Is Humpback Whale Song a Language?

Lewis Dartnell

1. Introduction

The humpback whale, *Megaptera novaeangliae*, produces the most complex vocalisations of all 77 cetacean species, which have been dubbed by Payne and MacVay (1971) as “songs”. These songs are hierarchical in nature, with rules seemingly governing their organisation and evolution over breeding seasons. No one hypothesis of the song's function adequately explains its complexity and structure, except perhaps for the theory that it constitutes the first non-human language yet discovered. Buck and Suzuki (1999) have applied Information theory to analyse a sample of humpback song converted into a stream of symbols using a self-organising neural network (SONN). This theory can be used to determine the maximum amount of information contained within a coded sequence by the unpredictability of the next symbol. Different assumptions can be made about the nature of the sequence; the next symbol is randomly determined (thus no hierarchical structure is possible within the sequence), or the probability of the next symbol is dependent on the previous one, or two symbols (0th, 1st and 2nd Order Markov models respectively). It was found that a first-order assumption couldn't reasonably model humpback song, meaning that humpback song possesses a hierarchical structure suggestive of language. The low rate of information transmission, about 0.1 - 0.6 bits per second, may ensure reliable communication over long distances in noisy, unpredictable acoustic conditions.

Language, as opposed to simple communicative signals, has a deep-structure which is both hierarchical and recursive. This recursiveness means that words sometimes greatly separated in a sentence must agree (for example in gender or plurality); so called “long-distance dependency”. Words of the same category (for example adjectives or nouns) behave similarly with respect to syntactically correct positioning within a sentence, and thus language also contains a lexical category structure. Elman (1992) has demonstrated that a partially recursive artificial neural network trained on a set of English sentences can detect and internally represent both recursive and lexical category structure.

This study constitutes a novel approach to analysis of humpback song for signs of language. A network architecturally identical to Elman's was used to test song for evidence of lexical category structure, detection of which would provide support for the language hypothesis.

The rest of this introduction will describe in greater detail pertinent aspects of whale ethology and song, linguistic structure, artificial neural networks in general and Elman's net in particular.

1.1 Songs of the Humpback Whale

Songs span the frequency range 8Hz - 4KHz (Mercado, 1998) and are composed of a hierarchically structured non-random sequence of distinct sound units. A unit is defined as any sound that is continuous to the human ear. If a spectrograph later reveals that it is in fact comprised of multiple discrete pulses or tones in quick succession then they are named subunits. Several units are arranged in a specific pattern to make up one type of phrase, with a series of

similar phrases termed a theme, and many themes making up an entire song of about 10 minutes. The same song is repeated without pause several times within a singing session, which can last up to about 10 hours (Payne and Tyack 1983). Although there are subtle acoustic differences within the units and the repetition number varies at all levels of the hierarchy, songs within one session are very similar. Payne and Tyack (1983) state that very strict rules govern which units can appear in which themes and although sometimes omitted, themes are never sung out of order. It is believed that only lone males sing on the winter breeding grounds, whilst 20m deep and adopting a typical head-down posture. Mercado (2000) has used a computer model of sound propagation to show that males tune their song to frequencies of optimal-propagation and modulate in response to ambient noise.

Humpbacks spend the summer months feeding at high latitudes on krill, but migrate to the equatorial breeding grounds for the winter. There is some degree of mixing in the feeding waters but discrete populations are maintained during breeding, each with its own song (although similarities suggest acoustic contact for some portion of the migratory cycle).

The structure of song changes gradually over weeks, but the entire population adopts alterations synchronously, so that all males sing virtually identical songs. Over the course of several years however, the song becomes almost unrecognisable from an older version. Even these introduced modifications appear to be highly regular, as the pitch, duration, spacing and configuration of units evolve at different rates, and themes are gradually phased out, differentiated into two "daughter" themes, or are created *de novo* (Payne and Tyack, 1983). Little change occurs over the summer, and so it is unlikely to be due to the whale's forgetfulness, and an environmental influence is ruled-out as the trends are not cyclic. The shifts are also far too rapid to be attributable to genetic changes or turnover of individuals, and it appears that they are in fact due to cultural evolution, with new variations transmitted by individual learning.

The function of this humpback song is unknown, as is the adaptive significance of the continually changing details, or the reason why an entire population performs the same song. Many hypothesises as to the function have been proposed, including courtship (Payne and McVay 1971), synchronisation of ovulation, sexual advertisement, spacing of males, migratory beacon, and establishment of dominance (all cited in Mercado 1998). Mercado (1998) presents evidence against these theories, proposing his own hypothesis of sonar echolocation, which explains how singers locate non-vocalising females, and also the synchronous and dynamic nature of the song. "Cocktail party processors", assuming they exist in the humpback auditory cortex, enhance discrimination between neighbour's emissions and echoes if the original transmitted sounds are known, hence the similarity of song throughout the population. An individual cheater benefits if he modifies his own song slightly, as that maintains the efficacy of his own sonar whilst impairing neighbours. Other singers are now under pressure to pick up on the modifications, resulting in a dynamic equilibrium. The fact remains however that the degree of structure and complexity within humpback song is far in excess of what would be required if sonar were its primary function. Could these vocalisations therefore be some form of language?

The precise motor patterns required for human speech are generated within Broca's area of the cerebral cortex. Wernicke's area, close to the auditory cortex, is involved in the comprehension and also production of speech. It is not known if analogous structures exist within the humpback's brain, but their discovery would strongly support the language hypothesis. The auditory system

is certainly very sophisticated, with five times the human number of ganglia in the auditory nerve, cochlea with twice the cellular density of a human, and sound localisation abilities believed to be comparable to the best terrestrial examples, such as bats.

1.2 What is Language?

Language is still believed to be the only uniquely human attribute, and Table 1.2a shows the characteristics of non-human communication. All 6,000 human languages however are based on a discrete combinatorial system (grammar) that can potentially produce an infinite number of sentences by specifying rules that determine how to arrange words.

Finite repertoire of signals	E.g. vervet monkeys have a different alarm call for eagles, snakes and leopards
Continuous analogue signal representing magnitude	E.g. the activity of a honey bee's waggle dance
Series of random variations upon a theme	E.g. the majority of bird song

Figure 1.2a - The characteristics of non-human communication. Compiled from information contained within Vehrencamp (1998).

Language is compositional, meaning that each of the combinations has a different meaning, determined by the meaning of component words and the grammatical rules that arrange them. Language is thus composed of words and rules, with grammar determining both the form and meaning of sentences. Syntactical rules specify where particular word types appear in a sentence; in English for example adjectives precede nouns and subjective nouns precede the verb. These rules give rise to a lexical category structure.

As can be seen in 1.2b, grammatical structure, or "deep structure" is hierarchical; a complete branch of any depth can be substituted to create another grammatically correct sentence.

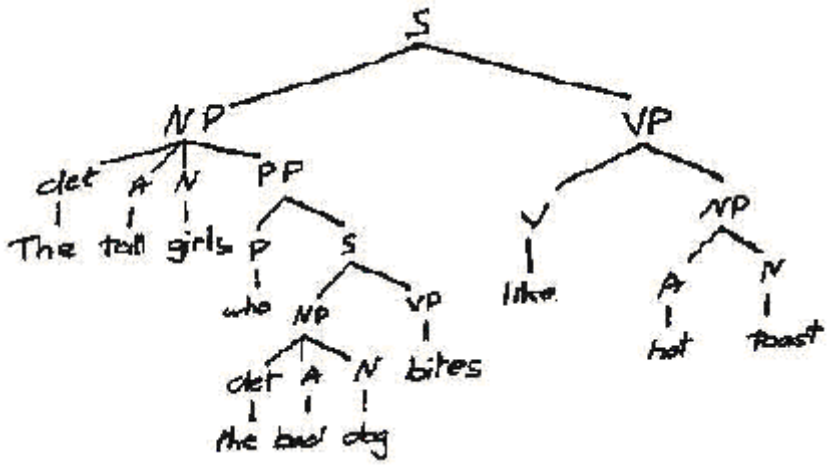

Figure 1.2b - The hierarchy of deep-structure and its recursiveness: an entire sentence is nestled within the proposition phrase. An example of a long-distance dependency is that `like' must agree with the plural `girls', and not the nearest noun, singular `dog'. S=sentence, NP=noun phrase, VP=verb phrase, PP=preposition phrase, det=determiner, A=adjective, N=noun, P=preposition,

V=verb.

The real creative power of grammar however lies in its recursiveness, whereby nestled branches can theoretically form limitless sentences. This recursiveness results in the need for grammatical agreement of words sometimes greatly separated in the sentence, so-called "long distance dependencies" (Chomsky 1957, cited in Pinker 1994). Chomsky furthermore showed that recursiveness prohibits the emulation of language by a word chain system, whereby the next word or phrase is chosen from a list of possibilities with predefined probabilities. Such a "finite-state Markov process" would require lists of infinite length. Language thus possesses both recursive and lexical category structures.

1.3 The Neural Network (NN)

Modern artificial neural networks (NNs) are computer simulations used to model biological nervous systems. Elman used such a model to demonstrate that children could indeed learn about correct grammatical structure merely by hearing example sentences - it need not be innate. NNs consist of a set of nodes (neurons) interlinked with connections (synapses), each with its own value of weighting (synaptic strength). Each individual node is activated to a certain extent depending on the summation of activations of all other nodes it receives input from, and the weight of the connections, which can be either excitatory or inhibitory. Input nodes receive extrinsic activations (from outside the network), which pass to all other linked nodes, being modified by the connection weights, and eventually results in a pattern of activations in the output nodes. In addition a bias node can form a connection with all hidden and output nodes. The input data must consist of a sequence of basis vectors, whereby only one bit is turned ON in each vector, thus ensuring that the vectors representing different things are all orthogonal to each other. All input vectors are equally disparate, and no bias is introduced that may affect the internal representations. A nonlinear logistic activation function is usually used to calculate a node's activation resulting from the net input of all connected nodes, whereby:

$$\textbf{activation} = 1 / (1 + e^{-\text{net input}})$$

The graph of this "squashing function" is sigmoidal in form and ensures the activation of a node is always between 0 and 1, as shown in Figure 1.3a.

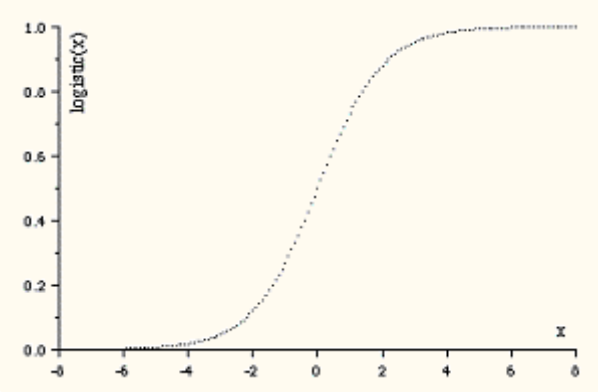

Figure 1.3a - The sigmoidal squashing function

During the training phase an input pattern is supplied with a target output pattern, and a back-propagation of errors algorithm is applied to modify the

connection weights. The error for each output node (the extent its actual activation differs from the desired activation) is calculated, and then the weights of all connections that contributed to that activation are modified, recursively proceeding back through all lower layers of the network. The extent of modification depends on how much of the error can be attributed to that connection; worse offenders are punished more severely. The *learning rate* can be altered by using different proportions of the error derivative to make changes to the weights, and the *momentum* parameter determines what proportion of the previous weight change is used in the current cycle. The training set is presented to the network many times, until its global error (root mean square of all output nodes) has dropped to an acceptable level, or it has satisfactorily solved the problem. Back-propagation is a gradient descent algorithm, in that it attempts to optimize NN performance by seeking the lowest point on the error hypersurface, or *landscape* (see Figure 1.3b).

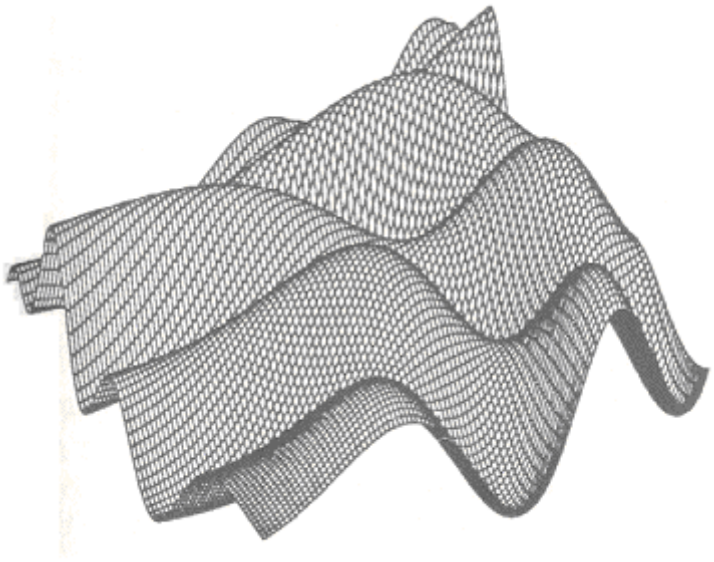

Figure 1.3b - A 3-D representation of an n-dimensional error landscape

It is vulnerable to getting stuck in a local optimum, and so the NN must be retrained from scratch a few times in order to ensure that the solution found is in fact the global minimum. Thus, through the gradual process of weight modification, a NN in an initially randomized state learns how to perform the correct vector transformations. The solution of linearly inseparable problems such as the Boolean function XOR requires an additional layer of “hidden nodes” to allow the formation of a internal representation. This representation does not reside in any specific node, but is *distributed* across them all.

If the network is trained for too few sweeps it will obviously under-perform. NNs can also be over-trained, and the reason for a low error may not be that it has successfully characterized a pattern or generalized between examples, but has simply been presented the same sequence too often and has memorized the exact order of the entire training set. The desired trade-off point between these two extremes is known as Occam's Cliff. In order to discount the possibility that a NN has been over-trained and has memorized the entire training set, it is often tested (the phase after training, when weights are frozen and the output patterns studied) with input patterns that were not present in the training set. The network will perform relatively poorly on these patterns if it has not discovered general patterns but simply learnt the training set.

The architecture of the network is crucial to its effective functioning. The optimal number of hidden nodes must be discovered heuristically. Too few hidden nodes and the network will have too few resources to form useful internal representations, too many and it will not be forced to discover generalizations or will simply learn the training set. The arrangement of these nodes is also important; if they are ordered into too many layers then the effectiveness of back-propagation and development of useful representations will be compromised. Connections between all layers or only to adjacent ones will also affect performance. All the networks described so far have been feed-forward (FF), in that activations proceed in only one direction, from input to output nodes. One exceedingly common feature of biological networks however is recursive pathways, whereby nodes connect back to themselves or lower layers, allowing a multidirectional flow of information. Partially recurrent networks, such as Elman's, have a degree of memory of previous states, and are able to recognize patterns in time as well as space.

1.4 Elman experiment

Languages possess both recursive and lexical category structure. The former results in long-distance dependencies and the latter restricts word types to particular positions within a sentence, with the result that certain categories are frequently juxtaposed. Elman (1992) showed that NNs could correctly identify and internally represent these two forms of structure. Only the lexical category structure will be tested for in humpback song.

Elman used a program to generate a corpus of 10,000 two or three-word sentences from 29 nouns and verbs. Each word was converted into a basis vector and presented without break between sentences to a 3-layered simple recurrent network; the Elman net. The hidden nodes were reciprocally wired one-to-one to an equal number of "context nodes", with the connection weights fixed at 1.0. This fixing of weights allows the use of back-propagation, even though the network is partially recurrent. See Figure 1.4a for an example of the partially recursive architecture with contextual nodes, although Elman's actual net had many more nodes.

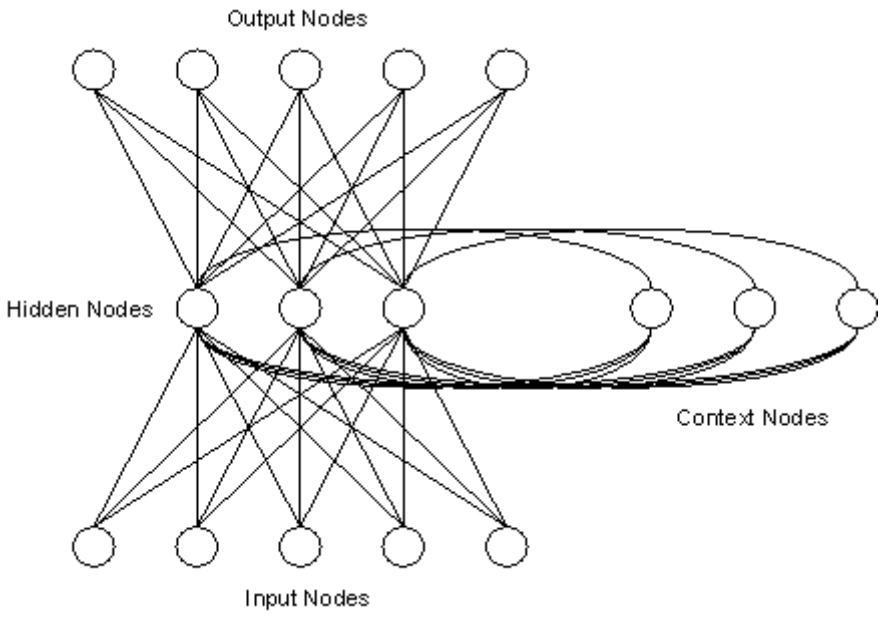

Figure 1.4a - Generic architecture of an Elman net

Each time step the hidden node activations were copied into the context nodes, and the context nodes are combined with the new input vector to activate the hidden nodes. The context nodes therefore contain information on prior states of hidden node activation - a memory of previous words that provides the hidden nodes with the context of the current word. The network was trained to predict the word after the one currently presented to its input nodes, and so contextual information is crucial. Although the choice of next word is constrained by grammatical rules, there is still a great range of permitted successors. The NN, unless it learns the entire training set, cannot therefore be expected to guess accurately every time, and in this instance the RMS error graph is not a useful indicator of whether the network has developed effective representations or not.

During the testing phase, weights were frozen, and every word was re-presented in each of its various contexts. The hidden node activations produced were averaged for each unique word, and then all 29 mean vectors were subjected to hierarchical clustering analysis, which allows analysis of the internal representation. N-dimensional Pythagorean geometry calculates which pair of input vectors produces the most similar hidden node pattern, and clusters them together. Clusters are then grouped on their similarity, or closeness in representational space, until a diagram of the whole test dataset has been built up. The task is one of prediction, so a cluster of words signifies that the network considers them to be related by all preceding the same other words. Figure 1.4b shows the resulting diagram, with a distinction made between nouns and verbs, and between verbs for which a direct object is absent, optional or obligatory. Importantly, the category structure is hierarchical: boy is a member of human, is a member of animates, is a member of nouns. Categories higher in the hierarchy correspond to larger regions of representational space.

The network, with access only to the behaviour of words in a sentence (with respect to co-occurrences and thus placement restrictions), with no bias from the way words were presented and no *a priori* knowledge of meaning, has detected the lexical category structure. If the hypothesis that humpback song also possesses a grammar indicative of language is correct, then the category structure as revealed by cluster analysis should also be hierarchical and logically organised.

2. Method

Elman (1992) has shown that a partially recursive NN with contextual nodes, as seen in Figure 1.4a can detect and internally represent the lexical category structure of a language. This same architecture will therefore be used to test humpback song for evidence of such structure. A NN program named Tlearn, created by Plunkett and Elman, was used to construct, train, and test the network, and analyse the resultant internal representation. Data can only be presented to a network as a pattern of input node activations, and so the raw hydrophone recording must be aurally and spectrographically analysed, units categorised and the song converted to a sequence of orthogonal vectors.

In total, three separate data sets were used. The first, 9:30 minutes of commercially available song recorded by Dr Roger Payne was of dubious quality. It was a relatively short recording (small dataset) and repetitive sequences (exactly what is of interest) had probably been edited out for commercial release. The Payne data was only initially used because despite repeated efforts (personal communications with Buck, Fisher, Mercado, and Suzuki) no alternative data set could be obtained. A twenty-minute recording

was bought from the National Sound Archives, but not preprocessed or analysed because Dr Robert Fisher responded and provided a large ready-characterised dataset. This data set consists of 450 units, representing 3 complete songs and the last presumably incomplete. These had already been categorised into a sequence of symbols by a SONN, with the division into themes and songs shown in Table 2a. The codified singing is comprised of 28 different units, although as in the Payne recording some are obviously alike and have thus been designated similarly (O, O' and O" for example). There are 7 distinct themes (I - VII), some more prominent than others (Theme I accounts for 208 of the units, whilst Theme IV contains only 2 units), and the patterns in some themes being more obvious than in others (eg. Theme I has a very regular pattern, Theme VII is much more obscure), see Table 2b.

Theme I	*tends to be*	A…B…[C…E]repeated…B…..A...B…*etc*
	or	A…B…[E'..C']repeated…B…..A…B…*etc*
Theme II	*tends to be*	D and D' units interspersed
	evolving to	Q and Q' units interspersed as the song progresses
Theme III	*tends to be*	F repeats broken up by single Bs
	or	F' or J repeats broken up by single B's
Theme IV		Insignificant theme consisting of only R units
Theme V		Insignificant theme consisting of only N units
Theme VI	*tends to be*	G repeats broken up by single Hs
	evolving to	S repeats broken up by T…I…T
Theme VII	*tends to be*	short P repeats broken up by O, O', O" or M units, no obvious pattern

Table 2b - The patterns evident in the Themes of Fisher's data

Table 2c show the unit sequence against position number.

Before the Fisher data was analysed however, an artificial dataset with structure similar to that of the Fisher data and with known pattern was created then analysed, to double-check that the NN methodology was valid. The dataset was composed of ten units (A-J) arranged in three Themes, as shown below.

Theme I: (ABCDCDCDCD) repeated 4 times
Theme II: (EFEEFEFEEF) repeated 8 times
Theme III: (GHHHH) repeated 4 times followed by (IJJJJ) repeated 4 times

The network successfully represented the lexical category structure, and so Fisher's data was analysed third. The conventional teaching method failed for this data, and so a second technique of incremental teaching (as described by Plunkett and Elman, 1997) was successfully applied. The three datasets used were thus:

1) Payne
2) Artificial song
3) Fisher (conventional + incremental teaching approaches)

Once processed into a vector sequence, and a teaching set of target activations created, the data sets were used to train the NN to predict the next unit. To solve this task, the NN would need to internally represent the patterns and lexical categories, if present.

2.1 Preprocessing of raw Payne data

An audio analysis program, SpectraLAB version 4.32.15, was used to display both the waveform (amplitude-time graph) and spectrogram (frequency-time) of the song, the latter produced by Fast Fourier Transformation (FFT) of the signal from the time domain to frequency domain. This process relies on the fact that any complex waveform can be approximated by a superposition of *n* sine waves with different amplitudes and wavelengths.

Synchronised waveform plot and colour spectrogram were printed-out as shown on preceding pages, so that units could be visually analysed and compared. The chronological position and putative category of units was marked on a black-and-white spectrogram. Audio (through headphones for improved reproduction) and visual analysis was performed to determine which units were most similar to other already categorised units, or dissimilar enough to warrant a new category. Similar units were played back-to-back to aid comparison. Additional spectrograms with expanded time-axes were created to aid discrimination between unit and echo (lower amplitudes on the waveform) in fast pulse-trains, such as Figure 2.1a. Harmonics and echoes blurring units in the frequency and time axis respectively and FFT artefacts all conspired to complicate the task. As more units were analysed categories were divided into two or occasionally merged.

Thus, through audio and spectrographic (eg. gradient of upsweeps, shape and size of tail) analysis, Table 2.1b with unit number against category (named by how they sounded or appeared on the spectrograph) was finally completed. Examples of the two naming strategies being *Brrr, Whistle, Whoop, Squeak, Grunt*, and *Levee, Hooked, Straight, Cup, Tick* respectively. Some phonemes were very rare, if not unique, and did not appear to be similar to any others. These were named by their chronological number, such as Type 1, Type 115 or Type 173. There would often be a smooth transition in form between one sound unit and another in a sequence, for example units 23 through 36, as shown in Figure 2.1c.

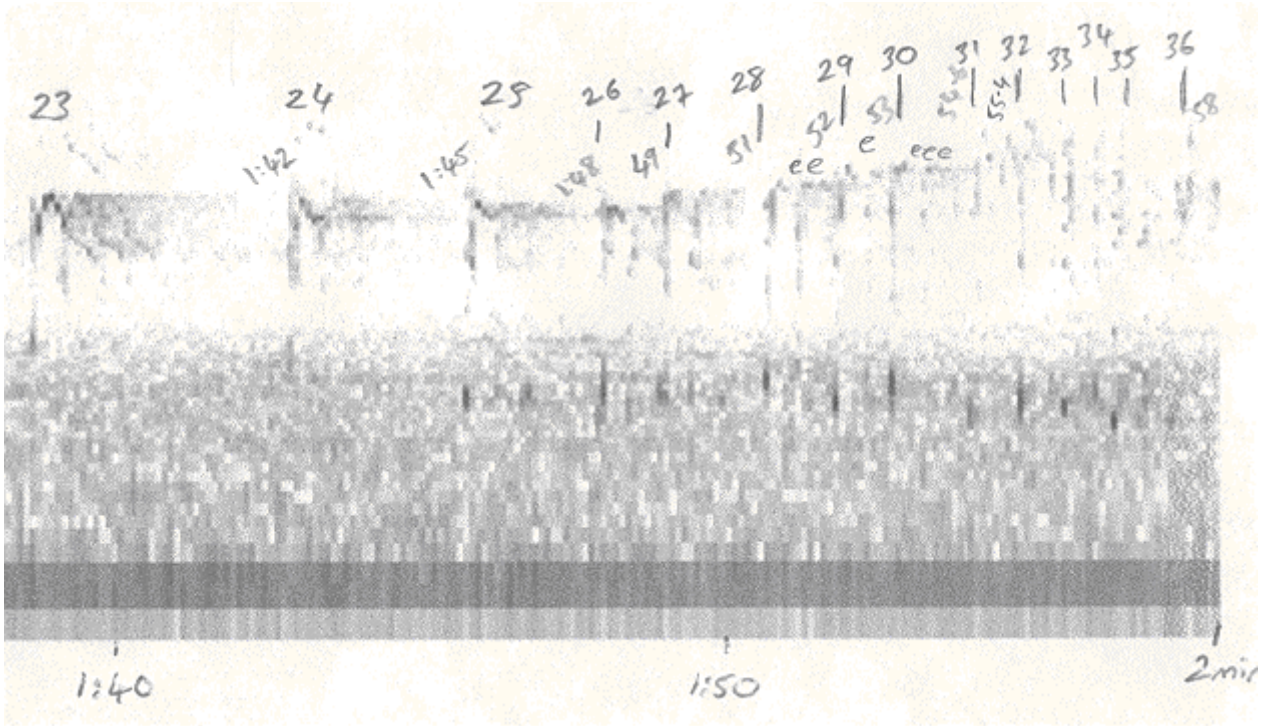

Figure 2.1c - Smooth transition between unit types. Unit number and time as marked. e = echo

In these cases the unit at the end of the sequence was obviously very different from the first one, but there was no clear distinction between consecutive phonemes. A fairly arbitrary division was therefore placed, with the units immediately before and after it given different labels even though they sounded and looked quite similar. Occasionally extrinsic sounds were heard, such as the thunderclaps at 4:07 and 4:50 minutes (as indicated on the monochrome spectrogram).

The word labels for the units are more convenient for a human brain, but are unwieldy when manipulating long lists of units in the song. For this reason the unit categories were arbitrarily ordered, and then assigned an alphabetical label. Some unit categories, although separate, were deemed similar enough to warrant acknowledgement of the resemblance; G, G' and G" for example.

For presentation to the input nodes of the NN, these categories were each given a 29element basis vector, as shown in Table 2.1d. A simple parse written by Torsten Reil performed this letter code to vector conversion.

2.2 Data Analysis

As described previously, a total of three separate data sets have been analysed. In all cases the Tlearn program and an Elman architecture were used, with the number of input and output units each equalling the number of different unit categories (29, 10, & 28 in Payne, Artificial and Fisher respectively). See Figure 2.2a for the network used to analyse the Payne data.

Figure 2.2a - The architecture of the initial neural network (Payne data)

The architecture is specified to the program in a .cf text file, with input vector sequence and expected output node activations in .data and .teach text files respectively. Only the Payne data required extensive preprocessing; Artificial and Fisher were simply converted to vector sequences before being taught to the NN. Like the Elman experiment, the task was one of prediction, and the expected output node activation is simply the next unit in the data sequence. The .teach file is therefore created by displacing every vector of the .data file one line upward, with the first vector now the last (see Figure 2.2b).

Configuration File:	**Sample of data file**	**Sample of teach file**
NODES: nodes = 48 inputs = 28 outputs = 28 output nodes are 11-38 CONNECTIONS: groups = 0	distributed 450 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	distributed 450 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

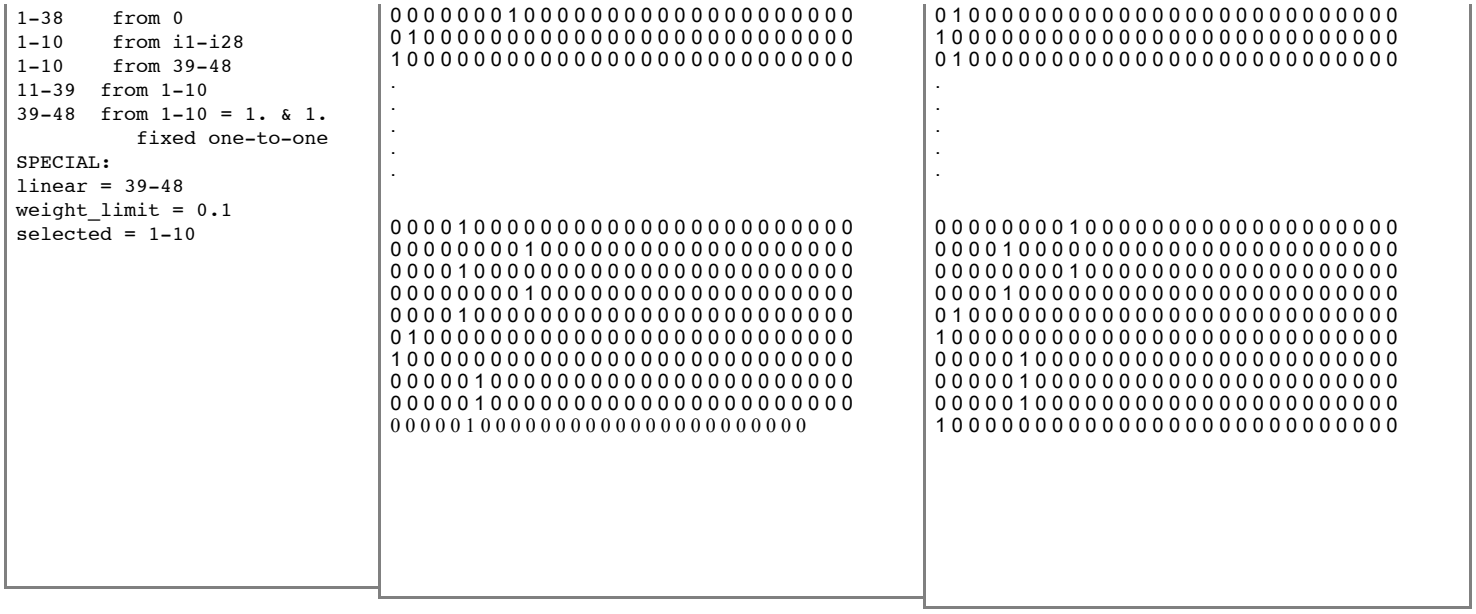

Figure 2.2b - Examples of the .cf, .data and .teach files used in analysis of Fisher data

Due to the problem of getting stuck in a local error minimum, training was repeatedly restarted with randomised weights and different values of learning rate and momentum, to check that the NN was consistently arriving at similar internal representations. If not, then the global minimum has not been found, and the internal representation and thus cluster diagram is effectively random and meaningless. The greater the number of hidden/context nodes or times the NN is presented with the training set, the greater the risk that the NN has not represented general patterns or lexical categories but has simply memorised the entire data set (again rendering the cluster diagram meaningless). The training phase was stopped as soon as the network was predicting accurately. The number of hidden and context nodes were varied for best results, but set initially at the same ratio of hidden:input nodes as Elman's net.

After the training phase, the network was presented with the training set again, and its prediction compared against the actual next unit, to check that the NN is predicting accurately.

Additional methodology on Fisher dataset

The conventional method of training the network with the entire data set from the outset failed to produce consistent cluster diagrams of the Fisher data. Plunkett and Elman (1997) describe similar problems in trying to train a network to form a representation of relative causes within a sentence. They found that by initially presenting simple, non-recursive sentences, and then working up though longer and more frequent recursions the network performed immeasurably better. It is almost as if the untrained network is overwhelmed by the complexity and thus fails to form representations of lexical categories. Incremental teaching, or “starting small” can thus be an effective remedy. This approach was applied to the problem of getting the network to form representations of the song units. The song was simplified to include only Theme I sequences in the input set, and trained until the network predicted accurately. Theme II units were then added, the network trained for 500,000 sweeps, and tested before the Theme III units were included, and so on. Thus the training set was cumulatively built up, in the order: Themes I, II, III, VI, VII, IV, V (the more prominent themes added first).

Comparing each of the 450 output activations (predictions) to the actual next unit is prohibitively laborious, and so a MS Excel spreadsheet was created with blocks of current unit, actual next vector, predicted next vector, and a fourth with cells conditionally coloured to indicate incorrect predictions. See Appendix II.

TLearn can plot up to about 200 hidden node vectors on the cluster diagram, and so cannot display a tree for the entire song of 450 units. To get round this, the list of hidden node activations during the testing phase was re-arranged using MS Excel into groups of the same theme, and only one theme clustered at a time. The activations of units that formed tight groups were then averaged, and the mean vector plotted on the grand tree (spreadsheet reproduced in Appendix III). The important aspect is how the units cluster, rather than the orientation of the clusters, and length of the branches is less important when average vectors are plotted. Thus the diagram has been cosmetically but not geometrically altered, standardising branch length to make the structure more obvious.

3. Results

3.1 Payne data

The cluster diagram representations of the hidden node activations were always completely different, without much evidence of a hierarchical organisation (Figure 3.1a). Inconsistent clustering and step-like organisation are symptomatic of a NN that is unable to find any pattern within the data. No reliable vector-vector transformation rules were found and so its internal representation (which the cluster diagram is based on) is disorderly, and effectively randomly arrived at each time the experiment is re-run. Although the large-scale structure was different each time, careful analysis of the tree diagrams revealed that certain categories did in fact tend to cluster together. Levee, Double Levee, and Half Levee are usually placed close to each other and reference to Table 2.2c will show that these categories were indeed considered to be very similar (with labels G, G' and G" respectively). Squeak, Squeal, Straight, Whistle and Whoop also tended to be placed close to each other.

3.2 Artificial data

The network's prediction was always correct (Table 3.2a), even when the unit could be followed by two different units (eg. E is followed by E or F with equal frequency). The network simply turned on the output nodes representing E and F each with about 0.5 activation. When multiple answers are possible (eg. H is most often followed by another H, but sometimes by G or even I) the network guesses the most common, so as to minimise the average error.

Current Unit	Actual Next Unit	Predicted next Unit
A	B	B
B	C	C
C	D	D
D	C / (B)	C
E	E / F	E / F
F	E / (G)	E
G	H	H
H	H / (G) / (I)	H

I	J	J
J	J / (A)	J

Figure 3.2a - NN predictions

The cluster diagram produced (Figure 3.2b) is also very logical, with units that precede the same next unit clustered together, and units of the same theme also clustered together. This is exactly the sort of precise, hierarchical structure that the true whale song would produce if it also has consistent lexical categories.

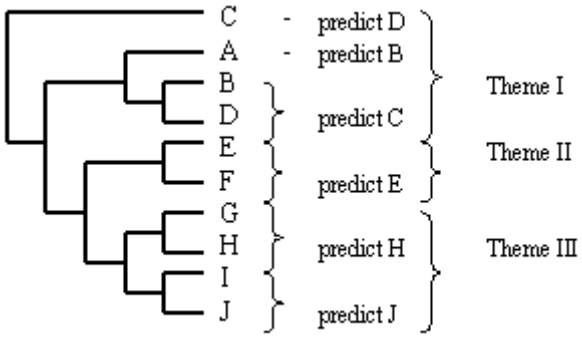

Figure 3.2b - Internal Representation of the Artificial dataset

Thus it has been demonstrated that the methodology is valid and in principle capable of detecting lexical category structure, if it exists, in humpback song.

3.3 Fisher data

126	A	
127	B	
128	C	
129	E	1st E in an EC repeat
130	C	
131	E	
132	C	
133	E	E in the middle of an EC repeat
134	C	
135	E	E at the end of an EC repeat

Figure 3.3a

After training the network on only Theme I, the cluster diagram was found to be highly ordered (see Figure 3.3), and all units had clustered with their own kind (no B unit was found within an E cluster for example). The technique of drawing cluster diagrams of one theme at a time and averaging the tightest clusters to produce the global diagram of the entire song is thus justified, since it is known that no dissimilar units ever invade a tight cluster.

Furthermore, the way in which the network had represented the different units was astounding. The E's all clustered together, with the ones that appear in a first CE repeat distinguished from those which appear in the middle or end of a repeating CE sequence (eg. positions 128-135, see Figure 3.3a). E' units were also ordered by their position in a repeated sequence.

262	A	Start-repeated-sequence B
263	B	
264	C	
265	E	
266	C	
267	E	
268	C	
269	E	
270	B	End-repeated-sequence B

271	A	
272	B	Start-repeated-sequence B
273	C	
274	E	
275	C	
276	E	
277	C	
278	E	
279	C	
280	E	
281	C	
282	B	End-repeated-sequence B
283	A	
284	B	Start-repeated-sequence B
285	C	
286	E	
287	C	
288	E	
289	C	
290	E	
291	C	
292	B	End-repeated-sequence B
Figure 3.3b		

Bs which started a CE or E'C' repeat were clustered distinctly from those that ended a repeating sequence (Figure 3.3b shows the two different contexts in which B is used).

Furthermore, Bs that ended a CE repeat clustered next to, but distinctly from those that ended an E'C' repeat.

37	A
38	B
39	E'
40	C'
41	E'
42	C'
43	E'
44	C'
45	B
Figure 3.3c	

It is obvious when the network has double guessed, for example position 44 (see Figure 3.3c). This is after a line of E'C' repeats and the network is unsure whether an "end-of-sequence" B will follow, or an E' of another repeat. Further inspection, for example of positions 116-137 (Figure 3.3d), reveals that the network has realised that CE or E'C' repeats stretch on average for about 3 repetitions, so in the position that would be the fourth C or E' it guesses both B and the unit which would start another repeat.

In unpredictable sequences, for example 48-73 where D' irregularly intersperses a D repeat; the network simply guesses D each time, as this is the most common unit and error is thus minimised. Where the distribution is more even, such as the Q

116	A	START
117	B	
118	C	1st repeat
119	E	
120	C	2nd repeat
121	E	
122	C	3rd repeat
123	E	
124	C	
125	B	END
126	A	
127	B	START
128	C	1st repeat
129	E	
130	C	2nd repeat
131	E	
132	C	3rd repeat
133	E	
134	C	4th repeat
135	E	

136	C	
137	B	END
Figure 3.3d		

or Q' sections in positions 324-342 and 383-292, the network tends to guess both. For regions with little pattern (eg. 225-235) the network makes no firm predictions (see Table 2c for all these sequences).
The network therefore performs very well on the whole, and only tends to predict inaccurately when a previously regular pattern breaks down, during relatively unstructured themes, or at the transition between themes. If the next unit is ambiguous then the network splits its guess, as was also done by Elman's network.

Now that the network's performance has been tested, the cluster diagram of averaged vectors must be studied (Figure 3.3e). Even random hidden node activations would produce a tree, and so a positive result is a non step-like organisation, with clusters logically nestled within each other. Although the tree is not particularly step-like, a perfect hierarchical structure is also not present. The clustering of like-predictors (eg. all the As, Cs, C's, Es, and E's which precede B) is also not perfect, and the unit categories from the less structured themes (such as Theme VII) are scattered apparently randomly through the tree. Thus although a very similar methodology to Elman (1992) was executed, the resultant internal representation is not as completely logical. The cluster diagram is not totally disordered however. Many like-predictors have indeed clustered together, and these clusters themselves are to a certain extent grouped by like-Themes. This, as described previously, is exactly the form of logical clustering and hierarchical nature suggestive of a lexical category structure. These attributes are more prominent in the cluster diagram of just Theme I (the most regularly patterned theme), as seen in Figure 3.3a, with one particular example emphasised in Figure 3.3f.

4. Discussion

4.1 Payne data

The failure of this first experiment was probably due to inadequacies of the data and not a fault in the architecture of the NN, or the experimental procedure as a whole. The data set was very small and probably inadequately characterised. Erroneous division of natural groups into several unit categories, or the forced fusion of several distinct categories would have obscured any pattern that was present. Quine (1953) describes the problem of categorising the phonemes of a previously unknown language. When utterances of similar length and acoustical form are encountered, it is unknown if the two are slightly different phonemes, or exactly the same phoneme with slight variance in pronunciation (insignificant variation to a native speaker). Every sound is unique, and so the problem becomes one of delimiting a multidimensional and continuous space of sound forms (with axis of duration, start frequency, gradient of upsweep, etc) into meaningful categories of similar sounds, with as much objectivity and consistency as possible. Sometimes natural clusters are not apparent, and arbitrary divisions must be set. At least as importantly, the song had almost certainly been amended for the commercial CD, editing out monotonous repetitions of units that are exactly what this research is interested in.

Although the many cluster diagrams plotted were all different on the large scale, certain unit types did tend to loosely cluster together (see Figure 3.1a). Levee, Double Levee, and Half Levee (G, G' and G") may have clustered together because the NN noticed that the units of these categories behave similarly

(constrained by the same placement limitations, or occur after similar other categories), hinting at the detection of a lexical category structure. Also possible however is that these are not natural categories, and in fact should be more properly merged together. The clustering may represent the actuality that these unit categories are in fact the same, and it is thus not surprising that they behaviour similarly in the song.

More intriguing however is the loose association of Squeak, Squeal, Straight, Whistle and Whoop common to many of the cluster diagrams (again see Figure 3.1a). The units of these categories are clearly acoustically very different, and this may therefore represent better evidence of lexical category structure - different units behaving in identical ways, such as `boy' and `girl' behaving similarly in the Elman experiment.

4.2 Fisher data

All network architectures and training parameters initially failed to usefully internally represent this data. The data is more extensive and obviously more structured than the Payne recording though, and thus more likely to contain a detectable lexical category structure. The success of Elman's similar experiment and the artificial training set has already established the validity of the NN methodology. Another technique, that of incremental teaching, or "starting small", however solved the problem. The network was able to characterise each theme at a time, and was not overwhelmed with complexity from the outset. The network performed well in its prediction task, especially considering the non-deterministic nature of the sequence. The units of Theme I were anticipated particularly well, and error-limitation methods were employed for the less patterned or effectively random sections. The only concern is that the network may not have extracted the patterns present, but merely memorised the order of the units in the entire training set. Training was stopped as soon as the NN was predicting reliably, and so this risk has been minimised as much as possible. Furthermore, testing revealed that the network predicted inaccurately during regions devoid of rigid pattern, suggesting that it is operating general principles, and has not memorised the entire training set. The only way to be sure though is to test the networks performance on a novel data set. No more data is available however (see communiqué with Fisher in Appendix I), and it was deemed better to train the network on the entire set than to divide it into two and train with only 225 units, especially since the pattern of unit occurrence evolves over the course of the three and a bit songs. It would be possible to create a pseudo-original set by counting the frequencies of unit transitions and using these as probabilities in a word-chain system to create an artificial testing set. This has not yet been attempted because of the labour and time involved, and would anyway be of limited value due to the similarity it would possess to the original training data. The best option would be to record two long sequences from the same whale on successive days, training the network on one and testing it on the other.

Confirmation that the network was logically characterising Theme I was provided by the highly ordered structure of the cluster diagram (Figure 3.3a). Not only was the NN representing the differences between the unit categories, but also differences between the same units in different positions in a repeating sequence. The network had also appreciated the syntactic difference between a B in the position …A**B**CECE…(START repeated sequence) from … E'C'E'C'**B**ABCE… (END repeated sequence). Thus B is analogous to a homonym in English - the same unit has two distinct behavioural patterns. If *quail* were included in Elman's training set, one would expect to find *quail* as in

game bird clustering with the other animal nouns, and *quail* meaning to falter to cluster among the verbs without a direct object.

Figure 3.3f shows the middle cluster comprised of both the Bs that precede a CE repeat and the Es involved in this repeating pattern. The Bs and Es are separate types, but both behave similarly by preceding C. Thus this region of the cluster diagram is both hierarchical and logically ordered; fulfilling the prerequisites of evidence for a lexical category structure.

Elman (1992) describes the distinction between types and tokens. A type is the unique word, such as boy, which appears in a slightly different context in each sentence; girl likes boy, monster chases boy, monsters eat boy, whereby each occurrence is a token. Elman found that all the tokens of a type clustered tightly together (see Figure 4.2a), with tokens of boy as an object clustering together and distinct from boy as a subject, just as was seen in Theme I with the E cluster divided into ones that appear at the beginning, middle or end of an EC repeated sequence. Thus the E category can be viewed as a type, and each of the 40 occurrences of E a separate token.

The diagram of the entire song (Figure 3.3e), however, shows that tokens of the same type do not in this instance form a single tight cluster, in contrast to Elman's result. Except in the case of B, this is unlikely to be because the type has a multiple function. The network probably has an imperfect internal representation, due to a short dataset, the sometimes obscure pattern or non-optimal network architecture. Elman was thus justified in representing an entire type as a single average vector, but this would not be valid for the unit categories of the humpback song, and it would also have lost the subtle difference between B in its two functions. Nonetheless, despite the high incidence of illogically positioned tokens and lack of tokens all forming one tight cluster, examples of hierarchical clusters suggestive of a lexical category structure do exist. Recognising logically nestled clusters, such as boy, humans, animates, nouns is easy in English, as the categories are already known. Identification of such lexical categories in humpback song when ignorant of what the categories might be is of course much more difficult, but possible in principle.

Units which precede S or I/S are clustered together, as are units that precede M, P, O” or O. Predictors of C', E and C are grouped together, and form a super-group of units within Theme I (see Figure 3.3e). This is comparable to the clustering in Elman's experiment; boy is a member of “humans”, is a member of “animates”, is a member of “nouns”. The tokens of C are a member of the “Predict E” cluster, which is a member of the “Theme I” group. Such a hierarchical organisation to the cluster diagram is much more prominent in the network trained only on Theme I, as recreated in Figure 3.3f. The “B: Start E'C' repeat” tokens seem at first to have been misplaced. It must be taken into account however that the network has no foreknowledge that E' actually comes next; it can only use general principles in predicting a non-deterministic sequence, and thus all 3 clusters of Bs are equivalent in the NN's internal representation.

Although evidence of an equivalent of the lexical category structure detected by Elman (1992) was found in the cluster diagram, a few hierarchically nestled clusters within a largely disordered tree does not constitute conclusive proof. As mentioned previously, designing the network architecture is a heuristic process, and it is possible that the one settled on was sub-optimal. The dataset was shorter than ideal, and may have been imperfectly characterised by Fisher's SONN. There is also the problem of deciding when the global error minimum

has been discovered. Even with the incremental teaching technique, the resultant cluster diagrams were never completely concordant, and the one used in the analysis was the one seemingly most successful. With no objective method of selecting the most accurate diagram, an element of choosing the one which best fits expectations is introduced. With many different diagrams, it can be expected that some of them will by chance contain clusters seemingly indicative of category structure.

As already mentioned, the use of two large datasets, one for training the other for testing, will eliminate a lot of these problems. Another possible extension to this novel approach of humpback song analysis would be to look for evidence of patterns in higher levels of the hierarchy, such as the occurrence of themes.

5. Summary

Words in a language are not positioned randomly, but in accordance to strict syntactic rules governing which word types can follow which other word types. Words that behave similarly can thus be deduced to belong to the same class. This lexical category structure is indicative of grammar, along with long-distance dependencies that reveal the existence of hierarchical recursiveness in the language. Elman (1992) showed that a simple recursive NN could detect and represent both of these indicators of grammar. Only lexical category structure was looked for in this study, and the hierarchical and logical structure of the cluster diagram of just Theme I, and the entire song to a lesser extent, suggests the presence of lexical categories. Although the result was not perfect, it still constitutes an indication of the existence of grammar within the songs of the humpback whale. Grammar is an attribute of language, but not simpler communicative signals, and so this result supports the hypothesis that the primary function of the song is language.

Scope for extension includes the use of two large datasets from the same whale and year to address the problem of memorization of the entire training set, and looking for patterns in higher levels of the song hierarchy.

Bibliography

1. Ball, P. Sounding out the science of whale song. *Nature* (1999).
2. Bechtel, W. & Abrahamsen, A. *Connectionism and the Mind* (Blackwell, 1997).
3. Bradbury, J. W. & Vehrencamp, S. L. *Principles of Animal Communication* (1998).
4. Churchland, P. M. *The Engine of Reason, the Seat of the Soul* (MIT Press, 1999).

Elman, J. Generalization, simple recurrent networks, and the emergence of structure. *Proceedings of the 20th Annual Conference of the Cognitive Science Society.* (1998).

6. Elman, J. L. Finding Structure in Time. *Cognitive Science* **14**, 179 - 211 (1990).
7. Elman, J. L. in *Connectionism: Theory and Practice* (ed. Davis, S.) 138 - 178 (OUP, 1992).

lman, J. L. in *Mind as Motion: Explorations in the Dynamics of Cognition* (eds. Port, R. F. & Gelder, T. v.) 195-223 (MIT Press, 1995).

9. Elman, J. L. in *The emergence of language* (ed. MacWhinney, B.) (1999).

). Elman, J. L., Hare, M. & Daugherty, K. G. Default Generalization in Connectionist Networks. *Language and Cognitive Processes* **10**, 601 - 630 (1995).

11. Fiske-Harrison, A. in *Finacial Times Weekend* 1, 3 (2001).

2. Frankel, A. S. Comparison of Alaskan and Hawaiian humpback whale song at the song-unit level. *Journal of the Acoustic Society of America* **108**, 2634 (2000).

Frazer, L. N. & III, E. M. A Sonar Model for Humpback Whale Song. *IEEE Journal of Oceanic Engineering* **25**, 160

- 182 (2000).
4. Helweg, D. A., Herman, L. M., Yamamoto, S. & Forestell, P. H. Comparison of Songs of Humpback Whales (*Megaptera Novaeangliae*) Recorded in Japan, Hawaii, and Mexico during the Winter of 1989. *Scientific Reports of Cetacean Research*, 1 - 20 (1990).
15. Johnson-Laird, P. N. *The Computer and the Mind* (Fontana Press, 1988).
McLeod, P., Plunkett, K. & Rolls, E. T. *Introduction to Connectionist Modelling of Cognitive Processes* (OUP, 1998).
17. Mercado III, E. (University of HI, Honolulu, 1998).
8. Mercado III, E. & Kuh, A. Classification of Humpback Whale Vocalizations Using a Self-Organizing Neural Network. *Proceedings of IJCNN'98*, 1584-1589 (1998).
9. Mercado III, E., Michalopoulou, Z.-H. & Frazer, L. N. A Possible Relationship Between Waveguide Properties and Bandwith Utilization in Humpback Whales. *IEEE Conference Proceedings*, 1743 - 1447 (2000).
20. Midgley, M. *Beast and Man: The Roots of Human Nature* (Routledge, 1995).
Payne, K., Tyack, P. & Payne, R. in *Communication and Behaviour in Whales* 117-118 (Washington DC, 1983).
22. Payne, R. S. & McVay, S. Songs of Humpback Whales. *Science* **173**, 585 - 597 (1971).
23. Pinker, S. *The Language Instinct* (Penguin Books, 1994).
24. Pinker, S. *Words and Rules* (Weidenfeld and Nicoloson, 1999).
25. Plunkett, K. & Elman, J. L. *Exercises in Rethinking Innateness* (MIT Press, 1997).
26. Quine, W. V. O. in *From a Logical Point of View* (Harvard University Press, 1953).
27. Seife, C. Deep Message. *New Scientist* (1999).
3. Suzuki, R., Buck, J. R. & Tyack, P. L. Information entropy of humpback whale song. *Journal of the Acoustic Society of America* **105**, 1048 (1999).
29. Vehrencamp, S & Bradbury J. *Principles of Animal Communication* (1998)
30. Waal, F. d. *The Ape and the Sushi Master* (Penguin Books, 2001).

Websites

Environmental Investigation Agency*, A Review of the Impact of Anthropogenic Noise on Cetaceans,* http://www.eiainternational.org/Campaigns/Cetaceans/.../noise.htm
Bioacoustics Research Program, Cornell Lab of Ornithology, *Effects of Human-made Sound on Behaviour of Whales,*
*h*ttp://www.ornith.cornell.edu/brp/LFAhb.html
Bioacoustics Research Program, Cornell Lab of Ornithology, *Humpback Whale Vocalisations,* http://www.ornith.cornell.edu/brp/WhaleSoundsHB.html
Natural Resources Defense Council, *The Proliferation of Undersea Noise,*
http://www.nrdc.org/wildlife/marine/sound/chap1.asp
Fisher, Robert, *Automatic Inference of Humpack Whalesong Grammar*, http://www.dai.ed.ac.uk/homes/rbf/WHALES/warinentry.html
Pacific Whale Foundation, *Hawaiian Humpback Whale Study, h*ttp://www.pacificwhale.org/learn/humpback.html
The Dolphin Institute, *Life Histories of Humpback Whales,* http://www.dolphin-institute.org/whale/research/lifehistories.html
South Pacific Humpback Whale Project, http://www.whalewatch.co.nz/south.htm
Robinson, Tony, *Practical application of the short-term Fourier transform,* http://svr-www.eng.cam.ac.uk/~ajr/SpeechAnalysis/node30.html
Robinson, Tony, *Windowing,* http://svr-www.eng.cam.ac.uk/~ajr/SpeechAnalysis/node26.html
SURTASS, *Acoustic Terms and Definitions*, http://www.surtass-lfa-eis.com/UAT&D/index.htm
Tabbott, M, *Information Theory*, http://www.amherst.edu/~mtabbott/current_projects/infotheory.html