

Input driven constraints on plurals in English noun-noun compounds

Jenny Hayes^{1,2} (J.Hayes@herts.ac.uk)
Victoria Murphy¹ (V.A. Murphy@herts.ac.uk)
Neil Davey² (N.Davey@herts.ac.uk)
Pamela Smith¹ (P.M. Smith@herts.ac.uk)

Departments of Psychology¹ and Computer Science², University of Hertfordshire, College Lane, Hatfield, AL10 9AB, United Kingdom.

Abstract

Native English speakers include irregular plurals in English noun-noun compounds (e.g. *mice chaser*) more frequently than regular plurals (e.g. **rats chaser*) (Gordon, 1985). This dissociation in inflectional morphology has been argued to stem from an internal and innate morphological constraint as it is thought that the input to which English speaking children are exposed is insufficient to signal that regular plurals are prohibited in compounds but irregulars might be allowed (Marcus, Brinkmann, Clahsen, Wiese & Pinker, 1995). In addition, this dissociation in English compounds has been invoked to support the idea that regular and irregular morphology are mediated by separate cognitive systems (Pinker, 1999). The evidence of the neural network model presented here is used to support an alternative view that the constraint on English compounds can be derived from the general frequencies and patterns in which the two types of plural (regular and irregular) in conjunction with the possessive morpheme occur in the input.

Introduction

The Compounding Phenomenon

Psycholinguistic research has shown that English compound words with irregular plural nouns in first position (e.g. *mice-eater*) are produced far more frequently than compound words with regular plural nouns in first position (e.g. **rats-eater*), (Gordon, 1985).

Explanation according to the Dual Mechanism model

The dual mechanism model (Pinker, 1999), proposes that irregular nouns and their plurals are stored as memorised pairs of words in the mental lexicon (e.g. *mouse-mice*) but that regular plurals are produced by the addition of the [-s] morpheme to the regular stem at a post lexical stage (e.g. *rat + s = rats*). Compounds are created in the lexicon. Thus as irregular plurals are stored in the lexicon they are available to be included within compound words. However, as only the singular stems of regular nouns are stored in the lexicon the plural form is never

available to be included within compound words (Marcus et al, 1995).

A Single Route Associative Memory Based Explanation of Compounding

An alternative explanation of the treatment of plural morphology in English compounds is that the dissociation between irregular and regular morphology in compounds can be explained by examining the nature and frequency of regular, irregular and possessive English morphology in the input (Hayes, Murphy, Davey, Smith & Peters, 2002). Frequency counts of a sample of the CHILDES (Child Language Data Exchange System) corpora (McWhinney & Snow, 1985) have shown that the plural [-s] morpheme is never followed by a second noun. Importantly, a different pattern is found with the possessive [-'s] morpheme since it is always followed by a second noun. Therefore, it might be that a noun rarely follows the regular plural [-s] morpheme (i.e. patterns such as “*rat[s] chaser” do not occur) because the pattern “*noun – morpheme [-s]- noun*” is reserved for marking possession (such as *rat’s tail*). Interestingly in other languages that do not have this competition between the plural and possessive morpheme such as Dutch (Schreuder, Neijt, van der Weide & Baayen, 1998) and French (Murphy, 2000), regular plurals are allowed within compounds. Irregular plurals may, however, appear in English compounds as they are not formed by the addition of the plural [-s] morpheme. Thus, irregulars do not compete with the possessive structure and as such may be followed by a second noun in a compound. The competitive relationship between the plural and possessive [-s] morpheme is examined here using a connectionist model trained on child directed speech taken from CHILDES.

Neural Net Modeling

Introduction

An associative memory-based account of inflectional morphology has been investigated in numerous connectionist models. Several models have

successfully simulated the putative dissociation between regular and irregular inflection for both verbal morphology (Daugherty & Seidenberg, 1994) and plural morphology (Plunkett & Juola, 1999) using a single learning mechanism and no explicit rules. Furthermore, as well as being able to learn mappings from input to output, connectionist models have also been able to learn sequential mappings (Elman 1990). Thus it is predicted that a single route associative memory system could learn that the inclusion or omission of the regular plural morpheme [-s] is influenced by where that [-s] morpheme occurs in a sequence of language input.

Experimental aims

The aim of this simulation was to investigate whether the fact that the possessive [-s] morpheme is always followed by a second noun but the plural [-s] morpheme is rarely followed by a second noun is sufficient to constrain compound formation in English. A simple recurrent network (SRN) was utilised so that at any point in time the state of the hidden units at the previous time step were used as additional input (Elman, 1990). Thus it was expected that the model would be able to learn sequential mappings. The network was trained on a large training set of real child directed speech in which the frequencies with which the various types of morphology occurred were not manipulated in any way. The syntactic type of each word was used as the input to the network. The network was required to predict the syntactic type of the next input it expected to receive given the syntax of the previous inputs. In previous modeling work (Hayes et al, 2002), it has been shown that an SRN is capable of learning about grammatical type from a training set in which each word token was encoded using a localist coding

scheme. In these earlier models items in the training set were not explicitly coded as being representatives of a particular syntactic type (e.g. as being nouns or verbs). Instead, learning about the distinct linguistic functions that the different syntactic types perform emerged during training. However, a disadvantage of these models was that it was only possible to use a small lexicon of words because of the complexity of the learning task. The model reported here was trained on a much larger training set than our previous models. This simulation sought to reproduce the behaviour of an older child, with a much larger vocabulary, who has knowledge, though perhaps not at a metalinguistic level, of the different functions that are performed by the different syntactic types.

The frequency in which regular and irregular plurals and possessives were included in the training set was determined by the frequency in which they appeared in the child directed speech that was used as the input to the model. Table 1 illustrates that some items appear in sequence with other items in the input (e.g. possessives are always followed by singular nouns) but other items do not appear in sequence with other items (e.g. regular plurals are not followed by singular nouns). The performance of the network was investigated using a syntactic type prediction task in which one of three syntactic types was input (a possessive, a regular plural or an irregular plural) and the network predicted which syntactic type it expected to see next in the input stream. The difference (error) between this predicted output and the output for noun, verb, other and word ending was calculated. It was predicted that the error would be high for all items after possessives except nouns. Conversely it was predicted that there would be a high error on predicting a noun after a plural of either kind.

Table 1. Frequency with which items from various syntactic categories followed irregular plurals, regular plurals and possessives (percentage frequency shown in brackets) in the training set.

Item following plural or possessive	Irregular plurals (n=9)	Regular plurals (n=95)	Possessives (n=39)
Others	3 (33)	40 (42)	0
Sentence ending marker	0	30 (32)	0
Singular nouns	2 (22)	0	39 (100)
Verbs	1 (11)	24 (25)	0
Regular plurals	1 (11)	1 (1)	0
Irregular plurals	2 (22)	0	0

Training set and coding scheme

The exact composition of the training set is shown in Table 2. Irregular and regular plurals and possessives

form less than 1% of the input. Items coded as “others” included anything that was not a noun or a verb (e.g. adjectives, determiners, adverbs and prepositions). 2182 sentences, made up from 9999 words, from the

Wells study from the CHILDES corpora (MacWhinney & Snow, 1985) were concatenated and used as input. A sentence ending marker was also included in the training set. The frequency with which items from various syntactic categories followed irregular plurals, regular plurals and possessives is shown in Table 1.

Table 2 Composition of training set

Item	Number of tokens in training set	Cumulative total	Percentage of tokens in training set	Cumulative percentage
Irregular plurals	9	9	0.09	0.09
Possessives	39	48	0.39	0.48
Regular plurals	95	143	0.95	1.43
verbs	624	767	6.24	7.67
Sentence ending markers	1415	2182	14	22
Singular nouns	3014	5196	30	52
others	4803	9999	48	100

Possessives were only ever followed by singular nouns in the input. Regular and irregular plurals were followed by a range of items but never by a singular noun. Each item was encoded using a 7 bit vector. Three input units encoded syntactic category (noun, verb, other) and two inputs encoded whether the item was plural or not. Two input units encoded the presence

or absence of the [s] morpheme. Thus for both regular plurals and possessives the input units for noun and [s] morpheme present would both be activated? A possessive was disambiguated from a regular plural, however, because the plural input unit was “yes” for a plural but “no” for a possessive. Examples of how items from different syntactic categories were encoded is shown in Table 3.

Table 3. Examples of coding scheme

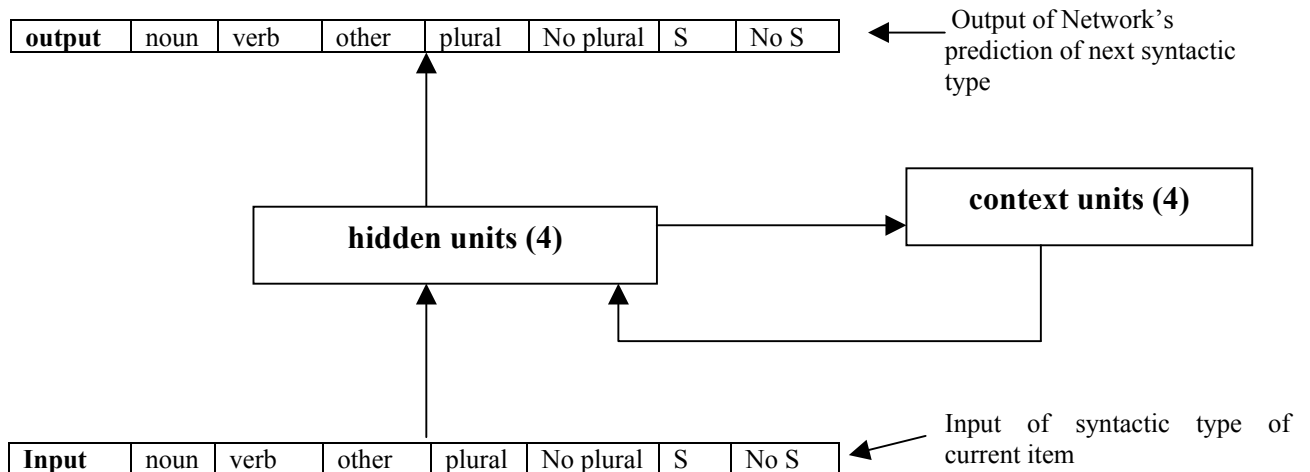
Item	Syntactic category			Type of noun		S present or not	
	noun	verb	other	singular	plural	S	No S
rats	YES	NO	NO	NO	YES	YES	NO
mice	YES	NO	NO	NO	YES	NO	YES
rat's	YES	NO	NO	YES	NO	YES	NO
chaser	YES	NO	NO	YES	NO	NO	YES
the	NO	NO	YES	NO	NO	NO	NO
chase	NO	YES	NO	NO	NO	NO	NO
Word ending marker	NO	NO	NO	NO	NO	NO	NO

Architecture

The architecture of the network is shown in Figure 1. The network had 7 input units, 4 hidden units, 7 output units and 4 context units. A simple recurrent architecture was adopted so that at any point in time the

state of the hidden units at the previous time step were used as additional input (Elman, 1990). The SRN was trained using a learning rate of 0.1 and a momentum of 0.3.

Figure 1. Network architecture



Task

In both the training and test phases, the network was required to predict the next input. (i.e. the target output was one time step behind in the input).

Training

The network was trained on 10,000 repetitions of the training set. This high number of presentations of the input was necessary because the training set was large and items of particular interest i.e. possessives (0.39% of the input), regular plurals (0.95% of the input) and irregular plurals (0.09% of the input) formed such a low proportion of the input.

Test Phase

After training, the network was presented with the following sequences:

- *possessive followed by singular noun*
- *possessive followed by verb*
- *possessive followed by other*
- *possessive followed by word ending.*
- *Regular plural followed by singular noun*
- *regular plural followed by verb*
- *regular plural followed by other*
- *regular plural followed by word ending.*
- *Irregular plural followed by singular noun*
- *irregular plural followed by verb*
- *irregular plural followed by other*
- *irregular plural followed by word ending.*

Thus, for example, in the test pattern *possessive followed by singular noun* the code for *possessive noun* was input and the target output was *singular noun*. However the network might not output *singular noun*. The actual output and the target output were compared and an error figure was calculated based on the difference between the two output weight values.

Results

The error on producing the target output was recorded after the network was presented with the test

sequences. Many runs of the simulation were carried out but each produced almost identical results.

Figure 2. Error on producing nouns, verbs, other items and word endings after possessives, regular plurals and irregular plurals

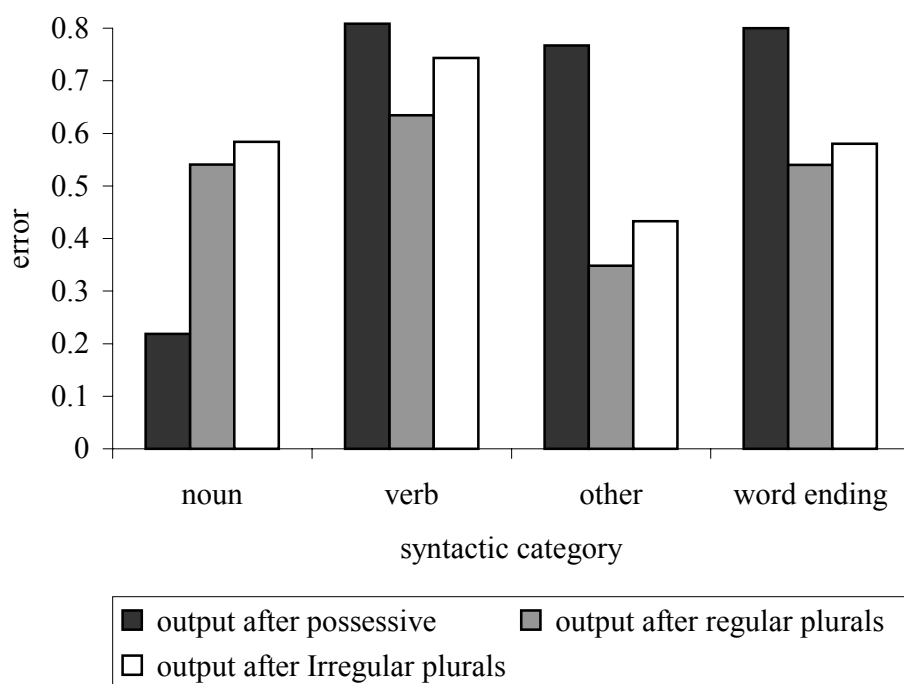


Figure 2 illustrates that at a descriptive level the error on producing a singular noun after a possessive was about half as high as the error on producing a singular noun after a plural of either type¹. The network also learnt that the syntactic categories that make up other items and sentence-ending markers can follow plurals but not possessives. The network, produced a high rate of error when the target output after a plural noun was a verb, despite the fact that in the input verbs followed regular plurals (25% of the time that regular plurals occurred) and irregular plurals (11% of the time that irregular plurals occurred). However, the training

et contained very few verbs (6.24 % of the training set). Given that verbs were so underrepresented in the input it was unlikely that they would be predicted as the next item in a next word prediction task to any great extent.

Discussion

This neural network was trained using naturalistic child directed speech. Gaining this advantage, however, meant that the syntactic type of each token rather than individual tokens were used as input to the network.

¹ It was not possible to carry out a statistical test on the error rates shown in Figure 2 as the figures shown relate to the output of 1 test rather than to the output of several tests.

This means that syntactic type did not emerge during training (see Hayes et al, 2000 for a discussion of how syntactic type did emerge in a neural net model trained on a smaller training set). However, this model offers an insight into how learning might take place when the frequencies of items in the input are more accurately represented.

The syntactic category prediction task showed that the error on producing a singular noun after a plural, of either kind, was twice as high as the error on producing a singular noun after a possessive. This suggests that the network easily learned the sequence *possessive [-'s] – noun*. The network also learnt that the syntactic categories that make up other items and sentence-ending markers can follow plurals but not possessives. This learning seems to have occurred because these items appeared in consistent patterns in the input despite having very low frequencies in the messy context of child directed speech.

Regular plurals and possessives were disambiguated in the input by the fact that the plural input unit was on in the case of a regular plural but off in the case of a possessive and from the patterns in which they occurred in the input. Thus it seems that the network was able to learn that the *noun –morpheme [-s]* pattern occurred in different patterns when it was plural to when it was singular. Some items follow one pattern (i.e. a second noun follows the *noun [-s] morpheme* pattern when it is singular but not when it is plural) while other items follow the reverse pattern (i.e. word ending markers and other items follow the pattern *noun- [-s] morpheme* when it is plural but not singular). That a neural network model with no explicit grammatical structure was able to learn these linguistic patterns is further support for the idea that there is sufficient evidence in the input to constrain learning that a second noun is not included after a plural because the pattern *noun-morpheme [-s]- noun* is used to denote possession not plurality.

General Discussion

In the neural network model described here, the syntactic type of each word token was explicitly encoded in the input. Real child directed speech was used as input and thus the different syntactic types were represented by the actual frequency in which they occurred in child directed speech. Under these realistic input conditions it is clear that a neural network model was able to learn that the *noun –morpheme [-s]* pattern occurs in different patterns when it the noun is plural to when it is singular. The network learned to predict that nouns follow possessives but not plurals of either type and also learned to predict that “other items” and word ending markers followed plurals of either type but not possessives.

The results of this neural network model together with previous models (Hayes et al, 2000) provide some insight into how an input driven constraint on compound formation might develop in the human language learner. Both the possessive [-s] and the plural [-s] sound the same phonetically but the child may learn from the syntactic patterns in which the two different types of morpheme appear in the input that one type of morpheme is appropriate in some sentential contexts but not in others. It might be that the child learns that the possessive morpheme can be followed by a noun so when forming compound words it is not appropriate to follow the plural morpheme with another noun. When faced with a noun-noun compound the language user may delete the plural morpheme from the end of the first noun, not because regular items of morphology are different in kind from irregulars and represented as “rules” in the brain, but simply because this pattern is used to denote possession not plurality in English.

References

- Daugherty, K. G. & Seidenberg, M. S. (1994). Beyond rules and exceptions. In S. D. Lima, R. L. Corrigan & G. K. Iverson (Eds.), *The reality of linguistic rules*. Amsterdam: John Benjamins.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Gordon, P. (1985). Level-ordering in lexical development. *Cognition*, 21, 73-93.
- Hayes, J.A., Murphy, V.A., Davey, N., Smith, P.M. and Peters, L., (2002). The /s/ morpheme and the compounding phenomenon in English. In W.D.Gray & C.D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.
- MacWhinney, B. & Snow, C. E. (1985). The Child Language Data Exchange System. *Journal of Child Language*, 12, 271-296.
- Marcus, G. F., Brinkmann, U., Clahsen, H., Weise, R. & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, 29, 189-256.
- Murphy, V. A. (2000). Compounding and the representation of inflectional morphology. *Language Learning*, 50, 153-197.
- Pinker, S. & Prince, A. (1992). Regular and irregular morphology and the psychological status of rules of grammar. In L. A. Sutton, C. Johnson, & R. Shields (Eds.), *Proceedings of the 17th Berkeley Linguistics Society* (pp.230-251).
- Plunkett, K. & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23, 463-490.
- Schreuder, R., Neijt, A., van der Weide, F., & Baayen, R. H. (1998). Regular plurals in Dutch compounds: Linking graphemes or morphemes? *Language and Cognitive Processes*, 13, 551-573.