A Connectionist Model of Lexical and Contextual Influences on Ambiguity Resolution in Human Sentence Processing

Frank Keller
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, UK
keller@cogsci.ed.ac.uk

Klaus Zechner
Department of Philosophy
Carnegie Mellon University
135 Baker Hall
Pittsburgh, PA 15213-3890, USA
zechner@andrew.cmu.edu

Abstract

The lexicalist model of human sentence processing (MacDonald *et al.* 1994) provides an account for the interaction of lexical frequency effects with contextual information in the resolution of syntactic ambiguities.

In this paper, we present an implementation of a connectionist network which evaluates the predictions of the lexicalist model for NP/S garden paths. Our network is trained using a corpus tagged for argument structure and contextual information. It exhibits processing preferences which are interestingly correlated with argument structure frequency, but is also sensitive to information from the syntactic and semantic context.

1 Introduction

1.1 The Garden Path Model

The dominant view on human sentence processing is the garden path model (Frazier and Rayner 1982; Frazier 1989), which considers parsing as the incremental construction of phrase structures. In this model, it is assumed that parsing is guided by general principles which make reference to the structural complexity of syntax trees.

Evidence for the garden path model comes from examples as (1), which contain an MV/RR ambiguity, i.e., the verb *raced* can be interpreted either as a main verb (MV) or as a verb belonging to a reduced relative clause (RR). Readers experience processing difficulties with the RR reading like in (1b). It is assumed that the human parser has to perform a re-analysis (garden path effect).

(1) a. The horse raced past the barn.
b. The horse raced past the barn fell.

To explain when exactly this re-analysis takes place, the garden path model postulates that the human parser incrementally builds up a phrase structural representation for incoming material, and that this process is guided by principles which yield a single structure also for ambiguous input. The key principle involved is Minimal Attachment (MA), which requires that minimal phrase structure trees are constructed. Due to MA, the MV reading is generally preferred over the RR one, and hence the parser has to perform backtracking for examples like (1b). The crucial prediction of the garden path model is that parsing preferences are determined by syntactic factors only.

Apart from the MV/RR examples, several other cases of ambiguous input are discussed in the literature (for a review cf. MacDonald *et al.* 1994), including the noun phrase/sentential complement (NP/S) ambiguity:

(2) a. John knew the answer very well.
b. John knew the answer was right.

For NP/S examples as in (2), the garden path model predicts that the syntactically simpler NP reading is preferred.

1.2 The Lexicalist Model

Recent studies show that human parsing preferences are not really as independent of non-syntactic factors as the garden path model predicts: It was found that the resolution of MV/RR ambiguities is influenced by the frequency of the of the ambiguous verb occuring with a certain argument structure (MacDonald 1994) and by the frequency of past tense versus past participle occurences of the ambiguous verbs (Burgess and Hollbach 1988). Holmes *et al.* (1989) produced evidence for argument structure frequency effects with NP/S stimuli.

These and other findings have led to an alternative way of modelling human sentence processing: the lexicalist (or interactive, or constraint-based) model, as set out, e.g., by MacDonald *et al.* (1994). Their key assumption is that human sentence processing is largely based on lexical rather than phrase structural information. Their model provides rich lexical representations which incorporate syntactic knowledge, along with a processing mechanism based on an interactive activation account. This enables them to explain how probabilistic and contextual constraints guide syntactic processing.

MacDonald *et al.* (1994) postulate lexical representations which incorporate information crucial to syntactic processing:

- **Argument Structure:** The argument structure (AS) specifies the complements of a lexical entry, including syntactic (e.g category) and semantic (e.g. thematic role) information. Many lexical entries are ambiguous wrt. AS, cf. the MV/RR and NP/S examples in (1) and (2).
- **Morphological Information:** This part of the lexical entry encodes inflectional features like tense, finiteness, number, person, and gender.
- **Phrase Structure:** No separate phrase structure component is assumed, the lexicon rather contains partial X' representations to encode phrase structure: Each lexical entry incorporates the partial syntactic tree it projects.[1]

To account for the cited frequency effects in syntactic processing, lexical items need to include probabilistic information: The assumption is that the lexical representations for AS, morphology, and phrase structure also store information about the respective frequencies.

Concerning contextual influences, the lexicalist model predicts that context facilitates the decision between different types of lexical ambiguity. Typically, context cannot isolate a single alternative in advance, however. Context effects are overridden by lexical influences such as frequency bias.[2]

According to the lexicalist account, syntactic processing does not require the construction of phrase structures, but is a matter of connecting partial syntax trees provided by the lexicon. Hence, it is claimed that no phrase structure rules are processed but rather syntactic constraints which have the effect of enforcing or inhibiting connections between the lexical representations involved. Such a mechanism can be implemented by a connectionist network, for example.

1.3 Predictions for the NP/S Ambiguity

The main prediction of the lexicalist model for NP/S ambiguous verbs is that processing preferences towards the NP or the S reading are mainly determined by the frequency of the respective AS. Contextual factors should only have a limited effect on AS preferences.

Experiments reported by Holmes *et al.* (1989) provided results that are compatible with this prediction. Contradictory findings of Frazier and Rayner (1982) were re-assessed by Trueswell *et al.* (1993), who were able to show that two further effects influence the resolution of the NP/S ambiguity: (a) The relative frequency of alternative ASs (some NP/S verbs can occur transitively or with infinitival complement), and (b) the preference of a verb to omit the complementizer *that* (since the NP/S ambiguity arises only if no *that* is present).

In the present study, we try to show that NP/S preference is correlated with AS frequency. We also investigate the influence of contextual factors such as the information about the presence of *that*.

2 Previous Implementations

2.1 Pearlmutter et al. (1994)

Pearlmutter *et al.* (1994) build on the lexicalist model proposed by MacDonald *et al.* (1994) and present a connectionist implementation to simulate the influence of lexical preferences and context effects on the processing of the MV/RR ambiguity.

They use a three-layer feed-forward network, with the input units representing animacy of the subject, voice, and presence of a direct object. Each verb is identified by a set of semantic features and a unique verb ID. The output units represent the AS of the verb.

For training, Pearlmutter *et al.* (1994) used a set of 60 verbs. All occurences of these verbs in the Wall Street Journal (WSJ) corpus were extracted and coded for input features and corresponding ASs. The resulting set of 176 tokens was presented probabilistically according to WSJ frequency to train the network using back-propagation.

After their model successfully learned on the (unambiguous) training set, Pearlmutter *et al.* (1994) conducted a set of experiments using ambiguous input. To simulate a garden path setting, the same tokens as in the training set were presented, but the units for voice and direct object

[1] Such a lexicalist representation has its analogy in strictly lexicalized grammars as used in computational linguistics (tree-adjoining grammar, categorial grammar).

[2] This claim refers to a very general notion of context: All information which is not part of the lexical entry of the ambiguous verb is regarded as contextual, including, e.g., the case of the argument NPs, their semantic properties, or the presence/absence of a complementizer (cf. sec. 3.1 for details).

were set to neutral values.

They found a significant correlation between the AS frequency in the corpus and the activation of the relevant AS in the model. This shows that the network has extracted corpus frequencies correctly and uses this information for ambiguous input. But the model is also sensitive to context information: It showed a clear preference for ⟨agent, theme⟩ over ⟨cause, theme⟩ if the subject was animate. Furthermore, the model tended to prefer passive AS for inanimate subjects, reflecting the MV/RR preference found in humans.

In this paper, we present a study which extends the results of Pearlmutter *et al.* (1994) to the NP/S ambiguity and also takes syntactic context into account.

2.2 Juliano and Tanenhaus (1994)

Juliano and Tanenhaus (1994) present a recurrent network which models NP/S attachment preferences. They use a localist encoding, i.e., their network has as input only the verb ID and the immediate syntactic context (the word following the verb). From this information, the complement type of the verb is predicted.

Their network was trained on a corpus containing the past tense occurences of 176 verbs extracted from the University of Pennsylvania Treebank (UP) corpus. To test it on ambiguous input, the net is presented only with information about the verb ID. The error scores it achieves under these conditions reflect the attachment preferences for verbs with different frequency biases: A clear preference for NP attachment is found in verbs which can only occur with an NP complement. Verbs which are ambiguous between NP and S complement but have a frequency bias towards NP show an NP preference as well. Even verbs with a frequency bias towards S show a slight NP preference, and also if the net is given no input at all, a global NP preference is observed.

The modelling experiment presented in this paper has a focus quite different from the one of Juliano and Tanenhaus (1994):

- Rather than in absolute corpus frequencies (i.e. the number of occurences of a certain verb in the overall corpus), we are interested in relative frequencies (i.e. the distribution of AS within the occurences of an individual verb). Therefore, we use a subcorpus which is equi-biased for absolute NP and S frequency. This avoids the conditions of Juliano and Tanenhaus (1994), where the global NP preference they find seems predetermined by the way they set up the training corpus (high absolute NP frequency with 44% NP tokens vs. 15% S tokens).

- We try to model the interaction of frequency and context effects. Therefore we encode a set of syntactic and semantic features for the pre- and postverbal context, rather than only the item immediately following the verb. To allow for the investigation of individual verbs, we use a hand-tagged corpus containing the occurences of a small number of verbs found in the psycholinguistic literature.

3 A Model for the NP/S Ambiguity

3.1 Network Architecture

The present study tries to cover a broad range of features possibly relevant to ambiguity resolution. We augmented the feature set used by Pearlmutter *et al.* (1994) and adjusted it for the requirements of the NP/S ambiguity. The factors which are potentially important wrt. resolving the NP/S ambiguity are verb syntactic features (e.g. tense morphology, polarity, presence of modals in the VP), ontological properties of the subject NP and the NP following the verb (inanimate/animate, person/non-person, etc.), and words or phrases in the postverbal context.

Our model is a three-layer feed-forward network (26 input, 16 hidden, and 4 output units), the input layer of which is organized as shown in table 1.

The input units are mapped on a set of output units which represent possible AS of the verb. The ASs occuring in the set of verbs we chose for our subcorpus are the following:[3]

Argument structure
⟨agent, theme⟩
⟨agent, proposition⟩
⟨agent, patient, theme⟩
⟨agent, gerund⟩

3.2 Encoding

We used the UP corpus, which is tagged for POS and contains approx. 4.8 Mio. words of American English, mostly taken from newspaper articles.

A group of nine verbs was chosen for the training and testing sets. Each of the verbs falls into one of the following categories according to its frequency bias:[4]

[3] The first AS corresponds to an NP, the second to an S complement.

[4] Bias is calculated as the ratio of NP occurences to S occurences in the corpus, which has to be larger than 1.5 for NP bias and smaller than 2/3 for S bias.

Function	Units	Type
Verb syntactic features:		
Tense	4	0/1
Polarity	1	0/1
Presence of modal verb	1	0/1
Verb ID	10	0/1
Preverbal contextual features:		
Animacy of NP	1	0/1
Reference to person in NP	1	0/.5/1
Postverbal contextual features:		
Animacy of NP	1	0/1
Reference to person in NP	1	0/.5/1
Case of NP	3	0/.5/1
Presence of *that*	1	0/1
Presence of *to*	1	0/1
Presence of an adverb	1	0/1

Table 1: Features in the input layer

(a) S-biassed verbs: *admit*, *assert*, *imply*; (b) NP-biased verbs: *deny*, *maintain*, *recognize*, *reveal*; (c) equi-biased verbs: *confirm*, *observe*.

We extracted all occurences of these nine verbs from the UP corpus.[5] The occurences were then hand-tagged for the input and output features (as described in sec. 3.1) to provide our training and testing sets. The sets obtained in this way were equi-biased for NP vs. S frequency.[6]

Since we were only interested in cases which can give rise to the NP/S ambiguity, we had to exclude about 30% of the tokens (table 2). These involved structures where the relevant ASs precede the verb (such as passive sentences, sentences starting with a direct quotation, topicalized phrases, object relative clauses), and also sentences where the VP is formed by a conjunction of two verbs, as such tokens are likely to blur the frequency effects for individual verbs.

3.3 Training

Our model was trained using the quickprop algorithm by Fahlman (1988), an improved version of the standard back-propagation algorithm.

The training set reflects the frequency in the original UP corpus adequately, since all relevant UP occurences are contained in the subcorpus used.

Table 2 shows the number of tokens (input vectors) and the frequency distribution of the features (in percent of the total sum of input vectors).

[5]An exception are the verbs *admit* and *deny*, where only the first 50 and 40 occurences were encoded, respectively.

[6]Note that this does *not* match the overall AS preference for NP in the entire corpus. However, for the purpose of our study, we wanted to focus on individual verbs with different AS bias rather than to simulate the absolute AS frequency shown in the corpus.

3.4 Testing

From the 582 tokens, we created a training set of 523 tokens and a testing set of 59 tokens (10%). Both sets were taken from a randomized file. The feature distribution in both sets was equal to the distribution in the overall token set as given in table 2.

After 128 epochs of training the algorithm converged and the global net error was 18.1 (about 10% of the mean error). The error for the testing set with the same weights was 3.6 (about 18% of the mean error).[7]

Then, we investigated the relative influence of different contextual conditions on the network's performance. Three masks were used to set selected groups of input units to default values:[8] (a) ALLOFF: All contextual features were set to default values, i.e., only the verb ID was unmasked; (b) POSTOFF: Just the postverbal context features were set to default values; (c) THATON: Like POSTOFF, but the feature indicating the presence of *that* was not masked.

We first tested the initial training and testing sets for all of these three masks.

Thereafter, we looked at the verbs individually in order to validate the model of MacDonald *et al.* (1994), i.e., to evaluate (a) the strength of the frequency effect, and (b) the relative influence of contextual information on the network performance. Table 3 shows the net error for the individual verbs, both as global net error and in percent of the mean net error.

Table 4 gives the average activation of the output units for the original token set and then for the same set masked with ALLOFF, POSTOFF, and THATON.

The net performance can be determined by correlating the average activations in table 4 with the frequency distribution in the corpus given in table 2 (last three rows). We present the Pearson correlation between frequencies and activations across verbs for each AS in table 5.[9]

4 Discussion

4.1 Context Dependency

Table 4 shows the influence of pre- and postverbal context on the output activations. For the S-

[7]The mean net error is calculated on the assumption that all output units are equal to .5 (random error).

[8]As default value for an input unit, we used the global mean, i.e. the average value of this feature in the overall token set.

[9]We excluded the data for *observe* from the computation of the correlation. This is justifiable, since the sample size is very small (7 occurences).

	S-biased			NP-biased				equi-biased		
	adm.	ass.	imp.	deny	mnt.	rcg.	rvl.	cnf.	obs.	total
UP frequency	109	71	36	81	213	97	49	181	18	855
encoded tokens	50	44	25	40	172	62	38	144	7	582
past	30.0	31.8	28.0	.0	19.2	21.0	44.7	73.6	14.3	35.4
present	60.0	54.6	68.0	75.0	33.1	53.2	39.5	18.8	85.7	45.1
future	4.0	2.3	.0	.0	7.0	3.2	2.6	2.1	.0	3.6
infinitive	6.0	11.4	4.0	25.0	40.7	22.6	15.8	5.6	.0	20.1
preverbal animacy	100.0	90.9	36.0	87.5	91.9	67.8	31.6	77.8	100.0	79.9
postverbal animacy	46.0	38.6	40.0	32.5	9.9	14.5	23.7	41.0	28.6	27.3
presence of *that*	42.0	70.5	60.0	15.0	24.4	29.0	34.2	43.8	57.1	36.6
presence of *to*	12.0	.0	.0	.0	.0	.0	2.6	.0	.0	1.2
presence of adverb	4.0	9.1	8.0	5.0	.0	.0	2.6	5.6	.0	3.3
NP complement	14.0	11.4	24.0	55.0	68.6	59.7	57.9	42.4	42.9	48.3
S complement	72.0	77.3	76.0	30.0	30.8	33.9	34.2	54.2	57.1	46.4
others	14.0	11.4	.0	15.0	.6	6.5	7.9	3.5	.0	5.3

Table 2: Feature distribution (in %)

	orig. vect.	ALLOFF	POSTOFF	THATON
admit	.8 (4%)	10.6 (59%)	14.2 (79%)	10.2 (57%)
assert	1.0 (6%)	6.8 (43%)	5.2 (33%)	1.4 (9%)
imply	.6 (7%)	4.4 (49%)	5.6 (67%)	2.9 (32%)
deny	2.8 (19%)	18.1 (126%)	13.6 (95%)	7.8 (54%)
maintain	5.9 (1%)	58.8 (95%)	65.0 (105%)	12.1 (20%)
recognize	4.1 (18%)	19.2 (86%)	21.8 (98%)	7.5 (34%)
reveal	.1 (1%)	9.8 (72%)	13.3 (97%)	3.3 (24%)
confirm	6.4 (12%)	44.0 (88%)	45.1 (87%)	18.7 (36%)
observe	.0 (0%)	2.4 (97%)	1.4 (58%)	.0 (0%)
training	18.1 (10%)	156.0 (83%)	162.0 (86%)	54.6 (29%)
testing	3.6 (18%)	19.2 (90%)	20.6 (97%)	9.4 (44%)

Table 3: Error distribution

biased verbs, the net exhibits a clear preference towards the S argument structure for the original vectors. This preference is preserved under the condition ALLOFF, where the net has to rely solely on the verb ID to decide between ASs. For the conditions POSTOFF and THATON, we find similar activation patterns.

For the NP-biased verbs, the net shows the expected NP preference for the original vectors. In the ALLOFF condition, however, it exhibits an S bias. This effect is even stronger for the condition POSTOFF. This is unexpected, as one would assume that features such as animacy help the net to determine the correct AS.[10] Interestingly enough, the activations for THATON are similar to the ones for the original vectors, i.e., the net shows an NP preference. This result indicates the relevance of syntactic context (information about the presence or absence of a complementizer) for determining the correct AS.

For the equi-biased verbs, the results are not uniform, which is probably due to the small sample size of the verb *observe*. The results for *confirm* resemble those for S-biased verbs.

	NP	S	other
original vectors	$.997^a$	$.996^a$	$.826^a$
mask ALLOFF	$.515^c$	$.558^c$	$-.242$
mask POSTOFF	$.583^c$	$.640^b$	$-.022$
mask THATON	$.891^a$	$.981^a$	$.087$

Significance: (a) $p < .01$, (b) $p < .05$, (c) $p < .1$

Table 5: Frequency to activation correlation

4.2 Frequency Dependency

In order to analyze the net behavior wrt. to frequency information consider table 5. Here, we correlate the AS frequency in our subcorpus as given in table 2 (last three rows) with the average activations of our network under the different conditions shown in table 4.

The first row gives the correlations for the unmodified input vectors, i.e., the net is exposed to the complete context. The correlation of activation with frequency is highly significant ($p < .01$ for both NP and S), which is not surprising since this simply indicates that the net has learned the frequency distribution of the input set.[11]

Row 2 shows the correlations under the condition ALLOFF, i.e., all input features with the ex-

[10] A possible explanation might be that the high correlation between *that* presence and a propositional AS leads to very strong connections between the *that* node and the S output units. Thus, this is sufficient to produce an S bias even if the input activation at the *that* node is only the default value (global average).

[11] Note that the correlation for non-S and non-NP argument structure ("other") is less significant here and non-significant for the three masking conditions, which is probably due to data sparseness in the training set (containing only 5% of other ASs).

	S-biased			NP-biased				equi-biased	
	adm.	ass.	imp.	deny	mnt.	rcg.	rvl.	cnf.	obs.
orig. vect.:									
NP	.19	.18	.25	.52	.60	.57	.53	.40	.41
S	.71	.74	.74	.36	.40	.40	.37	.57	.57
other	.18	.16	.02	.24	.09	.21	.16	.13	.05
ALLOFF:									
NP	.20	.23	.13	.22	.26	.24	.32	.15	.74
S	.69	.70	.72	.70	.64	.50	.37	.70	.24
other	.02	.11	.04	.06	.11	.38	.10	.23	.03
POSTOFF:									
NP	.13	.26	.10	.30	.24	.26	.34	.17	.43
S	.83	.65	.74	.58	.70	.61	.49	.71	.45
other	.03	.19	.04	.10	.07	.22	.08	.23	.02
THATON:									
NP	.32	.17	.25	.51	.51	.46	.53	.29	.40
S	.68	.73	.72	.35	.43	.42	.35	.58	.57
other	.05	.17	.02	.26	.18	.38	.19	.26	.02

Table 4: Average activations

ception of the verb ID are switched off: The net has to rely solely on the frequency characteristics of the individual verbs. Here, the correlation is marginally significant ($p < .1$), which shows that the net is able to predict the correct AS reasonably well even without any context. This is in line with the claim of the lexicalist model that AS frequency is guiding parsing preference, while context only provides adjustments to the initial bias (sec. 1.3).

The POSTOFF condition (third row) brings about a higher correlation ($p < .1$ for NP and $p < .05$ for S): This indicates that AS preferences are sensitive to preverbal context (e.g. animacy). We conclude that contextual factors can complement frequency information (condition ALLOFF), hence yielding the higher correlation. This is predicted by the lexicalist model (sec. 1.2).

The last row shows a highly significant ($p < .01$) correlation for the THATON condition, both for NP and S: As we saw in sec. 4.1, syntactic information about the complementizer plays a crucial role in determining the correct AS, especially for the S reading (higher correlation), as this reading is signalled by the presence of a complementizer. This is in line with the experimental findings on *that* frequency reported by Holmes *et al.* (1989).

5 Conclusion

Our connectionist model of lexical and contextual influences on NP/S ambiguity resolution supports a lexicalist account of sentence processing (sec. 1.2): It shows that the argument structure frequency of different verbs is correlated with the respective processing preferences. Furthermore, it indicates how frequency information interacts with influences from the semantic and, more strongly, from the syntactic context.

References

Burgess, C. and Hollbach, S. C. (1988). A computational model of syntactic ambiguity as a lexical process. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*, 263–296.

Fahlman, S. E. (1988). An empirical study of learning speed in back-propagation networks. Technical Report CMU-CS-88-162, Carnegie Mellon University, Pittsburgh, Pa.

Frazier, L. (1989). Against lexical generation of syntax. In W. D. Marslen-Wilson, editor, *Lexical Representation and Process*, 505–528. MIT Press, Cambridge, Mass.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, **14**, 178–210.

Holmes, V. M., Stowe, L. A., and Cupples, L. (1989). Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language*, **28**, 265–274.

Juliano, C. and Tanenhaus, M. K. (1994). A constraint-based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research*, **23**, 459–471.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, **9**, 157–201.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, **101**, 676–703.

Pearlmutter, N. J., Daugherty, K. J., MacDonald, M. C., and Seidenberg, M. S. (1994). Modelling the use of frequency and contextual biases in sentence processing. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, 699–704.

Trueswell, J. C., Tanenhaus, M. K., and Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **19**, 528–553.