Learning the Arabic Plural: The Case for Minority Default Mappings in Connectionist Networks.

Neil Forrester
Department of Experimental Psychology
University of Oxford
South Parks Road
Oxford OX1 3UD
UK
Tel: 0865 - 271397
naf@psy.ox.ac.uk

Kim Plunkett
Department of Experimental Psychology
University of Oxford
South Parks Road
Oxford OX1 3UD
UK
Tel: 0865 - 271398
plunkett@psy.ox.ac.uk

Abstract

Connectionist accounts of inflectional morphology have focussed on the domain of the English Past Tense (e.g. Rumelhart & McClelland 1986; Plunkett & Marchman 1993). In this inflectional domain, the default mapping process (add /ed/) reflects the process of suffixation adopted by the majority of the forms in the language. Connectionist models exploit the imbalance between English regular and irregular verbs when learning the past tense and when responding to novel forms in a default fashion. Not all inflectional systems have a default mapping which is characterized by a majority of forms in the language. The Arabic Plural System has been cited (Marcus et al. 1993) as one such system where a minority default mapping process operates. The Sound Plural in Arabic applies to only a minority of forms in the lexicon (~10%), yet it appears to adopt the role of a default mapping for novel nouns. We describe a connectionist model that can learn a minority default mapping analogous to the Arabic plural and discuss its performance in relation to type and token frequency effects, and their distribution within phonetic space.

INTRODUCTION:

A competent language user can apply both regular inflectional patterns (talk - talked) and irregular patterns (sing sang) to novel words productively (wug - wugged; ling - lang). Traditional generative accounts (Pinker & Prince 1988) attribute the development of such abilities to the application of rules. For example in English, we may posit a memory storage device that contains high frequency forms and irregular forms in the language. In addition, a rule based system appends the appropriate allomorph of /ed/ to the stem of the verb to form the past tense. Errors during language acquisition are explained by the interference between the two mechanisms. Specifically, the memory storage device fails to block the application of the regular rule to an irregular stem - leading to overgeneralisation. Mature performance is achieved by fine-tuning the conflict between the mechanisms: The memory trace for the irregular verbs is strengthened and the blocking effect thereby enhanced. In contrast, connectionist accounts (Rumelhart & McClelland 1986) postulate a single mechanism which performs the task of learning both regular and irregular forms of the past tense. Similarities between the phonology of novel forms and the phonology of forms in the training corpus, dictate the network's ability to generalise to new forms in the language, while still performing correctly on the irregular forms.

A critical feature of the English past tense is that not all verbs undergo the same type of transformation from stem to past tense form. Irregular verbs undergo a variety of changes, e.g., arbitrary change, no change, vowel suppletion, blending. A connectionist model must learn all these types without corrupting the 'default' mapping of the regular transformation (i.e. add /ed/). It has been argued (Pinker & Prince 1988, Marcus et al 1992) that the ability of connectionist accounts to learn this task is an artifact of the statistical composition of the English past tense system. The 'default' mapping is used by the majority of the forms in the language, hence the connectionist model simply reflects this asymmetric distribution of regular verbs when it produces overgeneralisations and responds to novel forms. More recent modelling of the acquisition of the English past tense (Plunkett & Marchman 1993) has investigated the role of changes in the irregular/regular ratio for profiles of learning in the early stages of acquisitions. In this paper, we investigate further the effects of the vocabulary balance between different word types for the acquisition of inflectional systems.

Not all inflectional systems are like the English Past Tense. In the Arabic plural there is no single inflectional type which applies to a majority of the forms in the language. The majority of plurals are characterised by a system of sub-regularities conditioned by the phonological characteristics of the noun stem - the so called Arabic Broken Plural (Murtonen 1964). The phonological shape of the noun in Arabic provides a reliable cue to the formation of its plural. The broken plural is formed by internal vowel changes, sometimes with the addition of prefixes and suffixes (e.g. malikun - mulukun). A small minority of nouns (~10%) take the Sound Plural, whereby a suffix (differing for gender) is added to the stem, much like adding an /s/ in English (e.g. hasanun - hasanuna). These nouns have few phonological characteristics that make them cohere as a

class and do not possess the phonological template associated with the broken plural types. The sound plural is used with novel forms that do not possess a broken plural template and also loan words from other languages. It is assumed that children learning Arabic overgeneralise with the Sound Plural, but there is currently no experimental evidence available on this issue.

The standard generative account provides a natural explanation of the behaviour of the default Sound Plural in Arabic: A symbolic system automatically appends a suffix to the noun stem when an exceptional form is not recognised and blocking goes untriggered. This rule-governed process does not depend upon a majority of forms conforming to the default process in the language itself. In contrast, a connectionist account would appear to be ill-designed for the task of extracting a default mapping process which is characterised by a minority of forms in the language: Suppose the Arabic nouns were distributed evenly across the phonological space of the language, it is more likely that a novel word would occupy a space closer to a broken plural type than to a sound plural, since broken plurals are in the majority. One would therefore expect any inflectional transformation based upon judgements of similarity (such as those made by connectionist models) to reflect the gross statistical distribution in the language. Thus, the default process should reflect the characteristics of the broken plural.

It is advantageous here to visualise language in spatial terms. Consider the phonological space of a language as projected onto two dimensions, bounded by the limits of the phonological rules which govern the particular language. In this space, words that are phonologically similar would occupy similar positions. Recall that broken plurals fit well-defined phonological templates. Hence, the system of broken plurals in Arabic would constitute 'islands' in a 'phonological sea' where the size of these islands is determined by the degree of phonological variation characterising the class. For example, in English, the sub-regularity exemplified by /sleep/->/slept/, /keep/->/kept/, /creep/->/crept/ also includes tokens such as /meet/->/met/ that are phonologically somewhat distant from the prototype for that group. In contrast to the broken plurals, the Sound Plurals have few cohering features and may be visualised as tokens on their own (or in small localised groups of a few tokens), dotted in the 'sea' between the groups of broken plural types. When the Arabic plural system is visualised in this manner it is easy to see why a novel stem is much more likely to fall in the 'sea' rather than on an island. Hence if the network can learn to associate the 'sea' to the default mapping it will generalise correctly.

We will demonstrate that it is not necessary for this 'sea' to be densely populated for the network to learn the mapping it comprises. Rather, it is the distribution of the tokens within the candidate default space which determines the generalisation abilities of the trained network. We describe a connectionist model that learns a minority default mapping analogous to the Arabic Broken Plural. In our first experiment we investigated how distribution affects the classification of plural types in a pseudo-Arabic plural system. In experiment 2 we implement a fuller model of the inflectional process

EXPERIMENT 1

A Connectionist model of the classification of Plural Types in Pseudo-Arabic.

A simple feedforward network with two input units, fifty hidden units and three output units was constructed and trained using the back-propagation learning algorithm (learning rate 0.15, momentum 0). The training sets consisted of real valued co-ordinates corresponding to points on a two dimensional grid. Each point in the training set was designated as belonging to one of three classes. The classes were two 'broken-plural' groups occupying two discrete areas in the space and one 'default-mapping' group sparsely distributed across the remaining space. The ratio between broken plural and sound plural 'words' mirrors the ratio found in Arabic (approximately 90% broken plurals, 10% sound plurals). Membership in a class was signalled in the network by unambiguous activation of just the appropriate output unit. Two training sets were devised:

Training set #1: (Fig 1.1) 2 broken groups containing 146 and 144 tokens respectively, within a well defined areas. 29 sound plurals spaced randomly across the remaining space (group C). All points had a token frequency of 1. This distribution is used as a baseline to investigate how a network will generalise to a complex space.

Training set #2: (Fig 1.2) 2 broken groups of 149 and 144 tokens each, with a uniform token frequency of 1. 30 sound plurals spaced randomly across the remaining space, however with a different distribution to training set #1. This distribution is used to investigate the effect of distribution of training tokens on the generalisation abilities of the trained network.

Test Set: 441 points spanning the entire space, allowing the classification zones of each of the three 'word' types to be determined.

RESULTS & DISCUSSION

The outcome of training the network on set #1 summarised in Fig 2.1. The figure plots the response of the three output units to all the patterns in the test set in three different points in training (500, 1000, 10000 epochs). Dark areas indicate high output activity. It can be seen that the network has generalised to the space for the two groups of 'broken plurals'. However, the network has also genera-

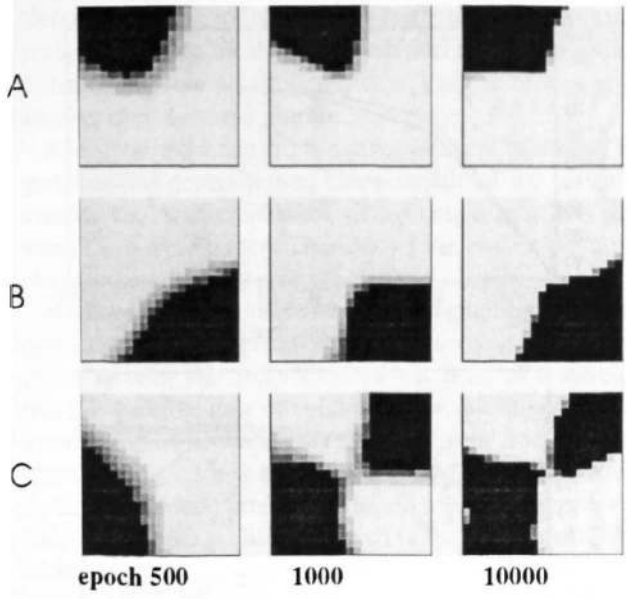

fig. 2.1 outputs for training set #1

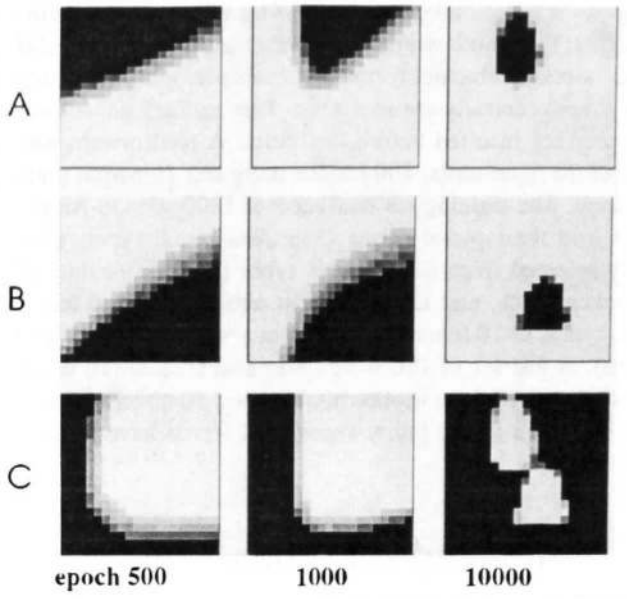

fig. 2.2 outputs for training set #2

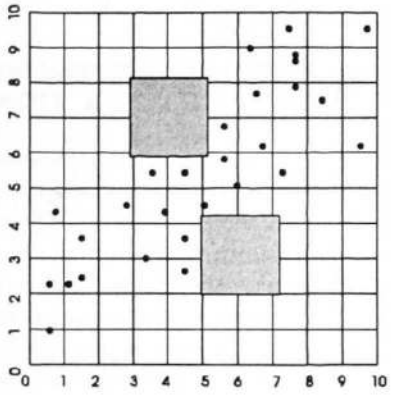

Fig 1.1 distribution of training set #1

Figure 1. Results of Experiment 1.

Results are given for each of the three plural types (2 broken plural groups, A & B, 1 sound plural group, C) Responses are shown for three points in training for each simulation, epoch 500, 1000, and 10000. Darker areas show higher responses

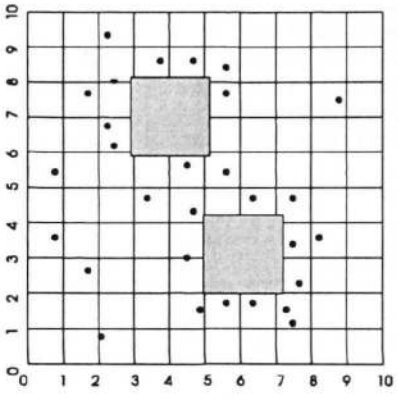

Fig 1.2 distribution of training set #2

lised the zones of space which were unpopulated by training exemplars. Consequently, these areas have not been treated in a minority default fashion. In fact, the minority default area has been restricted to just the zones populated by sound plurals in the training set. To improve performance a more representative distribution must be used.

The distribution of the 'sound plurals' in training set #2 (fig. 1.2) forces the network to generalise this plural type as the 'default' successfully. In particular, note that the peripheral zones of space are almost entirely generalised to the sound plural minority default, even though the sound plural training exemplars populate the 'phonological sea' as sparsely as in the first training set.

The distribution of the patterns is critical to ensure that good generalisation takes place. It must be remembered that the back propagation algorithm merely finds a solution whereby the correct transformations occur for the training data. Therefore a network cannot necessarily be relied upon to generalise to an area of data space if such a space is not defined appropriately in the training set. An area need not be densely populated for successful generalisation to take place: a representative spread of low token frequency forms is all that is required. Hence we have demonstrated that the problem of minority default mapping is solvable in connectionist models, providing care is taken to use a *representative* distributions.

WORK IN PROGRESS - LEARNING THE ARABIC BROKEN PLURAL

A second set of simulations was designed to implement the inflectional process of the Arabic plural directly in a network that takes a phonological representation of the noun stem and maps it to a phonological representation of the corresponding plural form. The words are represented using a phonetic coding scheme adapted from Plunkett & Marchman(1991) using 7 bits per phoneme rather than 6, the extra unit for 'pharengealised' to account for the guttural sounds in Arabic which are not present in English. The phonemic segments of the input were concatenated

using a scheme adapted from MacWhinney & Leinbach(1991) in which words are represented in consonant and vowel 'slots' to channel forms of variable length. The words were fixed centrally around their first radical consonant, with prefixes inserted before this point. A feedforward network of 70 input units, 100 hidden units and 71 output units was used. The training set consisted of 1300 pseudo-Arabic nouns and their plural forms (5 broken plural types, arbitrarily selected from the 40 or so types found in Arabic, of 234 tokens each, and 13 of the most frequently found sound plural types, of 10 tokens each, giving a 'default' area of 130 tokens). A test set of 180 words was also created (10 novel forms of each of the 5 broken plural types, 10 novel forms of each of the 13 sound plural types). All words have a token frequency of 1.

Training set 1300 words

Broken Plurals	Sound Plurals	
CaCaCun => 'aCCACun	CaCiC	MuCTaCiC
CaCiCun => CuCUCun	muCaCCiC	muCCiC
CaCCun => CiCACun	muCACiC	muCCaCC
CiCACun => CuCuCun	muTaCaCCiC	muSTaCCiC
CaCCun => 'aCCuCun	muTaCaCiC	CaCiC
	muNCaCiC	CaCCAN
		CaCCAC

NB 'C' denotes any consonant

TEST SET: 180 words

10 novel forms of each of the 5 broken types
10 novel forms of each of the 13 sound plural types

```
CVCVCVCCVC -> CVCVCVCCVC+ending
              e.g.
--CaCaC-un -> -aCCAC--un
muTaCaCCiC -> muTaCaCCiC+ending
```

Word Coding scheme
(adapted from MacWhinney & Leinbach(1991))

The broken plurals were generated around fixed vowel templates taken from five broken types in the language. The sound plurals were created from templates taken from the 13 most frequently used sound plurals in Arabic. The network was initially trained for 50 epochs, with a learning rate of 0.1 and a momentum of 0.1.

Network output is analysed by determining the closest legal phoneme in Euclidean space for all output phonemes and then comparing each phoneme 'slot' with its target.

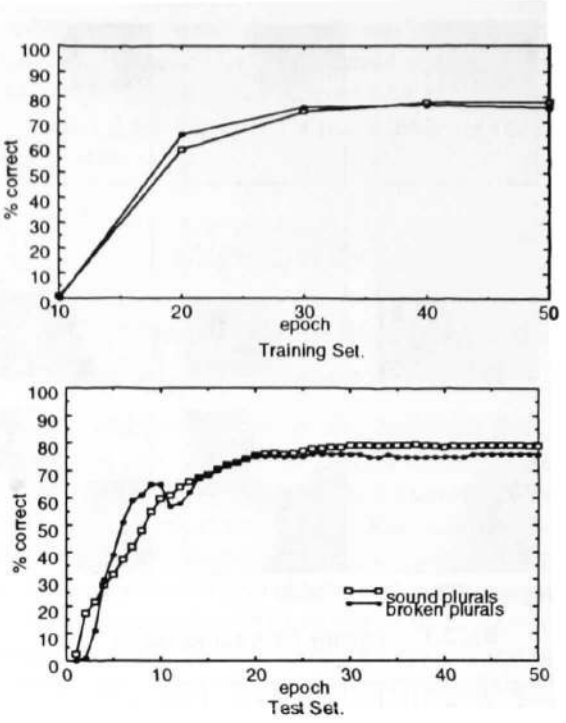

fig 3. Learning the Arabic Broken Plural - Performance of Network

RESULTS AND FUTURE DIRECTIONS

With current parameter settings the network successfully learns the mappings between singular and plural forms for 75-80% of the words in the training set. Each of the five broken plural groups are readily extracted from the training set and generalisation to novel broken plural forms is good (~75%). Mapping transformations from noun stem to plural form for novel sound plurals is also successful (~80%). Fig 3. plots the networks performance over the The majority of errors produced by the network in response to novel forms are errors on a single bit of the output vector, usually resulting in a slip from one vowel or consonant to a nearby phonological neighbour (e.g. b->d, a->A).

Although network performance and generalisation is good, it is clear that it could be improved. In these simulations we have not attempted to manipulate the token frequency of the training forms. From previous work (Plunkett & Marchman (1991)), we know that token frequency characteristics are crucial in determining overall performance and generalisation in problems of this type.Additional information about token frequency values for individual nouns in Arabic will permit us to evaluate whether token frequency effects are crucial for this language as well.

It is worth noting that the distribution of the sound 'default' group is not distributed across the entire phonological space. In the pseudo-Arabic model, sound plural forms were distributed widely across phonetic space. In real Arabic this is not in fact the case. The distribution is not simply a matter of broken groups and a homogenous

'default' sea. The sound plural is made up from many small groups displaying local similarity. Novel words which do not demonstrate a strong similarity to any of the broken groups are classified as sound plurals.

Anecdotal evidence from native speakers of Arabic suggest that this generalisation characteristic of the network is correct, i.e. Arabic speakers will produce a broken plural when the novel form matches one of the broken plural templates, otherwise they produce a sound plural.

We have not considered the developmental implications of this model for different stages in the learning of the Arabic plural system. Our main concern has been to demonstrate that it is possible to learn a minority default mapping within a connectionist network. However, it is clear from the results shown in Fig.2.2 that broken plural and sound plural forms have characteristic profiles of development. It remains to be seen whether the profiles observed in the model match those observed in young Arabic children. Empirical work with young Arabic language learners is projected for the near future.

Acknowledgments

We are grateful to MSC Thomas for his invaluable input, Denis Mareschal, Chrys Meula and Sharon McHale.The work in this paper is supported in part by an SERC grant.

REFERENCES

MacWhinney, B. & Leinbach, A. J.(1991) Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40, 121-157

Marcus, G. F., Brinkman, U., Clahsen, H., Weise, R., Woest, A. & Pinker, S. (1993) German inflection: The exception that proves the rule. *Occasional Paper #47, Center for Cognitive Science, MIT.*

Murtonen, A. (1964) *Broken Plurals, Origin and Development of the System.* Netherlands: E.J. Brill.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28, 73-193.

Plunkett, K. & Marchman, V. (1991) U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 38, 43-102

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

Rumelhart, D.E., & McClelland, J.L.(1986). On learning the past tense of English verbs. In J.L. McClelland, D.E. Rumelhart, & P.R. Group(Eds.), *Parallel distributed processing: Explorations in the Microstructure of Cognition, Vol.2: Psychological and Biological Models* (pp. 216-271). Cambridge, MA: MIT PRESS.