcting feedback and that having more speakers in the training patterns does not change this. It appeared that, without information that individuals appearing in the training and test patterns are a member of the same kind, networks learn rules for distinguishing first person pronouns from second person pronouns separately for each individual appearing in the training patterns and, thus, cannot generalize to child-speaking patterns without error-correcting feedback. With analogue coding, the networks are able to represent the kind (type) and its members (tokens) and this seems to be a key of their success in extending generalizations to untrained members of the same kind. Analogue coding appears to have advantages over distributed binary coding for capturing this fundamental characteristic of human cognition.

Several investigators have suggested that an ability to take speaker's point of view is necessary for the child to understand the shifting reference of first and second person pronouns (de Villiers & de Villiers, 1974; Fraiberg, 1977; Loveland, 1984). For instance, Loveland (1984) investigated the acquisition of personal pronouns in relation to the comprehension of spatial points of view on the assumption that understanding spatial viewpoints is a cognitive prerequisite to understanding speaker's point of view, which, in turn, is a prerequisite to the correct use of personal pronouns. Her data indicated that only children with full understanding of spatial viewpoints were able to correctly use all the forms of first and second person pronouns tested. However, evidence for the causal link between understanding of spatial viewpoints and acquisition of these pronouns is not conclusive, because the level of language development of the children in her study was not controlled. Subsequent studies done by other researchers (Issler, 1993; Girouard, Ricard & Decarie, 1995) have provided more complicated results, suggesting that spatial viewpoints and pronouns are not directly related, but develop simultaneously. Furthermore, no studies have directly examined Loveland's assumption that understanding spatial viewpoints is a cognitive prerequisite to understanding speaker's point of view. There is some evidence that at about 9 months of age children begin to understand another's perspective (Baron-Cohen & Ring, 1994; Tomasello, 1995). This would presumably be long before children understand spatial points of view.

We believe that children need to understand another's point of view or intention in order to correctly identify the referent of any word used by them. The major problem with first and second person pronouns is not only that the referent of the pronouns shifts with the speaker but also that the model for correct usage is not provided in speech addressed to the child. Thus, even though children could identify the person referred to by a pronoun used by others correctly, they may be unable to produce the pronoun correctly. Understanding the speaker's intention is not sufficient for learning the semantic rules underlying the correct use of these pronouns. Observing shifting reference of the pronouns in non-addressed speech is necessary.

Much previous language acquisition research has focused on mothers' speech addressed to the child as the primary linguistic input for language acquisition, and little theoretical attention has been paid to non-addressed speech (Oshima-Takane *et al.*, 1996). However, the findings of the present simulations as well as those from Oshima-Takane's psychological studies demonstrate that non-addressed speech is also an important resource for early language development. The present research also suggests that hearing many examples involving various referents would facilitate word learning and generalization. Future research should examine this hypothesis with children.

The present simulations are still incomplete because they do not contain

third person references. In actual language learning situations, children not only hear first and second person pronouns but they also hear third person pronouns referring to a person who is neither speaker nor addressee. A new simulation containing all three kinds of pronouns is now underway to simulate children's pronoun learning situations more realistically. Addition of third person references is particularly important because we could rule out a possibility that networks simply learn a partially correct function (i.e. partially correct with regard to the correct semantic rules) to produce the correct first and second person pronouns. That is, if the speaker is the referent, *me* should be produced; otherwise *you* should be produced. Or if the addressee is the referent, *you* should be produced; otherwise *me* should be produced.

A subsequent network analysis of *me–you* two pronoun learning (Takane, 1998) proved that this was the case. It appears that, for CC networks, speaker = referent is equivalent to addressee ≠ referent, and addressee = referent is equivalent to speaker ≠ referent when only *me* and *you* are to be distinguished. Because no training patterns were given for speaker ≠ referent and addressee ≠ referent, it is quite natural that the networks showed errors for these patterns. This implies, however, that pronouns such as *he* and *she* have to be included in the training patterns in order to learn to discriminate between speaker = referent, on one hand, and speaker ≠ referent and addressee ≠ referent, on the other.

Previous psychological research may suggest that children, too, learn a partially correct function before learning the full correct function. For instance, Brener (1983) reported that children show comprehension errors for second person pronouns indicating that they understand second person pronouns as referring to both addressee and non-addressee, even after they understand that first person pronouns refer only to the speaker. Furthermore, they show similar errors for third person pronouns: they interpret third person pronouns as referring to both addressee and non-addressee before they understand that third person pronouns refer only to non-addressee. Charney (1980) also reported that some children showed comprehension errors suggesting that third person pronouns refer to addressee. However, empirical evidence for such partially correct function is still inconclusive due to the cross-sectional designs used in these studies and the fact that there is very little research investigating the full developmental process of all three personal pronouns.

Although Oshima-Takane (1992) reports a case of consistent reversal errors across different pronoun cases, not all children making persistent reversal errors display such consistency. For instance, the boy in Chiat's study (1982) showed a significantly higher reversal error rate for the first person possessive forms than for its non-possessive forms. Oshima-Takane, Cole & Yaremko (1993) have reported that the hearing-impaired child in

their study correctly produced *I* in self-reference, while producing *me/my* incorrectly in reference with her mother or her mother as possessor. The networks producing reversal errors in the present simulation did not show any variability in error patterns because the study did not deal with the pronoun case distinction. Future research should construct a model which could account for such discrepancies in error rate among different pronoun cases observed in the course of acquisition.

An important difference between networks and children is that networks typically concentrate on a single task or on a restricted set of related tasks, and have little or no prior knowledge to draw upon. Children must deal with many unrelated tasks, but have considerable prior knowledge that might influence new learning. One definite advantage of computer simulation is, however, that we can test the hypothesized developmental mechanisms by setting up ideal environmental conditions that are impossible with humans for ethical reasons. Furthermore, we can specify prior knowledge the networks must have before the learning starts and test the effects of this prior knowledge. We can also directly examine internal representations of networks and their changes, which are again impossible with humans. The present work demonstrates that network analysis is important to understand what and how the networks learn. In particular, analysis of the different knowledge representations over time allows a close comparison between the behaviours of the networks and those of children. This is essential for determining whether the networks have arrived at the same degree of mastery as children and whether their development is similar to that of children.

Future study will examine the role of each unit in neural networks by conducting lesioning studies. We will eliminate connections in a network systematically and will examine how function approximation deteriorates. Elimination of a set of connections may entail elimination of direct or indirect effects of the unit. In this way we can isolate the total, direct, and indirect effects of a unit in function approximations, which would help determine whether the developmental changes observed in children can reasonably be approximated by the addition of hidden units in networks.

In sum, the present modelling study provided evidence in support of Oshima-Takane's theoretical analysis on the learning of English personal pronouns (Oshima-Takane, 1985, 1988, 1992; Oshima-Takane *et al.*, 1996). Children learn correct semantic rules for first and second person pronouns by observing the shifting reference of these pronouns in non-addressed speech, whereas they learn incorrect semantic rules if they simply observe the pronouns in addressed speech. In addition to speech role and referent information, the present study suggests that prior knowledge of the kind PERSON and exposure to examples involving various persons are important factors for improving networks' generalization capability.

REFERENCES

Baron-Cohen, S. & Ring, H. (1994). A model of the mind reading system: neuropsychological and neurobiological perspectives. In C. Lewis & P. Mitchell (eds), *Children's early understanding of mind: origins and development*. Hillsdale, NJ: Erlbaum.

Brener, R. (1983). Learning the deictic meaning of third person pronouns. *Journal of Psycholinguistic Research* **16**, 330–52.

Buckingham, D. & Shultz, T. R. (1994). A connectionist model of the development of velocity, time, and distance concepts. Proceedings of the sixteenth annual conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Charney, R. (1980). Speech roles and the development of personal pronouns. *Journal of Child Language* **7**, 509–28.

Chiat, S. (1982). If I were you and you were me: the analysis of pronouns in a pronoun-reversing child. *Journal of Child Language* **9**, 359–79.

Chiat, S. (1986). Personal pronouns. In P. Fletcher & M. Garman (eds), *Language acquisition: studies in first language development*. (2nd edition). Cambridge: C.U.P.

Clark, E. V. (1978). From gesture to word: on the natural acquisition. In J. S. Bruner & A. Garton (eds), *Human growth and development: Wolfson College Lectures*. Oxford: Oxford University Press.

Cole, E., Oshima-Takane, Y. & Yaremko, R. (1994). Case studies of pronoun development in two hearing-impaired children: normal, delayed, or deviant? *European Journal of Disorders of Communication* **29**, 113–29.

de Villiers, P. A. & de Villiers, J. G. (1974). On this, that, and the other: non-egocentrism in very young children. *Journal of Experimental Child Psychology* **18**, 438–47.

Fahlman, S. E. & Lebiere, C. (1990). The cascade correlation learning architecture. In D. S. Touretzky (ed.), *Advances in neural information processing systems* **2**. San Mateo: Morgan Kaufmann.

Ferguson, G. A. & Takane, Y. (1989). *Statistical analysis in psychology and education, sixth edition*. New York: McGraw-Hill.

Fraiberg, S. (1977). *Insights from the blind: comparative studies of blind and sighted infants*. New York: Basic Books.

Girouard, P., Richard, M. & Decarie, T. G. (1995). Perspective-taking skills and the acquisition of pronouns. Paper presented at *the biennial meeting of the Society for Research in Child Development*, Indianapolis, April.

Issler, D. (1993). *Comprehension of spatial points of view and acquisition of personal pronouns in Brazilian Portuguese*. Unpublished Master's thesis submitted to Pontifical Catholic University of Rio Grande do Sul, Brazil.

Kaplan, D. (1978). On the logic of demonstratives. *Journal of Philosophical Logic* **8**, 81–98.

Loveland, K. A. (1984). Learning about points of view: spatial perspective and the acquisition of 'I/you'. *Journal of Child Language* **11**, 535–56.

Macnamara, J. (1982). *Names for things: a study of human learning*. Cambridge, MA.: Bradford/MIT Press.

Macnamara, J. (1986). *Border dispute: the place for logic in psychology*. Cambridge, MA.: Bradford/MIT Press.

Macnamara, J. & Reyes, G. (1994). Foundational issues in the learning of proper names, count nouns and mass nouns. In J. Macnamara & G. E. Reyes (eds), *The logical foundations of cognition*. New York: O.U.P.

Mandler, J. M., Bauer, P. J. & McDonough, L. (1991). Separating the sheep from the goats: differentiating global categories. *Cognitive Psychology* **23**, 263–98.

Mareschal, D. & Shultz, T. R. (1993). A connectionist model of the development of seriation. Proceedings of the fifteenth annual conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Mareschal, D. & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development* **11**, 571–603.

Oshima-Takene, Y. (1985). *Learning of pronouns.* Unpublished Ph.D. thesis submitted to McGill University.

Oshima-Takane, Y. (1988). Children learn from speech not addressed to them: the case of personal pronouns. *Journal of Child Language* **15**, 94–108.

Oshima-Takane, Y. (1992). Analysis of pronominal errors: a case study. *Journal of Child Language* **19**, 111–31.

Oshima-Takane, Y. & Benaroya, S. (1989). An alternative view of pronominal errors in autistic children. *Journal of Autism and Developmental Disorders* **19**, 73–89.

Oshima-Takane, Y., Cole, E. & Yaremko, R. (1993). Semantic pronominal confusion in a hearing-impaired child: a case study. *First Language* **13**, 149–68.

Oshima-Takane, Y., Goodz, E. & Derevensky, J. L. (1996). Birth order effects on early language development: do secondborn children learn from overheard speech? *Child Development* **67**, 621–34.

Oshima-Takane, Y. & Oram, J. (1991). Acquisition of personal pronouns: what do comprehension data tell us? Poster presented at the biennial meeting of the International Society for the Study of Behavioural Development, Minneapolis, July.

Schiff-Myers, N. (1983). From pronoun reversals to correct pronoun usage: a case study of a normally developing child. *Journal of Speech and Hearing Disorders* **48**, 385–94.

Shipley, E. F. & Shipley, T. E. (1969). Quaker children's use of thee: a relational analysis. *Journal of Verbal Learning and Verbal Behaviour* **8**, 112–17.

Shultz, T. R. (1998). A computational analysis of conservation. *Psychological Science*, in press.

Shultz, T. R., Buckingham, D. & Oshima-Takane, Y. (1994). A connectionist model of the learning of personal pronouns in English. In S. J. Hanson, T. Petsche, M. Kearns & R. L. Rivest (eds), *Computational learning theory and natural learning systems, Vol. 2 : Intersection between theory and experiment.* Cambridge, MA: MIT Press.

Shultz, T. R., Mareschal, D. & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning* **16**, 57–86.

Shultz, T. R., Schmidt, W. C., Buckingham, D. & Mareschal, D. (1995). Modeling cognitive development with a generative connectionist algorithm. In T. J. Simon & G. S. Halford (eds), *Developing cognitive competence : new approaches to process modeling.* Hillsdale, NJ: Erlbaum.

Sorrentino, C. M. (1999). *Individuation, identity and proper names in cognitive development.* Unpublished Ph.D. thesis submitted to Massachusetts Institute of Technology.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (ed.), *Handbook of experimental psychology.* New York: Wiley.

Takane, Y. (1998). Nonlinear multivariate analysis by neural network models. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Takane & Y. Baba (eds), *Data science, classification, and related methods.* Tokyo: Springer.

Takane, Y., Oshima-Takane, Y. & Shultz, T. R. (1994). Approximations of nonlinear functions by feed-forward neural networks. In N. Ohsumi (ed.), Proceedings of the annual meeting of the Japan Classification Society. Tokyo: Japan Classification Society.

Tomasello, M. (1995). On the interpersonal origins of self-concept. In U. Neisser (ed.), *The Perceived Self.* New York: C.U.P.

Xu, F. & Carey, S. (1996). Infant metaphysics: the case of numerical identity. *Cognitive Psychology* **30**, 111–53.