J. Child Lang. **30** (2003), 305–331. © 2003 Cambridge University Press
DOI: 10.1017/S0305000903005622 Printed in the United Kingdom

A connectionist account of Spanish determiner production*

PAMELA SMITH
Department of Psychology, University of Hertfordshire, U.K.

ANDREW NIX, NEIL DAVEY
Department of Computer Science, University of Hertfordshire, U.K.

SUSANA LÓPEZ ORNAT
Facultad de Psicología, Universidad Complutense de Madrid, Spain

AND

DAVID MESSER
Division of Psychology, South Bank University, London, U.K.

(*Received* 10 *December* 1997. *Revised* 23 *October* 2002)

ABSTRACT

Evidence from experimental studies of Spanish children's production of determiners reveals that they pay more attention to phonological cues present in nouns than to natural semantics when assigning gender to determiners (Pérez-Pereira, 1991). This experimental work also demonstrated that Spanish children are more likely to produce the correct determiner when given a noun with phonological cues which suggest it is masculine, and more likely to assign masculine gender to nouns with ambiguous cues. In this paper, we investigate the phonological cues available to children and seek to explore the possibility that differential frequency in the linguistic input explains the priority given to masculine forms when children are faced with ambiguous novel items. A connectionist model of determiner production was incrementally trained on a lexicon of determiner–noun phrases taken from parental speech in

[*] Andrew Nix was funded by a University of Hertfordshire Research Studentship. The work was also aided by a British Council travel grant as part of the British/Spanish Joint Research Programme (Acciones Integradas) 1995/96. The ordering of the second to fourth authors is completely arbitrary and not indicative of amount of input involved in the preparation of this manuscript. Address for correspondence: Professor David Messer, Division of Psychology, South Bank University, Southwark Campus, 103 Borough Road, London SE1 0AA, UK. e-mail: messerdj@sbu.ac.uk

a longitudinal study of a child between the ages of 1;7 and 2;11 (López Ornat, Fernandez, Gallo & Mariscal, 1994) preserving the type and token frequency information. An analysis of the database of parental productions revealed that while regular feminine nouns were slightly more frequent than regular masculine nouns, irregular masculine nouns outnumbered irregular feminine nouns by roughly 2 to 1. On the basis of this, we made the prediction that as the training lexicon builds up, the network will perform better overall on masculine determiners than would be predicted from their forms alone and will tend to assign masculine gender to ambiguous novel nouns in a test set. The findings indicate that, at least in the case of Spanish gender agreement for determiners and nouns, a general associative learning mechanism can account for important characteristics of the acquisition process seen in children.

INTRODUCTION

Over the past fifteen years, several groups of researchers in cognitive psycholinguistics have been using the connectionist paradigm to investigate the processes and representations which govern the acquisition and development of language in young children (Rumelhart & McClelland, 1986; MacWhinney, Leinbach, Taraban & McDonald, 1989; Elman, 1993; Plunkett & Marchman, 1993). The appeal of connectionist systems has been their ability to produce developmental patterns of behaviour similar to those observed in young children, without the need for an explicitly programmed set of rules and parameters. Their ability to produce complex, dynamic patterns of behaviour using a simple mechanism and learning rule has led to a major reassessment of whether humans come innately pre-programmed with natural language or whether language is an emergent property of the dynamic interaction between the developing cortex and the linguistic environment (Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996).

The argument that general learning mechanisms can account for the acquisition of language requires supporting evidence from all aspects of language. Our work aims to show how a connectionist model, with a developmental dimension, supports this claim with respect to the acquisition of gender agreement in Spanish determiner-noun phrases.[1] Determiners are obligatory in most contexts in Spanish and must agree with the gender of the noun. Connectionist models can be vulnerable to the criticism that the data given to the model, and the training regime implemented, are far removed from real life phenomena and hence are not valid tests of theories of language

[1] We have restricted our discussion to monolingual children acquiring Castilian Spanish in accordance with the views expressed by López Ornat, on the dangers of conflating phenomena presented by monolingual and bilingual children learning a variety of Spanish tongues (López Ornat, 1988).

acquisition. We address these problems in two ways: the data given to the model is provided by real language spoken in the hearing of a real child and the training regime mirrors the incremental expansion of that language input over an eighteen month period.

Nominal gender systems have been classified by Corbett & Fraser (2000) into three main groups. In the first group are those languages where the gender of a noun is wholly predictable from semantic assignment without reference to form, for example, Godoberi, a minority language in Daghestan, which has three genders: male rational, female rational and other. Another example is Zande, a language spoken in central Africa, which has four genders: male human, female human, other animate, the rest (with some overlap between the latter two classes). In the second group of languages are those where gender is predictable from natural sexual gender and phonology, e.g. Qafar, spoken in Ethiopia, with two genders where final stressed vowels denote female sex and feminine nouns and all other endings denote male sex and masculine nouns. In the third group are languages such as Russian (and German although this is not an example given by Corbett & Fraser) which require, for all nouns outside natural sexual gender, a knowledge of the inflectional behaviour of a noun in order to identify gender. Spanish, not mentioned by Corbett & Fraser, lies between the transparency of the first two classes and the complexity of the third: semantic assignment is limited to some sexually differentiable referents and phonology only provides limited cues.

Thus gender agreement in Spanish is interesting because gender cannot be easily and reliably recovered from semantics, morphology or phonology although each plays a part. Indeed this lack of consistency in gender assignment to Spanish nouns has led Ambadiang (1999, p. 4874) to call for data on acquisition in the hope of identifying the cues which result in adult competence.

The assignment of gender to Spanish nouns has little connection with natural or semantic gender except in the case of some animate objects (Karmiloff-Smith, 1979; Pérez-Pereira, 1991). For these nouns the same stem is given an *-o* ending for the male and an *-a* ending for the female. This holds true for many of the people a very young child is likely to encounter, e.g. *hermano/hermana* 'brother/sister', *niño/niña* 'child', *chico/chica* 'boy/girl', *primo/prima* 'cousin', *tío/tía* 'uncle/aunt', *abuelo/abuela* 'grandfather/grandmother', *enfermero/enfermera* 'nurse'. It also holds for a few, frequently encountered, animals, e.g. *perro/perra* 'dog/bitch', *gato/gata* 'cat'. At first sight this looks as if it would be possible for children to identify a morphological rule during acquisition, with exceptions such as *padre/madre* 'father/mother', *papá/mamá* 'daddy/mummy', *nene/nena* 'toddler', *bebé* 'baby' (m&f), *dentista* 'dentist' (m&f) and *medico* 'doctor' (m&f) being learned by rote and stored as items in the lexicon in line with predictions from the dual-route model of the acqusition of morphology (Pinker & Prince, 1994).

TABLE 1. *The Spanish determiners used in the present study*

Determiner	Feminine		Masculine	
	Singular	Plural	Singular	Plural
Definite	la	las	el	los
Indefinite	una	unas	un	unos
Demonstrative 1	esa	esas	ese	esos
Demonstrative 2	esta	estas	este	estos
Comparative	otra	otras	otro	otros

But gender suffixes do not even extend to most animals. They have either two independent words, e.g. *caballo/yegna* 'horse/mare' or an invariant form, e.g. *girafa* (f) 'giraffe', *gorilla* (*m*) with sex identified by an adjective, e.g. *la girafa macho* 'the male giraffe', *el gorila hembra* 'the female gorilla'. Moreover once outside the animate class, nouns for inanimate objects seem to be ascribed gender in an arbitrary fashion. A small girl getting dressed is faced with masculine *vestido* 'dress', *abrigo* 'coat', *calcetines* 'socks' and *zapatos* 'shoes', while *falda* 'skirt', *camisa* 'shirt', *bragas* 'underpants' and *botas* 'boots' are feminine. Spanish nouns for knife, fork and saucer are masculine and for kitchen knife, spoon and cup are feminine. In this separation of semantics and linguistic gender, Spanish is similar to other languages. MacWhinney *et al*. (1989) give similar German examples where fork is feminine, knife is neuter and spoon is masculine. The word for soap is masculine in French and Spanish, while in German it is feminine and in Russian it is neuter (Karmiloff-Smith, 1979).

Semantics are thus of limited use to the learner of gender agreement (see Tables 2 and 3). Morphology is also of limited use: there are no gender morphemes in Spanish nouns: no suffixes, prefixes or infixes. It might be argued that final *-o* and *-a* are gender bearing morphemes but this is only true for a limited class of animates as discussed above. Outside this class words differing only in these final vowels signify unrelated referents as in *libro* 'book' and *libra* 'pound' and *suelo* 'floor' and *suela* 'shoe sole' (Ambadiang, 1999). It is perhaps more appropriate to describe these final vowels as phonological cues, an issue to which we return.

The determiners (which precede nouns in Spanish) also present problems. Although feminine singular and plural determiners are regular in form with an *-a* ending in the singular and an *-as* ending in the plural, masculine singular determiners take a variety of forms (see Table 1) although the plural ends in a regular *-os*.

The class of feminine nouns starting with stressed *a-* takes the masculine singular determiner but the feminine plural determiner, e.g. *el agua* 'the water' but *las aguas* 'the waters'. This form of the singular determiner can

lead to constructions such as *El* (*m*) *otro* (*m*) a*v*e (*f*) *está enferma* (*f*) 'the other bird is ill' (Ambadiang, 1999). In addition there is a third, neuter, form of determiner, which may be paired with an adjectival form to construct a noun, e.g. *lo bueno* 'the good' or an adverbial form in an exclamation, e.g. *¡lo bien que lee este niño!* 'How well the child reads!' All three forms, masculine, feminine and neuter also function as pronouns, e.g. *esa casa* 'that house' feminine; *ese mantel* 'that table-cloth' masculine, giving as pronouns *ésa* and *ése* 'that one' (the marked stress in the written form does not change the pronunciation) and *eso* which is neuter, used for example in the common phrase *Eso es* 'that's it'; 'O.K.'. When pronouns are used for inanimate items as direct objects there are further complications: while *la* can be used as a direct object 'it' for a feminine object: *¿conoces esta cancion?* 'do you know this song?' *si, la conozco* 'yes, I know it', for 'it' referring to a masculine object the form is not *el* but '*lo*' *: ¿ quien tiene el libro?* 'who has the book?' *lo tengo* 'I have it'. Hence determiner forms in the singular are not straightforward. Plural forms are *-as* for feminine and *-os* for masculine.

This lack of any single reliable route to gender agreement in determiner–noun pairs makes the acquisition of these forms an especially interesting test of the plausibility of associative learning mechanisms.

Phonological cues to gender assignment

Given that all Spanish nouns are either masculine or feminine in gender (or both as in the cases of *dentista* and *medico* cited above) and that the gender of the preceding determiner (and/or following adjective) must show concord with the gender of the noun, the question remains as to how a Spanish child acquires the gender of nouns and consequently the form of the corresponding determiner. If neither semantics nor morphology provide easy routes perhaps children can make use of phonological cues.

Phonological cues to the gender of nouns are restricted to the last phoneme or phoneme cluster and are unreliable. Tables 2 and 3 outline the variety of endings for each gender (Leathes, 1984). These tables are given to inform readers without knowledge of Spanish, but it should be noted that the examples would not necessarily be present in the linguistic environment of very young children.

In this paper we identify the masculine nouns ending in *-o* (pl. *-os*) and the feminine nouns ending in *-a* (pl. *-as*) as being regular and all other nouns as being irregular. These terms undoubtedly overstate the case: a term used by Spanish academicians for the former classes is *majoritariamente* (Alarcos Llorach, E & Real Academia Española, 1994). This is because, although the endings of regular nouns provide quite reliable cues, there also exist many nouns which do not possess such cues. Indeed there are a sub-class of nouns

TABLE 2. *Cues to feminine gender in Spanish nouns*

Semantic cues	
● Female people, animals, etc.: *madre* – 'mother', *tía* – 'aunt', *reina* – 'queen', *vaca* – 'cow'	
● Letters of the alphabet: *la be*, *la ese*	
● Names of islands: *Mallorca*, *Ibiza*	
● Countries, regions, continents and generally towns, etc. ending in unstressed *-a*:: *Francia*, *Europa*, *Galicia*, *la Haya*	
Phonological cues	
Nouns ending in:	Main exceptions:
-a	*el día*, *mapa*, *planeta*, *tranvía* and most words ending in *-ma*: *el tema*, *problema*, etc.
-ción, *-sión* and most other *-ión*	*el avión*, *camión*, *gorrión*
-dad, *-tad*, *-tud*	
-dez	
-ed	*el césped*
-ie	*el pie*
-itis	
-iz	*el lápiz*, *tapiz*
-sis	*el análisis*, *énfasis*, *paréntesis*
-umbre	

which have the opposite ending to what one would expect: *papá* 'daddy', *mapa* 'map' and *día* 'day' are all masculine while *mano*, 'hand', *radio* 'radio' and *moto* 'motorbike' are feminine.

In considering the range of phonological cues and the possible regularities amongst them, it is useful to draw on Bybee's model of a network of associations (1995, 1999). She argues that as the child hears, and begins to store in memory, specific words, networks of phonological and semantic similarity are developed. Within these networks subparts will emerge e.g. the 'play' in 'plays', 'playing' 'played' and the '-ed' in 'played' and 'spilled' (Bybee, 1999). These networks she calls lexical schemas. Very high frequency use of specific lexical items in the speech input will result in independent representations for those items but as exposure frequency decreases, access is more dependent upon activation of components of the schemas. Because of the activation of associations, the lexical schemas support generalization of two sorts: source schemas match base forms, and their derived forms, e.g. 'play' and 'played', while product schemas match the derived forms directly, e.g. 'played' and 'spilled'. Production of novel items will be dependent upon these generalizations.

It is useful to consider these ideas in relation to the case of a child hearing specific Spanish nouns. The child will usually hear nouns preceded by an obligatory determiner. In a developing network of associations one would expect that nouns and determiners would emerge as subparts with links among nouns (phonological and semantic), among determiners (phonological and semantic) and between the two classes (phonological and semantic).

TABLE 3. *Cues to masculine gender in Spanish nouns*

Semantic cues

- Male people, animals, etc.: *padre*, *rey*, *toro*, *gallo*
- Days of the week, months, years, numbers
- Mountains (except where the word 'montaña' or 'sierra' is used), seas, deserts, winds, volcanos, points of the compass and usually rivers: *Sahara*, *Caribe*, *Sur*, *Amazonas*
- Countries, regions, continents and generally towns, etc. not ending in unstressed *-a*:: *Canadá*, *Artico*, *Ferrol*
- Musical notes: *fa* etc.
- Trees (except *haya*, *higuera*, *palmera*)
- Other parts of speech used as nouns: *el sí*, *el adiós*, *un no sé qué*, etc.
- Most compound nouns: *paraguas*, *pasatiempo*, *altavoz*

Phonological cues

Nouns ending in:	Main exceptions:
-o	*mano*, *foto*, *moto*
-e	*ave*, *calle*, *carne*, *clase*, *corte*, *fe*, *fiebre*, *frase*, *fuente*, *gente*, *hambre*, *leche*, *llave*, *mente*, *muerte*, *mugre*, *nave*, *nieve*, *noche*, *nube*, *parte*, *sangre*, *suerte*, *tarde*, *torre*
-i	*bici*, *metropoli*
-l	*cal*, *cárcel*, *catedral*, *col*, *miel*, *piel*, *sal*, *señal*
-r	*coliflor*, *flor*, *labor*
-u	*tribu*
-y	*ley*

Bybee terms the ability of a lexical schema to generalize to new items the 'productivity' of a schema. The extent to which a schema will be productive depends on openness to items and strength, which is determined by the type frequency of the pattern. Considered in terms of her model, patterns of Spanish nouns related by the phonology of their endings will be very open, since the endings are compatible with many phonological elements for the rest of the word. What is of interest therefore is the type frequency of patterns and whether these are related to the construction of gender agreed determiner–noun pairs by children in the course of acquisition. It is also of interest whether type frequency can explain the errors in gender assignment made by children.

Developmental course

In a longitudinal study López-Ornat (1997) found that between the ages of 1;7 and 2;0 a higher proportion of masculine rather than feminine nouns were produced. Between 2;1 and 2;2 determiner–noun pairs became more frequent and some errors were made, there being fewer errors in feminine than masculine determiner–noun pairs. This was followed by error-free production. This developmental pattern has been confirmed by cross-sectional data (López-Ornat, 2003).

A case study carried out by Hernández-Pina (1984) focused on a child named Rafael, who produced errors in the assignment of gender between the ages of 1;9 and 2;3. In the Hernández-Pina study, Rafael showed systematic overgeneralization in the case of certain nouns at age 1;11, by putting the masculine determiner indiscriminately before feminine nouns such as *llave* 'key', and *leche* 'milk'. Word-final *-e* is shown in Table 2 as a masculine cue, although there are many counter-examples. This behaviour was followed for a short while by systematically using the feminine determiner before certain masculine nouns – *camión* 'lorry', *pez* 'fish'. Again, we can see from Table 3 that nouns ending in *-z* and *-ión* tend to be feminine. Adjectives were sometimes regularized, a prime example is *tierra azula* 'blue earth', where the adjective *azul* 'blue' (which takes the same form in masculine and feminine) has been inappropriately regularized in concord with the feminine *tierra*. Hernández Pina also recorded a very rare error where an irregular feminine noun *moto* 'motorbike' was regularized in concord with the feminine adjectival form *rota* 'broken'; *mota rota*. These findings suggest Raphael was overgeneralising when assigning gender to nouns and that the overgeneralization changed with development.

A study by Mariscal (1997) of gender agreement in the use of the determiner *otro / otra* 'other' in the spontaneous speech of María (the child to whom the speech used in our experiment was directed) together with the spontaneous and elicited productions of six other children found that when the child's use showed productivity (at around 2;0 to 2;6), there was a period of errors in agreement. These errors were not frequent, but included mismatches: *oto bola* 'other marble', overgeneralizations: *oto mano* 'other hand' and even token variability by the same child: *ota utara*, *oto utaro* 'other spoon'.

Correct gender agreement in spontaneous speech by the age of about three in Spanish is matched in the more complex German system (Mills, 1986), but this performance cannot be taken as evidence of full mastery of the regularities of gender in the language as a whole. Native speaking primary school children faced with nonsense nouns in Spanish (Pérez-Pereira, 1991), nonsense nouns in German (Wegener, 2000) and object names and nonsense nouns in Icelandic (Mulford, 1985) do not assign gender in accordance with the probabilities of phonological and/or morphological analogues in the language as a whole. The children in Pérez-Pereira's study showed a specific bias: they were better at producing the appropriate determiner when given a noun with masculine cues and more likely to assign masculine gender to nouns with ambiguous cues (Pérez-Pereira, 1991). This evidence of dominance by masculine gender, despite the lack of surface transparency in masculine determiners in the singular, might argue for a special 'default' or 'unmarked' status for the masculine in noun phrases. Pérez-Pereira writes: 'The tendency to attribute masculine gender to nouns more often, and the higher results obtained with items providing masculine clues, seem to

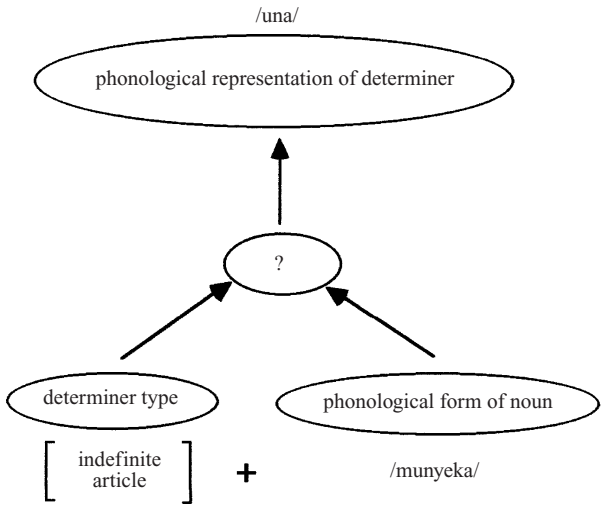

Fig. 1. The net, given the determiner type and a phonological representation of a noun, must produce a phonological representation of the required determiner.

support Greenberg's theory of markedness (Greenberg, 1966). According to this, the masculine is the unmarked term, and, thus, easier to acquire' (Pérez-Pereira, 1991: 584). The same effect on production might equally result from distributional aspects of the child's linguistic environment in the early stages, i.e. greater frequency of masculine types and/or tokens. This is the issue we investigate here.

It has been argued by Redington & Chater (1998) that the validity of computational tests of associative learning mechanisms depend crucially on whether the input accurately reflects the child's linguistic environment. To test the hypothesis we advance, one needs the actual linguistic input to a child over the relevant time period. Relative frequencies in adult speech and writing are of marginal relevance and cannot provide a strong test of acquisition by associative learning. We were fortunate in having a sample of speech input to a child over the period of gender agreement acquisition. Connectionist models build representations of the co-occurring regularities in the input. By restricting ourselves to the frequencies in a sample of naturally occurring speech by parents and grandparents of a young Spanish child, María, we take account of criticisms of connectionist models that involve direct manipulation of the input (Pinker & Prince, 1988). We also further maximize the ecological validity of the model by using a linguistically motivated phonological representation of words and by presenting the input incrementally to the computational model, to mirror the child's experience over time.

TABLE 4. *The number of determiner-noun pairs input at each incremental stage during training*

María's age	1;7	1;10	2;1	2;4	2;7	2;11
Incremental lexicon size	351	922	1398	1803	2048	2285

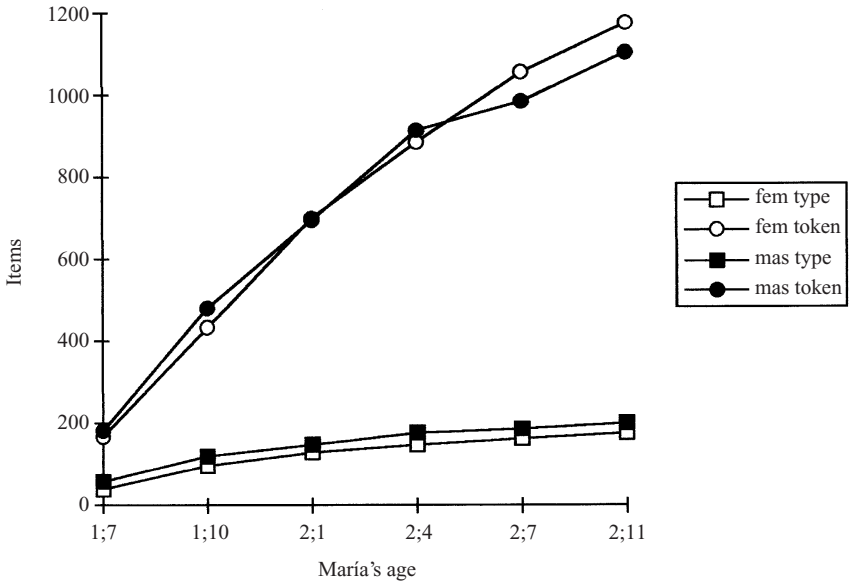

Fig. 2. Type and token frequencies of feminine and masculine NPs in the María database.

In our experiment a connectionist network has the task of acquiring the ability to produce phonological representations of the appropriate Spanish determiners when given phonological representations of Spanish nouns together with identification of the determiner type required. This might be thought of as analogous to children accessing a mental representation of a noun and having to produce the appropriate determiner–noun phrase.

METHOD

The connectionist network was presented with a phonological representation of a noun and an arbitrary code which identified the type of determiner (e.g. definite article, indefinite article, etc.). Using this information the network was trained to produce a phonological representation of the determiner (see Figure 1).

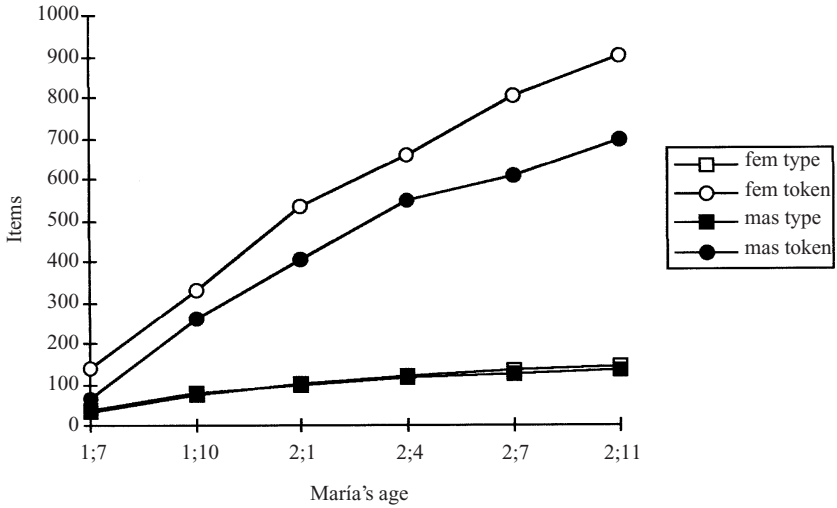

Fig. 3. Type and token frequencies of regular feminine and masculine NPs in the María database.

Training set

A longitudinal study of a child, María, conducted by López Ornat in Madrid (López Ornat *et al.*, 1994) was used as the basis of the simulation. María was recorded twice monthly in the presence of her parents and members of her family between the ages of 1;7 and 3;11. Sessions typically lasted for about half an hour. This study is available on the ChiLDES database (MacWhinney & Snow, 1985; MacWhinney, 1991).

Transcriptions of the parental productions in a machine-readable form were organized into three-monthly chunks which formed the basis of an incremental training regime for the network. No data were available for the month 2;10 and therefore the database consisted of 6 three-monthly transcription files which reflected the items that María had been exposed to around 1;7, 1;10, 2;1, 2;4, 2;7 and 2;11 (data from 3·0 to 3·11 were not included in this study). For practical reasons described below, any nouns which were longer than eight phonemes in their plural form were dropped from the training data (less than 1% of the entire training set). The number of determiner–noun phrases in each training set is shown in Table 4.

In order to test an associative hypothesis it is necessary to know the relative type/token frequencies of the María database. Figure 2 shows the relative numbers of items (tokens) and noun types (types) in the training set at the six incremental training stages. This picture shows that the training set is split roughly into an equal number of masculine and feminine nouns.

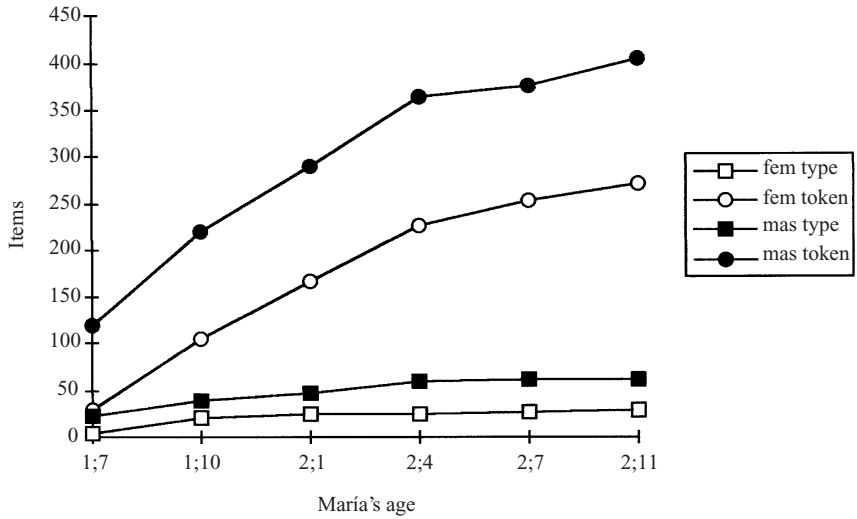

Fig. 4. Type and token frequencies of irregular feminine and masculine NPs in the María database.

Looking at the lexicon in terms of whether the nouns are regular or irregular, however, shows that this equal masculine/feminine split is misleading. Figure 3 indicates the type and token frequencies for the regular nouns and reveals that while there are roughly the same number of regular masculine and feminine types, the feminine types have a much higher token frequency.

A different pattern is seen in the case of masculine and feminine irregular nouns (Figure 4). At all stages there are roughly twice as many masculine types as there are feminine. Additionally there are far more masculine tokens than feminine. This suggests an asymmetry between regular and irregular nouns in the speech directed to Spanish children.

Based on these relative frequencies we would predict that after an early advantage for feminine determiners, the network would be increasingly more successful with masculine determiners, particularly for irregular nouns and would tend to assign masculine determiners to nouns with ambiguous cues.

Phonological representation

To maintain correspondence with actual input to the child, it was necessary to represent Spanish nouns in a theoretically plausible way. Each consonant and vowel in the input and output representations, therefore, was coded using a pattern of features distributed (a series of 1s and 0s) across seven units, reflecting standard phonological contrasts such as voiced/unvoiced, coronal/labial/velar, front/central/back, etc. to which Spanish speakers are

TABLE 5. *The seven bit phonological code used to represent phonemes*

	Vowel	Voice	Manner			Place	
/t/	0	0	0	1	1	1	0
/d/	0	1	0	1	1	1	0
/k/	0	0	0	1	1	0	0
/g/	0	1	0	1	1	0	0
/p/	0	0	0	1	1	1	1
/b/	0	1	0	1	1	1	1
/f/	0	0	0	1	0	1	1
/v/	0	1	0	1	0	1	1
/s/	0	0	0	1	0	0	1
/T/	0	0	0	1	0	1	0
/¢/	0	0	1	0	1	1	0
/h/	0	0	0	0	1	0	0
/y/	0	1	0	0	1	0	0
/r/	0	1	0	0	1	0	1
/l/	0	1	0	0	1	1	0
/w/	0	1	0	0	1	1	1
/m/	0	1	1	0	0	1	1
/n/	0	1	1	0	0	1	0
	Vowel	Voice	Front/middle/back			High/med/low	
/a/	1	1	0	1	0	0	1
/e/	1	1	1	0	0	1	0
/i/	1	1	1	0	0	1	1
/o/	1	1	0	0	1	1	0
/u/	1	1	0	0	1	1	1
–	0·5	0·5	0·5	0·5	0·5	0·5	0·5

sensitive (see Table 5). Distributed representations of this kind are useful in modelling aspects of language in that they encode information in such a way that similarities and dissimilarities between phonemes are preserved.

The precise nature of the phonemic code is described by Nix (1997) and is an adaptation of the phonemic code used by Plunkett & Marchman (1993). This phonemic coding was utilized for nouns at the input layer and determiners at the output layer. Thus, in order for the network to produce a correct phonemic representation of the determiner it had to extract gender information from the phonemic information present in the noun and combine this with information of the type of determiner required (represented at the input layer by an arbitrary pattern of activation over three units as shown in Table 6) to produce a full phonemic representation of the correct determiner at the output layer.

Each phoneme was represented by a distributed phonological code over seven processing units or nodes. With a maximum field of eight phonemes for each noun the input field consisted of 59 units, 3 for the determiner code and 56 for the noun code. The maximum length of a noun, therefore, was seven phonemes in its singular form and eight phonemes in its plural form

TABLE 6. *Arbitrary codes used for identifying determiner type in the input*

Article type	Input representation
definite	110
indefinite	010
demonstrative 1	001
demonstrative 2	101
comparative	111

TABLE 7. *Positioning of nouns at the input layer*

1	2	3	4	5	6	7	8
			k	a	s	a	
			k	a	s	a	s
m	u	n	y	e	k	a	
m	u	n	y	e	k	a	s

(see Table 7 for examples of how the words *casa* and *muñeca* are presented in their singular and plural forms). In order to align potential gender-carrying segments the nouns were presented so that the final phoneme of the singular form always occupied the seventh phoneme position. The eighth phoneme position was occupied by the number carrying segment, /s/, where relevant. Nouns of less than seven phonemes in length were padded to the left with intermediate values, 0·5 in this case, signifying that those units were neither on or off.

The output layer allowed for a five phoneme determiner field consisting of 35 units since the longest determiner was five phonemes in length in its plural form (see Table 8). This configuration is described in a graphical form in a later section (see Figure 5). The last phoneme of the singular form always occupied the fourth position, (except for the masculine form *el* 'the' which has an irregular syllable structure), while the last phoneme of the plural form occupied the fifth position.

This arrangement was used to add extra saliency to the gender-carrying parts of nouns and determiners. Simple feedforward networks of this type are not very good at identifying identical strings which are displaced in the input representation (referred to as translation invariance (Plunkett & Elman, 1997: Ch. 7)), whereas humans are very good at this task. On the reasonable assumption that the child being modelled has already learned to segment the speech stream into words, we may have a better model of human information processing by aligning the ends of words of varying lengths.

TABLE 8. *Positioning of determiners at the output layer*

		l	a	
	e	l		
	u	n	a	
	e	s	o	s
e	s	t	a	s

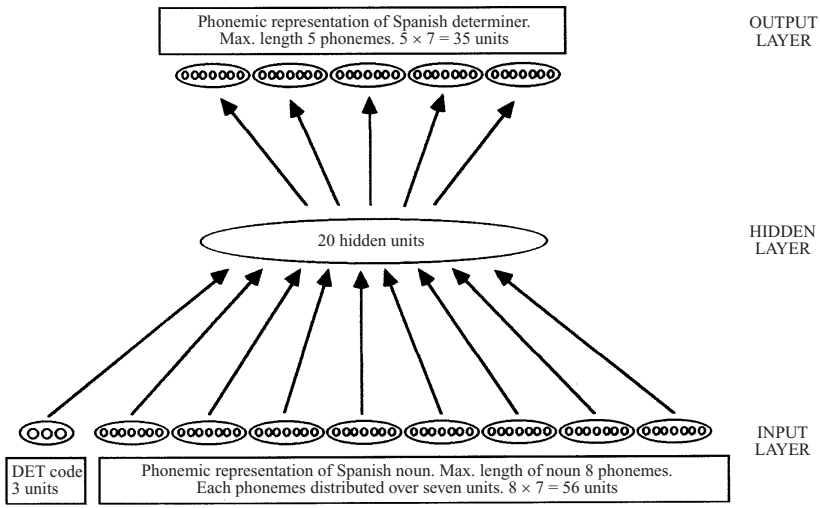

Fig. 5. The network architecture.

Network architecture and training schedule

The network had 59 input units, 20 hidden units and 35 output units (see Figure 5). The 59 input units were split up into three units defining the determiner type code, together with 56 units representing a potential eight-phoneme noun (as each phoneme was distributed over 7 units). Similarly the 35 output units represented a potential five-phoneme determiner.

The choice of training schedule and other variables such as number of epochs and values for learning rate and momentum is typically the result of experimentation rather than any principled criteria in connectionist methodology. In order to investigate this issue we initially trained the network for 100 epochs (100 presentations of each pattern in the training set) for each of the six incremental training sets. We then trained the network for 1000 epochs for each of the sets. The error signal was recorded at every epoch and the network's weights matrices were recorded at the end of each incremental training run to allow testing at strategic points during the training schedule. The error curves for the network trained for 100 epochs per increment and the network trained for 1000 epochs per increment are shown in Figure 6.

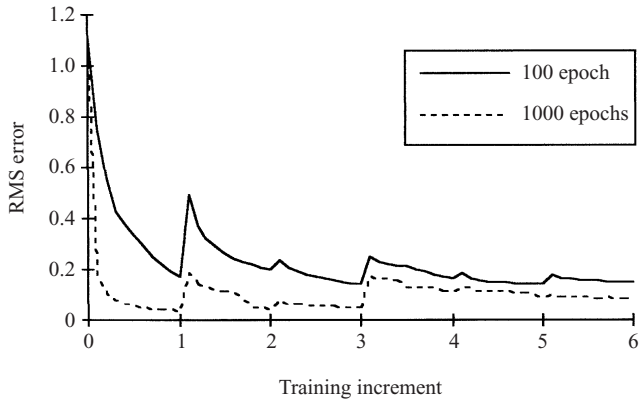

Fig. 6. Error curves for networks trained for 100 and 1000 epochs per increment.

The periodic peaks in the curve coincide with an influx of new items into the training set at the beginning of each training increment. The error curves show that initially the error is much lower for the network trained for 1000 epochs per increment. But as successive increments are made to the training file, this difference becomes less and less. It is well documented that driving down the error rate too low, early on in training, can result in 'over-training' which has the effect of limiting a network's ability to generalize to novel data (Tesauro & Sejnowski, 1988). Considering this and the fact that by the final increment the error rates are not too different, it was decided to analyse the results of the network trained for 100 epochs.

With learning rate and momentum values (the parameters which control the extent to which weights are changed after each pattern is presented) a period of experimentation revealed that a learning rate of 0·1 and a momentum of 0·5 with weights updated after each pattern provided reasonable error reduction over the 100 epoch increments. All simulations were run using the TLEARN simulator (Plunkett & Elman, 1997).

Testing

Test patterns (see details of the test items below) were presented to the network at the six successive training intervals using the saved weight matrices. This procedure allows testing for generalization at successive stages in learning. The output activations were recorded for each test pattern and were evaluated in the following way. Phoneme by phoneme, a closest fit in Euclidean space was calculated for each of the output patterns. Error analysis was carried out in terms of whether or not the network produced the intended phonemic sequence for the determiner. Even if the

TABLE 9. *High frequency nouns used in Test Set* 1

Fem irreg	Fem reg	Mas irreg	Mas reg
mano	cosa	nene	cuento
calle	niña	pié	perro
vez	vaca	dia	beso
leche	caca	coche	culo

network produced only one phoneme error for a word it was counted as being incorrect.[2]

Test Set 1 – high frequency combinations

Test Set 1 was based on the four highest frequency masculine regular nouns, masculine irregular nouns, feminine regulars and feminine irregulars, from the María database (see Table 9). These sixteen nouns were presented to the network with determiners that they had not been paired with in the training data. This test set consisted of 98 NPs. Although the nouns were familiar to the network, they had never been presented in these combinations to the network during training.

Test Set 2 – novel nouns

The second test set contained 12 nonsense words taken from the Pérez-Pereira experiment (Pérez-Pereira, 1991). In experiments involving children aged between three and eleven years of age, Pérez-Pereira varied three different variables of semantic, syntactic and phonological clues when presenting pictures of imaginary objects and creatures to the children. He then asked the children to say what the pictures were of, hoping to elicit a gender carrying NP.

The results of the Pérez-Pereira study (see Figure 7) reveal that Spanish children are more likely to assign masculine gender to a novel noun regardless of how many cues they are given. Figure 8 shows that the masculine/feminine distinction is preserved for items which have typically regular endings. The children tended to assign masculine gender to items with irregular endings even when the experimenter introduced these items with a feminine determiner.

[2] At the suggestion of anonymous reviewers, we also analysed the results by selecting the closest fit with the nearest *legitimate determiner*. This would eliminate errors of the type not found in children, e.g.'ulo'. In fact, while decreasing the error rate very slightly, this analysis did not affect the pattern of errors across the masculine/feminine; regular/irregular variables.

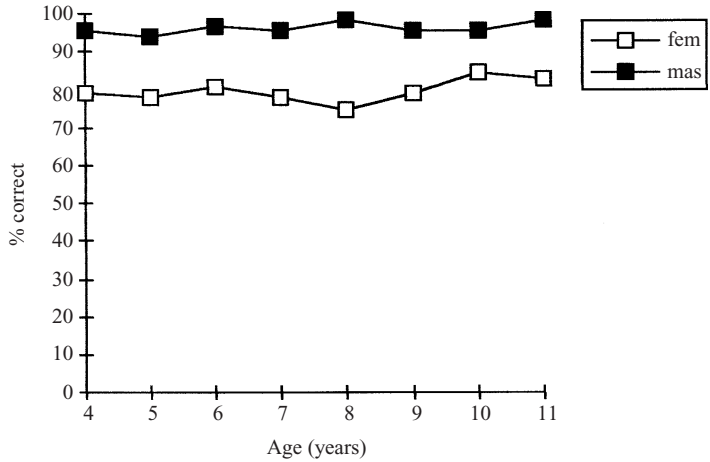

Fig. 7. Overall results of children's gender assignments to novel nouns in experiments by Pérez-Pereira (1991).

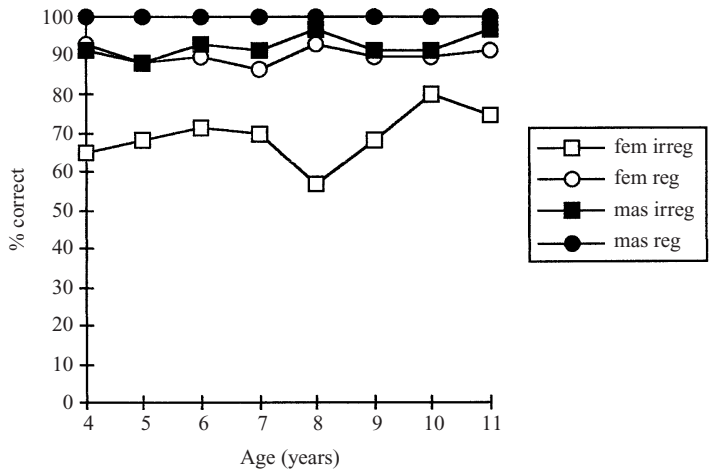

Fig. 8. Children's gender assignments for regular and irregular novel nouns from Pérez-Pereira (1991).

A subset of Pérez-Pereira's experimental items were used to make up Test Set 2 (see Table 10) below. All of the nonsense words were phonologically allowable strings in Spanish. These items were presented with the five determiners in singular and plural form. This test set, therefore, consisted of 120 NPs.

TABLE 10. *Novel nouns with regular feminine, regular masculine or ambiguous endings used in Test Set 2*

Fem reg	Mas reg	Ambiguous
anira	carepo	pifar
lodena	milipo	pilin
satila	linolo	liben
tica	nepo	tanten

Using novel nouns, to which the network had never been exposed, tested whether the network was assigning gender using the phonological cues present in noun endings and whether it would respond to novel nouns in a similar way to the children in the Pérez-Pereira study.

RESULTS

The network's performance was analysed in relation to masculine nouns compared with feminine nouns, and regular nouns compared with irregular nouns. Given the lack of straightforward phonological cues to gender for the irregular nouns, we expected to see lower performance on these than for regular nouns. However, within the regular group we expected to see more feminine determiners produced correctly than masculine due to the regularity of the feminine determiner system compared with the variability of the masculine singular determiners. Within the irregular group, we expected that the increasing frequency of masculine irregulars in the input compared with feminine irregulars (see Figure 4) would lead to better performance on that class compared with feminine irregulars.

Test Set 1 – high frequency nouns in novel combinations

Figure 9 shows the overall performance of the network when presented with the high frequency nouns from the test set. The performance on feminine determiners is only narrowly better than the performance on masculine determiners between 1 ; 7 and 2 ; 1 (increments 1 and 3 on the x-axis). After this point success in the production of masculine determiners overtakes the feminine reaching its peak at 2 ; 7 (increment 5).

Figure 10 shows the same information broken down into regular and irregular items. It is clear from this that the poorer overall performance on feminine determiners is largely due to the irregular items. Performance on regular feminine determiners peaks at almost 100% at 2 ; 1 (increment 3) only to be overtaken by masculine regulars at 2 ; 4 (increment 4). Despite this, performance is still remarkably strong – never falling below 80% after its peak. Our findings, like those of Pérez-Pereira, were that the highest

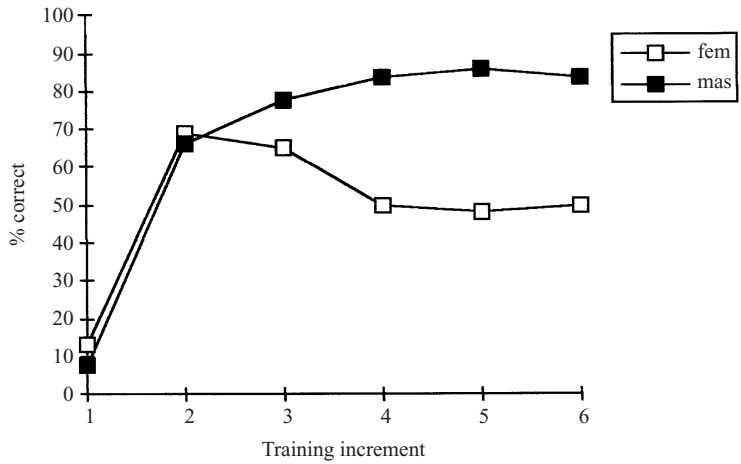

Fig. 9. The network's performance on feminine and masculine determiners for Test Set 1.

percentage of errors occurred with feminine irregulars. A closer look at the errors in the feminine irregular group revealed that two of the items *leche* and *calle* often were being treated as masculine. The *-e* -ending is an extremely ambiguous cue occurring in many irregular feminine and masculine nouns. In the training lexicon, however, a greater number of masculine nouns ending in *-e-* were present in the speech of adults to María. The network appears to be adopting a mapping strategy for these nouns in which the masculine version wins out. It is important to note, however, that some of the errors which occur for the masculine irregulars are of the type where masculine *-e* nouns are assigned a feminine version of their determiners, indicating inconsistencies in performance, reminiscent of the errors found by Hernández-Pina (1984). The masculine solution, however, dominates in the majority of cases.

The net's performance of correct output for Test Set 1 (trained nouns each paired with a determiner type with which it had not been paired during training) was entered into a gender (masculine vs. feminine) by regularity (regular vs. irregular) ANOVA with repeated measures of the training increment factor. There were main effects for gender, with masculine determiners being produced correctly significantly more often than feminine determiners ($df = 1$, $F = 9{\cdot}499$, $p < 0{\cdot}01$) and regularity, with determiners paired with regular nouns being produced correctly significantly more often than determiners paired with irregular nouns ($df = 1$, $F = 30{\cdot}788$, $p < 0{\cdot}001$), and increment ($df = 5$, $F = 66{\cdot}964$, $p < 0{\cdot}001$). There was an interaction between gender and regularity ($df = 1$, $F = 12{\cdot}489$, $p < 0{\cdot}001$) with irregularity of nouns decreasing performance on feminine items to a greater extent than on masculine items.

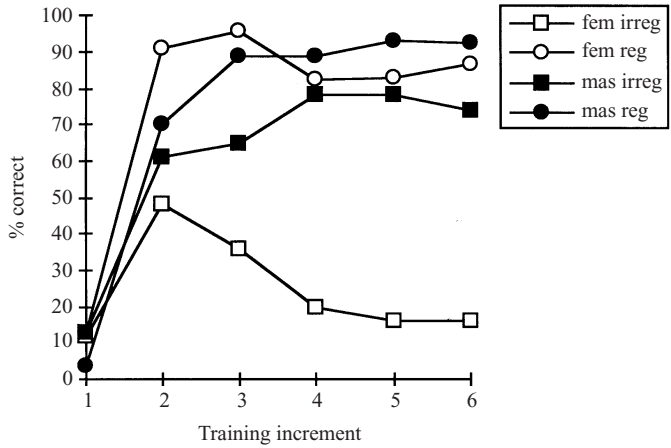

Fig. 10. The network's performance on regular and irregular feminine and masculine determiners in Test Set 1.

The main effect for increment is hardly surprising but of more interest is the interaction between increment and gender (df = 5, $F = 10{\cdot}414$, $p < 0{\cdot}001$) and increment and regularity (df = 5, $F = 10{\cdot}070$, $p < 0{\cdot}001$). The biggest increase of performance was between the first and second increment. Thereafter it can be seen that there is a decrease in performance for determiners paired with irregular feminine nouns, while determiners paired with regular nouns and irregular masculine nouns continue to improve (see Figure 10).

Test Set 2 – novel nouns

Test Set 1 assessed the network's performance when given nouns it had already learned in novel combinations with determiners. A further test involved presenting the network with novel nouns. Test Set 2, made up of twelve nonsense nouns from Pérez-Pereira's experiment (see Table 10), was given to the network with each of the five determiner types in singular and plural forms. Four items were novel strings ending in *-o* (*-os*), the dominant form of masculine nouns and four items were novel strings ending in *-a* (*-as*), the dominant form of feminine noun. The remaining four items had ambiguous endings *-r* (*-es*), and *-n* (*-es*) which were analogous to irregular nouns of either gender. Thus the net was tested on generalization to legitimate but novel strings. Of particular interest is the classification given by the net to the strings with ambiguous endings.

The results of the test with novel strings are given separately for those with regular masculine endings, those with regular feminine endings and those with irregular endings. Performance is classified according to the gender

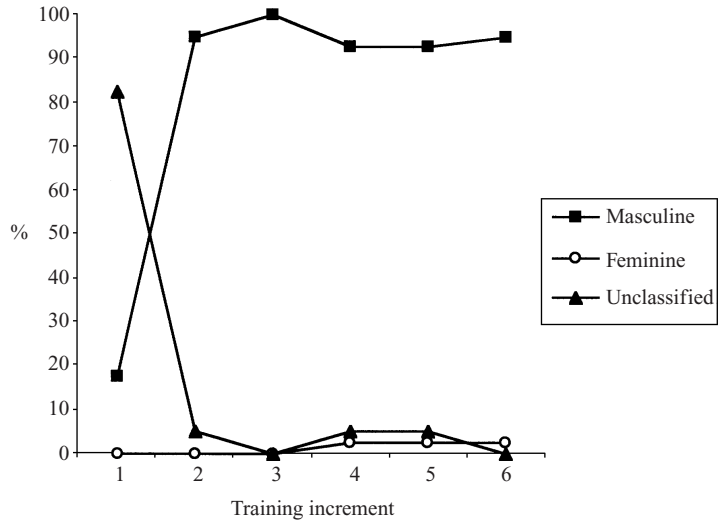

Fig. 11. The network's gender assignment for novel nouns ending in *-o* (*-os*), the dominant ending for masculine nouns in Spanish.

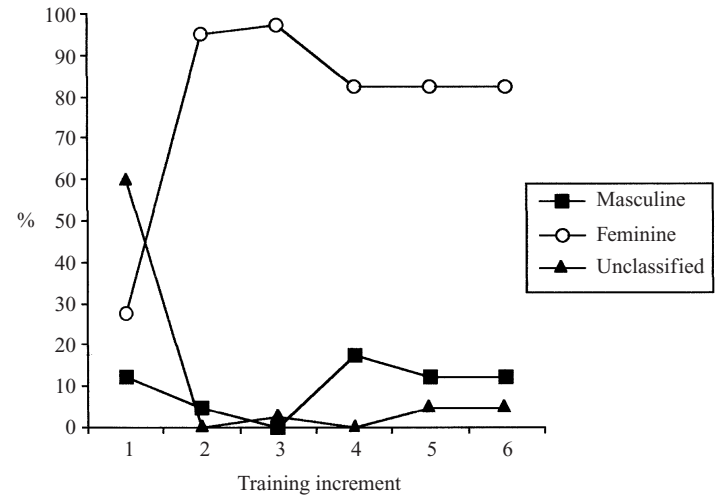

Fig. 12. The network's gender assignment for novel nouns ending in *-a* (*-as*), the dominant ending for feminine nouns in Spanish.

of the determiner produced in response to the input of that novel item and a determiner type. Any response which was not a fully formed determiner of the correct type (e.g. definite: plural) was placed in the 'Unclassified' category. Figure 11 shows the performance of the net given the strings ending with *-o* and *-os*.

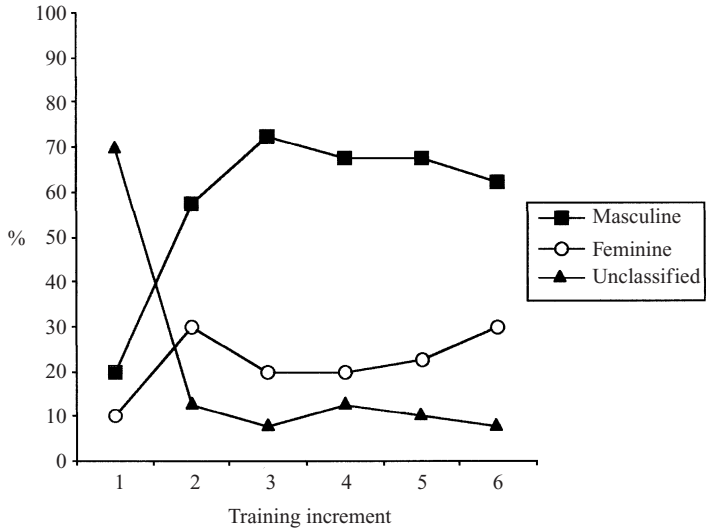

Fig. 13. The network's gender assignment for nouns with ambiguous endings.

As early as the second increment, the net consistently assigned the masculine gender to novel strings with the regular masculine ending. Figure 12 shows slightly less consistency in assigning feminine gender to novel strings with the regular feminine endings *-a* and *-as*.

This ability of the net to produce fully formed determiners to novel input with the regular endings is evidence of generalization from training set to novel items. It is thus interesting to see what was produced in response to novel items with irregular endings. Figure 13 shows some uncertainty with around ten percent being unclassified, but more striking is the discrepancy between the proportions assigned to the two genders. There is a clear bias towards masculine assignment as early as the second increment, a response firmly established by the third increment.

This pattern of results suggests that even when presented with novel items, the network has learned very early on to identify the phonological cues which mark gender in regular nouns. When faced with novel items with ambiguous endings, there is a strong bias towards assigning masculine gender.

DISCUSSION

The experimental research described in this paper set out to model the acquisition of gender agreement in determiner–noun phrases using an incremental training set which preserved the type and token frequency information found

in a child's linguistic environment. The results demonstrate the learnability of these forms from the phonological information alone. They also show that a network trained on data taken from child-directed speech can generalize production to novel combinations of determiner + noun and also to novel nouns, producing a pattern of performance similar to that reported for children and providing an explanation for the children's bias toward assigning masculine gender to novel items.

It is of interest to compare the performance of the network with that of Spanish children. Research by Mariscal (1997) focussed on gender agreement of nouns with the determiner *otro* 'other' and with adjectives. She looked at the spontaneous productions of María (the child to whom the speech used in our experiment was directed) and the spontaneous and elicited speech of six other children around 2;0 to 2;6 months. Mariscal's data suggest a period of instability in gender assignment when the child begins to use determiners and adjectives productively. This is followed by stable, correct use. Two aspects of her findings are of particular interest here. First, she found a low rate of errors, but all the possible errors in gender agreement were present, including incorrect assignment, e.g. *oto lupa* for *otra lupa* 'other magnifying glass', overgeneralization, e.g. *oto mano* for *otra mano* 'other hand' and even an invention *ota botona* for *otro boton* 'other button'. Our net could not, of course, change the noun but did produce all error forms early in learning and then continued to make errors in assignment on irregular nouns, particularly feminine irregular nouns which were less frequent in the training data than masculine irregulars. Second, she found the children's errors on *otro* disappeared at about the time that their omissions of determiners in obligatory contexts fell below 10%. This suggests that acquisition of gender agreement for any one determiner is part of learning the form and use of the syntactic class. Moreover this may be intrinsically linked with the learning of gender assignment in another class, adjectives. Mariscal found that although adjectives (which in Spanish usually follow the noun) had a lower error rate than determiners, there was broad concordance in timing between the correct gender agreement with *otro/otra* and the correct gender agreement of adjectives. The continued problem which our net had with irregular feminine nouns might be mitigated if feminine adjectives were added to the input.

However, research on children between 4;0 and 11;0, suggests that children are more likely to assign masculine gender to a phonologically ambiguous noun (Pérez-Pereira, 1991). In our data the greater frequency of masculine over feminine irregulars in the child directed speech provides an explanation for such a bias. We do not know of any study looking at this aspect in other samples of child directed speech in Spanish so this hypothesis remains tentative, but suggests a distributional explanation for Pérez-Pereira's findings.

Connectionist models of language acquisition have been criticized in the past for manipulating type and token frequencies to produce desired results (Pinker & Prince, 1988). By using type and token frequencies taken directly from a database of child-directed speech we avoid this problem. Furthermore, by maintaining realistic increments to the training set we can legitimately analyse the changing performance of the model against the changing performance of a child. The net's performance lags behind the child in that it does not reach the error free performance on familiar items that a child of 2;11 would probably exhibit (López-Ornat, 1997; Mariscal, 1997), but there are obvious reasons for this: the net has only phonological information whereas the child has supporting semantic information for at least some of the animate items; the incremental training is only a crude approximation to the child's age-related exposure and, perhaps of greatest significance given Mariscal's findings, the net is not given adjectives. Some adjectives in Spanish have the same form in masculine and feminine, e.g. *grande* 'big' and *azul* 'blue' but most have regular masculine and feminine forms, e.g. *amarillo* 'yellow' (m) and *amarilla* 'yellow' (f). The demonstration that children acquire gender agreement between 2;0 to 2;6 months in both determiners and adjectives suggests that these may provide mutually supporting cues.

In our experiment a connectionist network was incrementally exposed to determiner–noun pairs which occur in child directed speech, based on the linguistic environment of a specific child over time. Using these data it was able to successfully develop a strategy to assign gender to determiners when tested on novel nouns. Furthermore, the decisions about gender assignment were very similar to those shown by young children. These findings provide further support for the idea that language acquisition is influenced by the frequency distribution of relevant aspects of child directed speech. For example the tendency to assign masculine gender to novel nouns appears to reflect distributional frequencies.

By developing a connectionist model of gender assignment in Spanish determiner noun pairs we have been able to extend this form of research, ensuring that the model is dependent upon frequencies in actual child directed speech. We hope that the approach we have adopted and the findings from the modelling contribute to the debate on the part that connectionist modelling can play in accounting for language acquisition.

REFERENCES

Alarcos Llorach, E.: Real Academia Española. (1994). *Gramatica de la lengua española*. Madrid: Espasa.

Ambadiang, T. (1999). La flexión nominal. Género y número. In I. Bosque Muñoz & V. Demonte Barreto (eds), *Gramática Descriptiva de la Lengua Española*. *Vol. 3*. Madrid: Espasa.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes* **10**(5), 425–55.

Bybee, J. (1999). Use impacts morphological representation. *Behavioral and Brain Sciences* **22**(6), 1016–17.
Corbett, G. G. & Fraser, N. M. (2000). Gender assignment: a typology and a model. In G. Senft (ed.), *Systems of nominal classification*. Cambridge: CUP.
Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* **48**(1), 71–99.
Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: MIT Press.
Greenberg, J. H. (1966). Language universals. In T. A. Sebeok (ed.), *Current trends in linguistics*. The Hague: Mouton.
Hernández Pina, F. (1984). *Teorías psicosociolingüísticas y su aplicación a la adquisición del español como lengua materna*. Madrid: Siglo XXI.
Karmiloff-Smith, A. (1979). *A functional approach to child language*. Cambridge Studies in Linguistics, Vol. 24. Cambridge: CUP.
Leathes, R. N. d. M. (1984). *A concise Spanish grammar*. London: Murray.
López Ornat, S. (1988) On data sources on the acquisition of Spanish as a first language. *Journal of Child Language* **15**, 679–86.
López Ornat, S. (1997). What lies in between a pre-grammatical and a grammatical representation: evidence on nominal and verbal form-function mappings in Spanish from 1;7 to 2;1. In W. R. Glass & A.-T. Pérez-Leroux (eds), *Contemporary perspectives on the acquisition of Spanish*. Somerville, MA: Cascadilla Press.
López Ornat, S. (forthcoming in 2003). Grammatically-relevant variation in speech before 22 months; syllabic and (pre)morpholical MLU in Spanish. In S. Montrul (ed.), *Studies in the acquisition of Spanish and Portuguese*. Somerville, MA: Cascadilla Press.
López Ornat, S., Fernandez, A., Gallo, P. & Mariscal, S. (1994). *La adquisición de la lengua española*. Madrid: Siglo XXI.
MacWhinney, B. (1991). *The CHILDES project: computational tools for analyzing talk*. Hillsdale, NJ: Erlbaum.
MacWhinney, B., Leinbach, J., Taraban, R. & McDonald, J. (1989). Language learning: cues or rules? *Journal of Memory and Language* **28**(3), 255–77.
MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language* **12**, 271–96.
Mariscal, S. (1997). *El proceso de gramaticalization de las categorias nominales en español*. Doctoral Dissertation. Universidad Autónoma de Madrid, Spain.
Mills, A. E. (1986). *The acquisition of gender*. Berlin: Springer.
Mulford, R. (1985). Comprehension of Icelandic pronoun gender: semantic versus formal factors. *Journal of Child Language* **12**, 443–53.
Nix, A. (1997). A connectionist inquiry into the acquisition of gender harmony in Spanish noun phrases. Ph.D. Thesis. University of Hertfordshire, U.K.
Pérez-Pereira, M. (1991). The acquisition of gender: what Spanish children tell us. *Journal of Child Language* **18**(3), 571–90.
Pinker, S. & Prince, S. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**, 73–193.
Pinker, S. & Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In S. D. Lima, R. L. Corrigan & G. K. Iverson (eds), *The reality of linguistic rules*. Philadelphia: Benjamins.
Plunkett, K. & Elman, J. L. (1997). *Exercises in rethinking innateness*. Cambridge, MA: MIT Press.
Plunkett, K. & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition* **48**, 21–69.
Redington, M. & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: a distributional perspective. *Language and Cognitive Processes* **13**, 129–91.
Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tense of English verbs. In D. E. Rumelhart, J. L. McClelland & the PDP. Research Group (ed.), *Parallel*

distributed processing: explorations in the microstructure of cognition. Vol. 2: Theoretical issues. Cambridge, MA: MIT Press.

Tesauro, G. & Sejnowski, T. J. (1988). A 'neural' network that learns to play backgammon. In D. Z. Anderson (ed.), *Neural information processing systems* (*Denver 1987*). New York: American Institute of Physics.

Wegener, H. (2000). German gender in children's second language. In B. Unterbeck & M. Rissanen (eds), *Gender in grammar and cognition*. Berlin/New York: Mouton de Gruyter.