A Connectionist account of Spanish determiner production

Andrew Nix, Neil Davey, David Messer, Pamela Smith
University of Hertfordshire
Hatfield, UK

Abstract

A Connectionist Network that models the production of simple phonologically coded Spanish Noun Phrases is described. The training data uses type/token frequencies taken directly from a Spanish child's linguistic environment. The training set increases in size in a manner which mirrors the increasing complexity of the real linguistic environment. The results show that the model can learn the task and generalise to unseen Noun Phrase combinations. Moreover the generalisation performance is of a similar nature to that of Spanish children.

1 Introduction

Research into the acquisition of Spanish gender has revealed that masculine expressions are acquired with equal if not greater ease than feminine expressions despite the fact that masculine articles have irregular morphology. Evidence in support of this interpretation is present in previous research which has shown that, Spanish children pay more attention to morphophonological cues present in nouns than to natural semantics when assigning gender. They are better at producing the correct determiner when given a noun with masculine cues and are more likely to assign masculine gender to nouns with ambiguous cues (Pérez-Pereira, 1991).
The research described here is an attempt to model Spanish determiner production using a connectionist network and is characterised by the following key points:

- The training data uses type/token frequencies taken directly from a Spanish child's linguistic environment.
- The training set increases in size in a manner which mirrors the increasing complexity of the real linguistic environment.
- The results show concordance with the results of Pérez Pereira (1991) - the model produces similar behaviour, to a child, with respect to gender assignment generalisation.

2 Spanish Gender Harmony

Spanish nouns which end with **-a** are generally feminine, while those ending with **-o** are generally masculine. However, there are many exceptions to this pattern.

Spanish determiners have to agree in gender with the noun. Feminine determiners are regular in that they are all marked with the -a suffix. Masculine determiners, however, exhibit varying degrees of irregularity in the singular but are regular in the plural form.

Determiner	**Feminine**		**Masculine**	
English equivalent	*Singular*	*Plural*	*Singular*	*Plural*
The	la	las	el	los
a/some	una	unas	un	unos
This/these	esa	esas	ese	esos
That/those	esta	estas	este	estos
(an)other/other	otra	otras	otro	otros

These five different determiner types were used in this study. While the task of gender agreement might seem a trivial task with regular nouns, problems occur when children are confronted with nouns with ambiguous cues such as mano, día, calle, coche, etc. In a study of Spanish children Pérez Pereira (1991) found that they are better at producing the correct determiner when given a noun with masculine cues and more likely to assign masculine gender to nouns with ambiguous cues. These findings are hard to explain, given the greater complexity of the masculine determiner system.

3 Network Architecture

A feedforward network with 59 input units, 20 hidden units and 35 output units was presented with phonological representations of Spanish nouns together with a 3-bit code to distinguish which determiner (DET) type was to be produced (See Figure 1). Phonemes were encoded using a 7-bit phonological representation (Nix, 1997, Plunkett & Marchman, 1993). The task for the network was to produce the correct phonological representation of the determiner, from a specification of the type of determiner required, e.g.:

Input = <*definite article*> + */gato/*
Output = */el/*

Input = <*indefinite article*> + */pupa/*
Output = */una/*

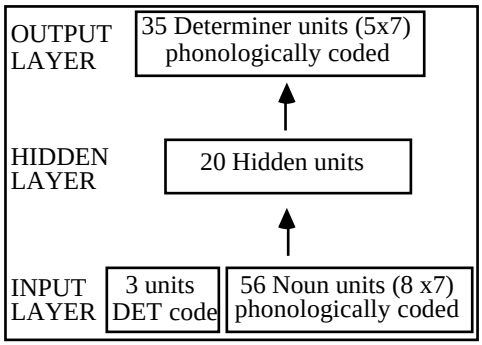


Figure 1: The feedforward network trained to produce phonologically coded determiners

4 Training

A longitudinal study of a child, María, conducted by Susana López-Ornat in Madrid was used as the basis of the simulation (López-Ornat, 1994). María was recorded in conversation with her parents and members of her family a total of 568 times between the ages of 1;7 and 2;11. Transcriptions of the *parental productions*, that were made available in a machine-readable form, were organised into three-monthly chunks which would form the basis of an incremental training regime for the network (Plunkett & Marchman, 1993). The database consisted of 6 three-monthly transcription files which reflected the items that María had been exposed to at 1;7, 1;10, 2;1, 2;4, 2;7 and 2;11:

Incremental training data - NP's per training file

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

Training took place using backpropagation with a learning rate of 0.1 and a momentum of 0.5 with weights being updated after each pattern. The network was trained for 100 epochs on the training set at 1;7 and the weights were saved. The lexicon was then increased to the 1;10 stage and was trained for a further 100 epochs and the weights were saved. This process was repeated until the network was being trained on the full lexicon of 2,285 patterns

5 Results

5.1 Test Set

To discover whether the network was able to generalise to novel DET+NOUN combinations, a test set was constructed. The 16 most frequent nouns were extracted from the lexicon and presented to the network in DET+NOUN combinations not

present in the training set:

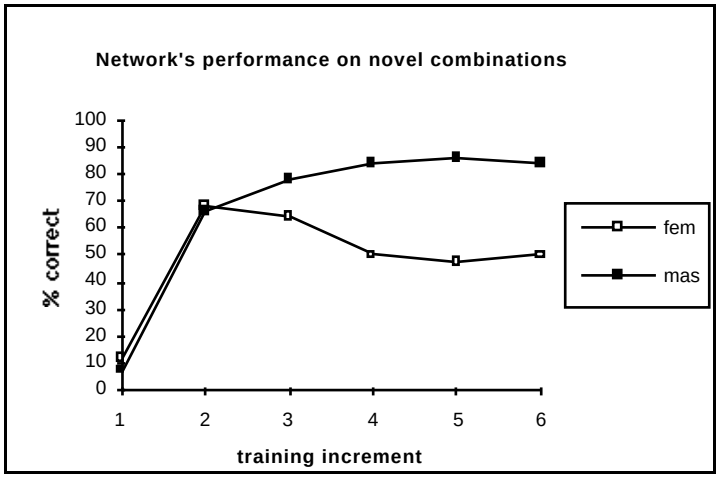
16 most frequent nouns used in the testing

Feminine		Masculine	
Regular	**Irregular**	**Regular**	**Irregular**
cosa	mano	cuento	nene
niña	calle	perro	pie
vaca	vez	beso	día
caca	leche	culo	coche

This resulted in a 98 pattern test set consisting of the various unseen DET+NOUN combinations. The network's performance on the test patterns was calculated at the six points along the incremental training regime where the weights had been saved.

5.2 Overall Results

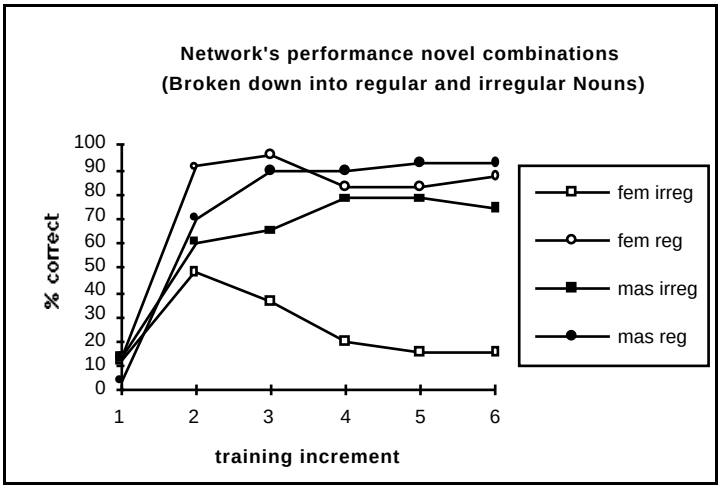
The results show that the network can learn to produce the correct determiner in unseen, novel, combinations. Although initially, feminine DETs are produced with a slightly higher degree of success, performance on masculine DETs soon overtakes:



After the 2nd increment, performance on the feminine nouns falls before settling at a level of about 55% at increments 4, 5 and 6.

5.3 Regular vs. Irregular

The next graph shows the same information broken down into regular and irregular items.



Performance on regular feminine determiners peaks at almost 100% at increment 3 only to be overtaken by masculine regulars at increment 4. Despite this, performance is still remarkably strong - never falling below 80% after its peak. It is clear from the graph that the poor overall performance on feminine determiners is largely due to the irregular items. Many of the irregular feminine DET+NOUN combinations presented to the trained network resulted in masculine DETs being produced at the output layer.

6 Discussion

The results of this experiment show that a connectionist network can be trained using child directed speech, to produce Spanish determiners. The successful performance on masculine noun-determiner pairs led us to analyse the type/token frequencies in the training set (see Figure 2).

It seems that the higher proportion of irregular masculine types and tokens is responsible for this pattern of results. Irregular nouns are more frequently presented to the child with a masculine determiner. Thus, when presented with the task of producing a determiner to accompany an ambiguous noun the child is more likely to produce a masculine one. Given that this training data are taken from speech directed to a young child, it suggests that the different frequencies in child-directed speech in Spanish, hitherto unnoticed, may account for the way in which the child learns the masculine forms at the same age as the feminine forms despite the differences in regularity.
This supports a theory that the acquisition of noun phrase morphology in Spanish may be a largely data-driven process owing much to the type and token frequencies of child-directed speech.

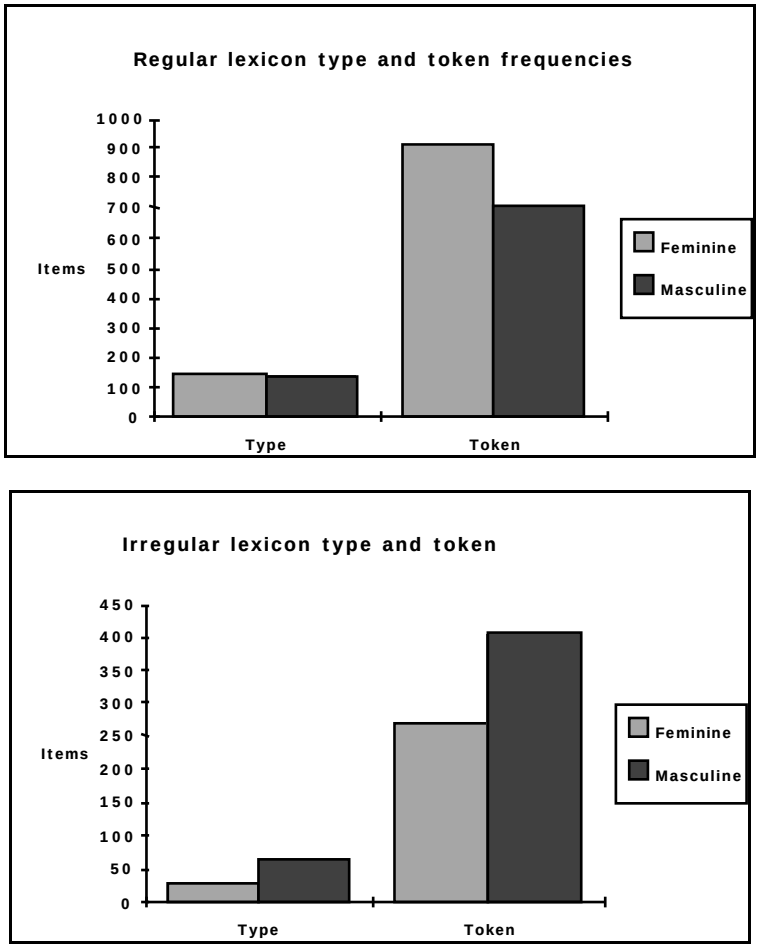


Figure 2: Type and Token Frequencies for both regular and irregular items in the training set

References

[1] López-Ornat, S. (1994). *La adquisición de la lengua española.* Madrid: Siglo XXI.

[2] Nix, A. J., (1997). A connectionist enquiry into the production of Spanish noun phrases. Ph.D. thesis. University of Hertfordshire.

[3] Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language,* 18(3): 571-590.

[4] Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition,* 48: 21-69.