DAVID M. W. POWERS

ROBOT BABIES

What can they teach us about language acquisition?

1. INTRODUCTION

Traditional Artificial Intelligence like traditional and modern Linguistics and Computational Linguistics tends to treat language as an isolated phenomenon, and syntax and semantics as separate areas of study. The approach taken to developing grammars, parsers and natural language systems has been more like the dissection of a cadaver than a study of live interactions in a complex ecology: the basic anatomical structure can be discovered and some educated guesses made about the physiological interactions, but there is no hope at all of understanding the functional operation of a system that involves multiple individuals living in a complex environment. In fact, Natural Language research has been even more hampered, as computer scientists ignored the collective wisdom of Linguistics, Psychology and Sociology, and relied on their own intuitions rather than real world data about language in action.

This is the context of the 25-year odyssey into Computational Language Learning that has led the present author to an increasingly ecological approach to the study of language acquisition and the engineering of learning systems. Each step along this path has involved the rejection of basic, if often implicit, assumptions of one or more of the cognitive sciences. Indeed, the emergence of Cognitive Science as a coherent field in its own right is itself a precursor of an ecological approach, recognizing that the complex interactions involved in cognition, and in particular language and learning, cross many disciplinary boundaries.

At the core of Cognitive Science is the further principle that theories of perception and cognition, and in particular language and learning, must be computationally viable. That is, the rules, processes and mechanisms proposed should be capable of effective realization by a computer of reasonable size and power, and preferably be capable of mapping to a neural model consistent with known neuroanatomy and neurophysiology: these computer implementations should be capable of generating predictions which are verifiable using empirical techniques from any of the cognitive and behavioural sciences, under naturally occurring as well as experimental conditions.

Some of the assumptions that have had to be abandoned include: that phonology, morphology and syntax are distinct; that syntax is independent of semantics and ontology; that there is a universal grammar; that while lexicon and grammar are

T.H.E. Editor(s) (ed.), Book title, 1—6.

finite language is infinite; that semantics and the lexicon can be acquired by learning but grammar cannot; that grammaticality is absolute and probabilities pertain to performance, not competence; that closed classes/functional forms are learned late. In the shift of focus from language as a closed system to the complex interrelationships that constrain its acquisition in an ecological setting, we have also seen than an adequate identification of the interactional 'context' of learning is fundamental to a formal analysis of learnability.

2. FIRST STEPS

The starting point for the Robot Baby research program was the rejection of the idea that a human, computer scientist or linguistist, could devise an adequate model of language. We have powerful computers, so why not let our computers analyze the data and come up with their own accurate grammars, morphologies and phonologies? Why not get the computer to learn like a baby does?

2.1 Learning to learn

An unwarranted assumption: there is a universal grammar.

This formulation of the problem is clearly somewhat naïve, as certain assumptions must be made about the kind of structures we look for, or the kind of innate mechanisms we postulate. The basic idea of constituent structure was assumed, although it was not expected that we would be learning grammars that fitted neatly into the Chomsky hierarchy, and it was absolutely clear that unparsimonious theories such as Transformational Grammar and Universal Grammar were not physically realizable.

On the other hand, it was expected that some sort of Feature Grammar would emerge, and Pike's Tagmemic Grammar [Pike,1977] was of particular interest as he had developed it as a generalization of his theory of Phonemics [Pike,1949], and had shown how it generalized further to a theory of human behaviour [Pike,1967]. This was the precise antithesis of modularity, and the simple methods for discovering Contrast in Identical or Analogous Environments (CIE/CAE) and Complementary Distribution (CD) provided a simple computationally and biologically plausible mechanism for learning about language and behaviour. Other structuralist ideas that preceded Chomsky, e.g. [Harris, 1960], also seemed to hold considerable promise as computationally viable mechanisms, providing they could be expanded to encompass a broader sensory-motor understanding of language - what we now refer to as an ecological approach.

A very similar constructivist mechanism had also been proposed by Piaget [1955,1971]: a 'sticky mirror' model whereby recognizers for particular features or percepts or concepts allow the interrelationships between them to become apparent, and the new recognizers to emerge for common relational patterns. In Piaget's terms, the reflected relationships became a reflection – they stuck to the mirror and became conceptual building blocks in the representation of more complex concepts.

Piaget's model was overtly ecological in nature, and language was only one aspect of the child development he was seeking to explain. This approach is moreover highly consistent with a self-organizing connectionist model, and that of von der Malsburg [1973] was particularly appealing as it demonstrated the self-organization of patterns (basic line/angles) that were well attested in studies of the visual cortex. This connectionist model provided a mechanism to formalize the notions of contrast and similarity that underlie both Pike's and Piaget's theories in terms of basic neural mechanisms of cooperation and competition, excitation and inhibition. Nearby neurons in a layer tend to excite each other: they "cooperate" and therefore tend to recognize/learn similar things. Neurons further apart tend to "compete" and inhibit each other and thus recognize/learn contrasting concepts. The gradual transition from excitation to inhibition as a function of distance, the lateral interaction function, gives rise to a graded series of recognizers for intermediate concepts.

In the original von der Malsburg experiments, the neurons at the distance where inhibition was maximum recognized lines that were mutually perpendicular – e.g. vertical and horizontal – and the neurons in between recognized the angles in between – e.g. reflecting similarity with vertical and horizontal in proportion to the level of excitation or inhibition at the respective distances. A related and more well known but less neurobiologically motivated approach is that of Kohonen [1982].

These influences led to six batteries of experiments based on different initial assumptions and different basic mechanisms. Most of these focussed solely on the learning of phrase structure constituents, but one applied the learning mechanisms to parsing both sentences and ontological representations in a simple hand-simulated robot world [Powers, 1983; 1984]. We will look first at the results of the experiments that focussed purely on a textual stream of words – which produced results that were quite surprising at the time, forcing us to take a more ecological perspective. Subsequently we will look at the first seeds of the robot baby project. But the most remarkable thing about these experiments is the way they led away from standard theories of grammar.

Word level results and their implications

Two of the pure grammar learning experiments are of particular interest at this point: one was based on a purely statistical learning mechanism that weighed hypotheses as to whether a unit should combine to the left or the right, and whether it should combine with a word or a phrase; the other was based on a von der Malsburg network augmented with delays and decays to introduce a time element and allow representation of sequence. In both of these models, the more a recognized structure occurred, the more useful it was deemed to be, and the higher the value of associating with it to form a new constituent: in the neural net this was implicit whereas in the statistical model it was explicit.

In developing both of these models, we also incorporated the idea of tight cognitive restrictions. In the connectionist model, the decay parameter filled this role – e.g. after two cycles, a neuron's activation might have halved. In the statistical model, we introduced an explicit parameter representing "the magical number

seven" [Miller,1956], maintaining the seven most likely candidate structures for a recent subsequence of the words seen so far. Whenever a new parse tree was needed to incorporate the next word, the least likely of the seven stored putative partial parse trees was dropped and a new composite tree added. The proposed new parse tree would also be examined to see whether it combined usefully with stored parse trees that adjoined, and would again supplant a stored parse tree if its utility was calculated as being higher. This reflects closely the way in which independent parse trees (e.g. for a noun phrase and a verb phrase) for adjacent sequences of words are joined into a full parse tree in a traditional approach.

Both of these models succeeded in learning to parse small phrases and clauses hierarchically, but proved to be extremely limited and quite unreliable as utterance length increased. But what was interesting was that in both cases the end-of-sentence punctuation was classified first, then articles, then sequences of closing punctuation followed by an article, then a structure in which that combination was combined with a following noun, essentially recognizing the subject of the sentence.

This rather strange construct seemed disappointing at first. We had been seeing the open class, 'content' words as the keys to language, and had indeed also experimented with learning to parse telegraphic sentences – without much success. But here, instead of recognizing the noun or verb as head of a phrase or clause, and finally augmenting with those pesky 'function' words, we found that it was the closed class, function words that were the seeds around which the crystalline constituent structures grew. We had to revisit one of our implicit assumptions concerning the relative importance of open and closed class words in learning.

This finding also meshes in with the perennial suggestions that articles may be the head of the noun phrase [e.g. Hewson, 1991]. It also represents empirical evidence of a bias in traditional linguistics and psycholinguistics due to lack of ecological perspective. The role of these functional words in our purely structural analysis has nothing to do with the particular semantics of the word. It has solely to do with the relatively high frequency and low entropy of these words, and this is what makes them useful as pointers to the open class words that follow them and in predicting the category of these associated words, but it has ecological implications.

The occurrence of a specifier – an article, possessive adjective or demonstrative adjective – is a strong indicator that the following words will identify concrete entities in the environment. The following word is most likely a noun, but an adjective such as a size or colour description is also very common and would usually be followed directly by a noun in child-directed speech. The specifier therefore has a dual deictic role, a sentence-internal role alerting to the kind of word expected, and an associated ecological role alerting the child to look for a concrete referent to associate with the following word. Note that the pronominal class, a closed class that shares words forms with the specifier class, has the ecological deictic function but is marked for completion of the utterance to the extent that a mutually exclusive form is employed. In the case of a form that can be used adjectivally or pronominally, such as 'that', the syntactic role is distinguished prosodically, viz. the demonstrative pronoun has high stress, a rising intonation, etc. In writing, it will usually be followed by another closed class form, viz. a punctuation mark reflecting this prosody.

An unwarranted assumption: closed classes are learned late.

A fundamental bias arises in our interpretation of children's speech due to the relative difficulty of assessing a child's comprehension of adult speech in its ecological context. When we gloss the words of a child, we tend to associate them with open class words, primarily nouns and secondarily verbs. But the child's word is frequently acknowledged to act more like a sentence, referring to the whole scene or desire, and our interpretation of the word as noun or verb or something else may be mediated by accidental resemblance to as much as deliberate emulation of a word.

The interpretation of child utterances is very subjective, and whilst the mother is typically aware of much of the ecological context of the utterance both contemporaneous and historical, the psycholinguist usually has a far more limited picture and is likely to be more focussed on transcribing the sentence than understanding the scene. Similarly in interpreting a child's response to an utterance, the observer must take into account the full range of linguistic and non-linguistic cues and has no direct evidence as to which aspects the child is attending to, or which aspects of a sentence he is recognizing and responding to.

Our results lead to a number of predictions which have the potential to be verified or refuted empirically: first, we would expect to see closed class or functional usages being produced at an early stage; second, we would expect to see evidence that the child is recognizing closed class words at an even earlier stage; third, we would expect to see that the child's comprehension of an utterance is enhanced by the presence of these words; fourth, we would expect to find that the general structural, prosodic and deictic role played by these words is more significant than their precise meaning or specific identity.

The first syllabic sounds a child makes (typically /ma/, /na/, /da/, /ba/, /pa/, /ta/, /ka/ - especially reduplicated) are universally associated with members of the family and other events/objects that are particularly salient to the child, and the associations are clearly encouraged by us (the person 'named' especially). However, many of the usages are primarily deictic in nature, and are aimed at attract attention and accompanied by appropriate gestures (e.g. /da/ with pointing, /na/ with looking under a chair at a fallen object) and there is some evidence that these earliest protowords are generalized deictics or prepositions.

For my own daughter, /na/ represented 'in' when she was out, 'out' when she was in, 'under' when the ball went under something, etc. This was the first consistently used protoword, and although it has some characteristics of an imperative verb (it was usually produced with an imperative or interrogative tone), it could also be interpreted as a kind of generic preposition. Similarly /da/ accompanied pointing related to interest and attention-directing behaviour, typically aimed at a concrete object, but is it a noun or a demonstrative? Conversely /nana/ represented food, particularly perhaps her mashed banana favourite - or does it mean eat? However /mama/, /dada/ and the like came later. The first name of a person she reproduced was Ann, a visitor for a few days. Shortly afterwards a new visitor named John was also /æn/! It is well known that homing in on the right level on generality is one of the hardest things a child has to do, but of the first three monosyllabic 'words', only this last is at all like a noun or a name, although it is

clearly generic in nature. In fact all three are generic in nature, so perhaps none of them is truly an open class word!

Although we are here focussing on our interpretation of the child's first words, the child's language ability is over a year old by this point. Already prior to birth the child recognizes and responds to the mother's voice. At birth, even very premature birth, the child differentiates between his mother tongue (literally) and other languages [Mehler et al., 1992]. There is also recent evidence of newborns having the capability to make functional/lexical distinctions [Shi,1999]!

Well before the child's first words, comprehension is seen to be better for full sentences than for telegraphic sentences that omit the closed class forms and disturb the prosody. It would seem that the rhythm of the sentence and the closed class forms play a kind of sentence-internal deictic role even then. They alert the child to where the words/morphemes are that might correspond to external stimuli, objects, colours, activities, locations. They are very frequent, and indeed characteristic of the language. They are also integral to the syntactic structure of the language, along with affixes and other forms of derivational morphology.

Empirical evidence: the utility of closed class words

Why have closed class words been neglected in earlier research into language learning? We have cited some evidence that they are recognized early, and there is room for considerable further exploration. Indeed, it is very easy to discover these characteristic closed-class words and inflections, and it is worth considering how useful they are to the language learner. We do not explore here the evidence that they are utilized by the adult language user, but this evidence goes back over a hundred years to Huey's work studying reader's eye movements in the 1890s [Powers,1989]! Rather we focus on understanding how they may be useful to the language learner and the language user from an ecological perspective.

The robot baby project uses two classes of algorithms with both implicit and explicit assumptions about closed class words. Our earliest algorithms made no assumptions about the existence of open and closed classes, but closed classes of words emerged first and acted as seeds around which larger phrases and clauses were built [Powers, 1983-4]. Generalizing across linguistic levels, emergent closed classes include the vowels as well as the articles [Powers, 1991-2]. The first of these subsequent algorithms were deliberately designed to bias for small classes of high frequency elements that provided strong structural cues, but these closed classes were still essentially emergent. Other experiments have explicitly examined how parsing can be carried out solely on the basis of these kinds of classes - the open class information is thrown away entirely, and parsing is completed using only the closed class words and affixes [Entwisle and Groves, 1994].

We have also found closed classes to be of considerable utility in devising a grammar corrector that is capable of correction of over 18000 distinct confused word errors in unseen text, with only a 0.4% rate of error introduction (e.g. 76% of random 'from'/'form' substitutions are found and corrected). This system is based on unsupervised training on five years of the Wall Street Journal using only statistics on closed class information in the immediate context [Huang & Powers, 2001] - as

discovered automatically using the techniques outlined in this chapter [Powers, 1989, 1991, 1992, 1997b].

If we consider the early and very common deictic use of /da/ glossed 'there', we note that our gloss includes the relatively rare and difficult /dh/ phoneme. This is rare across languages, but is characteristic of English and difficult for both native and non-native speakers to master (children learn its correct pronunciation very late). Most importantly its word initial usage *exclusively* marks closed class words (the, this, that, these, those, there, they, then, thus, thee + derivatives) the most frequent of which would all be part of the broader compass of /da/. In German it is /d/ that has this role, and in French it is /l/, in each case covering both the 'there' gloss and the articles, and giving rise to the characteristic sound of the language. Anywhere we hear this closed-class deictic marker, we are likely to have our attention directed at an object, and the following stressed word is likely to mark that object. On the negative side, the pseudo-deictic may not be stressed ('the dog') and even the deictic may not be (e.g. in French it combines with 'voir': 'Voilà un chien!). On the positive side it is often duplicated ('Là! Voilà le chien!' – triplicated here!) and the words that capture attention like ' look' vɔ̃ir/regarder', 'gucken') will also be associated with the deictic function and become frozen into attention-drawing phrases ('Look!', 'Guck mal!').

These examples of the deictic nature of closed class words emphasize the essential role of ecological information in their interpretation. The open class words have a semantics associated with them before taking into account the ecological context, but the closed class words have none. The meaning of a closed class word is almost exclusively derived from the *present* ecological context. The meaning of a specific closed class word incorporates the whole ecological history of the word and the relation to the present ecological context may be relatively indirect. Generic words (like 'person', 'thing', 'do', 'go') act more like closed class words than content words in that they must pick up their precise meaning from the context, and the fact that they have been used implies that no more specific term is available. These words may also pick up specific functional roles, e.g. in compounds ('chairperson', 'everything') or as auxiliaries or modals (e.g. 'I do want to', 'I am going to') and such a word thus tends to form a highly idiosyncratic closed class and will typically be the sole member.

Thus the assumption that open class words are learned before functional forms seems to be biased by our preconceptions regarding their lack of ecological utility, an emphasis on role rather than form, and a focus on production rather than comprehension, all compounded by the difficulty of assessing comprehension and our lack of accuracy in recognizing exactly what was intended or just what was understood.

An unwarranted assumption: phonology, morphology and syntax are distinct.

The above experiments were inspired by and based on algorithms that originated in Pike's phonemic analysis, Piaget's ecological study of language development and von der Malsburg networks for self-organization of visual features. This suggests that the Fodorian and Chomskian ideas of modularity should also be discarded as

assumptions - it may be that physical localities emerge, but the general cognitive mechanisms we are exploring discover useful structure in many different contexts and thus phonological, morphological and syntactic distinctions would seem to be at best emergent phenomena, if they have even that level of separate reality.

In the next section we will look at how these mechanisms have been applied at lower levels and discover hierarchical structure that cross all three levels of linguistic structure, discovering phonological, morphological and syntactic structure and classes. In fact, we not only applied them at the phonological level where the basic CIE/CAE idea originated, we applied them at the level of English, French, German, Dutch and Esperanto orthography, where the connection to phonology is somewhat tenuous, and we applied them to raw speech code vectors, that is the sets of frequencies present in each 50ms of a speech signal - with virtually identical results in each case.

In the section that follows this discussion of the phoneme/speech level work, we consider the broader application of these mechanisms in an ecological sensory-motor context in learning basic ontological and semantic concepts, in separating sound from different sources, in speech recognition and in speech reading.

Parsimony tells us that the simpler theory is best. Why have separate modules and mechanisms for phonology, morphology, syntax and semantics when one will do? We accept that it is quite likely that specialized mechanisms will be needed for specialized tasks, but for the moment we are concentrating on seeing how far we can get with our simple cognitive mechanisms inspired by CIE/CAE and connectionist self-organization, as well as some closely related correlates of these in statistics and machine learning.

We have also discovered another compelling reason why we don't have to assume different mechanisms for different aspects of language. The learning of basic structure at all of these levels seems to be very robust and largely independent of precisely which variant algorithm or mechanism we use. Whereas we initially demonstrated useful learning of linguistic structure with two different alternative mechanisms (a simple statistical grammar approach and a self-organizing connectionist approach) we have now demonstrated useful structural learning with of the order of a thousand different algorithms and variations [Powers, 1997].

Both the statistical and the learning models we have used (and indeed all the other algorithms we have successfully tested) reflect directly or indirectly occurrence probabilities conditioned on their context - that is they are overtly ecological in nature, and later we examine how they operate in simulated and real ecologies. Unlike the Markov Models popular in speech recognition, both the left and the right linguistic contexts, the future as well as the past, are able to influence the choice of the best structure. In fact, in the experiments we will describe in the next section, we found that three segments from right context (the future) and two from the left context (the past) gave us the most convincing grammars. We predict that the broader ecological context will further enhance our learning of syntactic structure as well as enabling learning of semantics.

An unwarranted assumption: grammaticality is absolute.

Although the original concept in the design of the algorithms was that there would be multiple hypothesized parses for each sentence, it was assumed that in general there was one correct parse and that this would be reflected by majority opinion – the most frequent grammar rules would be the correct ones and define the correct parse. Clearly this assumption is at best a simplification, given the existence of ambiguous sentences even in the absence of overt homonymy. Take 'Visiting professors can be dull!' which is grammatically ambiguous as to who is doing the visiting, but would be ecologically unambiguous in almost every conceivable context: was it pronounced by a student after enforced attendance at a seminar (more likely situation) or by a husband to his wife following a visit to some new neighbours?

The grammars derived by most of the algorithms we are using have probabilistic information associated with them by the algorithm. These probabilities are automatically conditioned on the immediate context of roughly seven word/phrase-level chunks, but could equally well be conditioned on a history of larger chunks of context, both linguistic and ecological. Even those approaches that aren' t explicitly probabilistic in nature, like CAE/CIE, have the implicit idea that some errors and exceptions can occur and so there is some threshold that must be achieved before a rule can be accepted

Rejecting the notion of the absolute nature of grammar leads to a different concept of parsing, in which the object is to allow every conceivable parse of the sentence and the expectation is that every sentence will have multiple parses. The constraint parser of Entwisle and Groves [1994] operates just like this. The closed class and inflectional information is used to specify a set of possible roles for each word, and the null inflection in general corresponds to a word that could be used as a noun, a verb or an adjective.

In English, the rule is that any open class word may be pressed into service as any of these parts of speech except that this becomes less likely/acceptable when there are alternate, irregular or derived forms that take on these roles (and sometimes English marks this derivationally, e.g. with voicing, length, stress and/or spelling – 'advice/advise', 'practice/practise'). For example, in the absence of a specific verb for an action undertaken with a body part, the noun may be pressed into service: 'shouldered', 'headed', 'elbowed', 'kneed', 'nosed' vs 'foot/kick' and contrasting further with 'ear/hear', 'nose/smell' and 'tongue/taste'. Note that distinct words tend to be used for the *function* of an organ, whilst the incidental usages of parts of the bodies for purposes they weren't designed for allow the nouns themselves to be employed as verbs. This productive use of language belies the traditional lexicon-based approach where words are assumed to have a specific part of speech.

Entwisle [1997] notes further that even apparently unambiguous sentences tend to have alternate readings that will be picked up by the constraint parser and require some thought to tease out as being possible – often involving such extended uses of a word. His favourite example, 'The red eyes water!', is short and contrived but illustrates the point well: until you consider the gloss 'communist' for 'red', it is difficult to see an alternate parse, notwithstanding that this sentence uses the less common verbal use of 'water' that is unavailable to most parsers we have tested. The average level of ambiguity seems to average close to three readings per clause

in experiments on War and Peace, Alice's Adventures in Wonderland and the Wall Street Journal.

It is important to note that we are not claiming that grammar, morphology or phonology is overtly *probabilistic* in nature - that rules sometimes apply or sometimes don' t in some arbitrary way. Rather we are saying there is a set of things that are *possible* and a set of things that are not - a sharp *possibilistic* distinction rather than a fuzzy *probabilistic* one. The appropriate alternative is selected for pragmatic and ecological reasons. Entwisle discusses several examples where a sentence appears ungrammatical until a possible meaning is identified, and many alternate parses that his system throws up have this same character - they are not seen as grammatical until glosses are added that make these alternate readings make sense. Deane [1992] similarly discusses a number of examples where entrenchment [Silverstein, 1976] makes the difference between whether sentences are seen as grammatical or not.

Our full ecological learning model can be visualized in terms of multiple series of blackboards as in a mid-20th century lecture theatre: at the front of the theatre there are columns of blackboards that may be raised or lowered or rolled vertically. In our model, one column of blackboards might correspond to the word-level input and hypotheses; others would correspond to the different sensory-motor modalities associated with vision, audition, movement, etc; yet others would correspond to short-term memory and associations relating to the total ecological context – partly filled in by direct sensory-motor experience, partly filled in from long-term memory, and partly filled in from linguistic information. As blackboards are filled they are pushed up, till eventually they are recycled. The learning algorithms have seamless access to everything on the blackboard, whether it originated in vision, audition, touch or motor feedback, whether in the immediate past, in the present, or in some tentative construction or expectation based on these. Neurophysiologically, the blackboards might correspond to layers of the cortex (and possibly other areas of the brain) that receive a certain kind of input.

In the next series of experiments we examine how by dropping down below word level, we find that the same basic idea as we used in our word level experiments also operates at character, phonological and speech levels, and that all of the closed class and inflectional/derivational information used by Entwisle for constraint parsing of unrestricted sentences can be automatically discovered.

2.2 Learning to hear and understand!

The problem of dealing with text is in some ways much simpler than dealing with real world linguistic input. In particular, written languages are already segmented into characters and/or words, as well as larger chunks including phrases, clauses, sentences and paragraphs. The experiments described so far have been performed on text, working in European languages written as words, although in our discussion of certain assumptions we have hinted at successful results from experiments at other levels. In word level text, the issues of phonology have been dealt with, and the text is basically phonemic. In some languages, issues of homophony have also

been dealt with, with different spellings teasing apart the homophones of English and the grammatically inflected but similarly pronounced verb forms of French.

On the other hand, perhaps the orthography is doing us a disservice. There is no consistent definition of a word that applies across all languages, and even English and German have quite different concepts of what a word is, at least in terms of how the spaces are placed. The written sentence also has quite a different character to the spoken sentence, and the reader can use the page as a memory device in comprehending complex sentences. In some languages and styles of writing, it is quite common to have extremely long sentences containing a dozen or more clauses. In the spoken language, recapping techniques would typically be used to assist the reader with complex material, although it may not formally obey the standards of grammar prescribed for the written word.

This is well illustrated with the multiple translations of the Bible, translated from the original languages into multiple registers in hundreds of languages. Some individual sentences in the Greek may extend for many verses, and the more colloquial the language the more this will be split up into simpler units. The way clauses constitute a sentences in one version will correspond to a similar structure in which sentences constitute one or more paragraphs in another.

Why is it that we give such a special place in linguistics to the word and the sentence, when both are hard to define for even one language, and such differences in usage exist across languages and genres? The prescriptive spellings and stylistics that go with them cause problems for human language learners and computer speech recognition alike.

The next logical step in our research was to see what our learning algorithms could do with speech input, phonetic or phonemic input, and character level input. Given that Pike's Phonemics [1949] was one of the inspirations for the methodology, which is closely related to looking for CAE and CD, we were hopeful that some interesting results would be obtained, and indeed predicted that the vowel class should be one of the first to emerge.

Character level results and their implications

The character level experiments initially used a fairly direct implementation of the CAE/CD procedure [Powers, 1991]. The sets of local contexts for individual characters or small groups of characters were collected – on the basis that there is a decision about segmentation to make too! This way diphthongs and other digraphs (like 'th' in English) or trigraphs (like 'sch' in German) can be treated as phonemes if that gives the algorithm the best analysis. If applied to a phonetic transcription, it allows for detection of mismatched segmentations and thus expects the more atomic variant to be used in cases of doubt. For example, in German, 'tcsh' is hotly debated as to whether it is one or two segments, and although my analysis may suggest it is a single segment, transcribing it as two allows the algorithm to make that decision objectively.

The next step was to find the set of segments that could occur in each context (CIE step), and then to look for contexts that accept more or less the same set of segments (CD step). This defines a class (such as vowel or liquid) of phonemes that

have a number of variant subsets defined by the different contexts. When applied to phonetic transcriptions this will include allophones, both those that occur in free variation and those that include in complementary distribution. The individual class variants will always tend to suffer from depletion due to possible combinations (e.g. words or morphs or syllables) that just do not happen to occur in the corpus or the language.

Interestingly, this approach handles segmentation (into useful units) and classification (into classes of units that behave similarly) automatically, unlike the original techniques proposed by Harris [1960] and exploited by a number of researchers in various guises and applications [Brent, 1997; Powers, 1997b].

There were a number of alternative versions of the algorithm explored, using different heuristics to decide how many differences there had to be between sets of segments before they were regarded as distinct classes, and how to weight the utility of a putative class (e.g. by number of distinct contexts or number of distinct occurrences). However, all variants showed a strong preference for the five standard vowels as the pre-eminent class in both English and French, and most of them also found the same five vowels in Dutch and German. Interestingly, the optimal context seemed to be two units of left context and three units of right context, with one or two segments in contrastive focus.

The work in Dutch also experimented with the specifying of hyphenation points, finding their inclusion actually reduced the effectiveness of hierarchical analysis; the work in French also experimented with phonetic input and speech input (in the form of speech code vectors) and found the vowels were resilient in both cases [Schifferdecker, 1994]; whilst related work in German experimented with different ways of representing the umlaut and scharfes-ess characters, e.g. using the 8-bit ISO representation (‘ä’), the ‘e’ representation (‘ae’) and the ‘quote’ representations (‘a"’), finding the five vowels were still relatively resilient [Powers, 1997a]. Another related technique applied to twenty different alphabetic languages [Boy,1977] failed to find all the vowels only in Russian, suggesting that an inappropriate segmentation had been enshrined in the Cyrillic orthography (there is both a palatalized and an unpalatalized set of vowels which should have been collapsed and distinguished with a separate ‘y’ segment in an orthogonal phonemic analysis). We are currently researching what the algorithm does with Chinese and Japanese texts in various representations. Note too that the Kohonen network has also been used to self-organize phonological structure in speech input [Ritter and Kohonen, 1990].

Once a strong class has been identified, the segments concerned can be replaced by a new class symbol and the process repeated, leading to the development of a context-free grammar that takes us from the character/phoneme level to the phrase/clause level, exhibiting intuitively appealing structure at each level. Note that the issue of characters/phonemes/segments that belonging to multiple classes has not been addressed here and some sort of context-sensitivity (based on the contexts of the variant classes) should probably be built in. For example, ‘y’ acts as a consonant in word-initial position, and can always be interpreted a vowel in word final position but after a vowel a consonant interpretation is also possible (‘by’ vs ‘play’), and it is clearly context-dependent medially (e.g. ‘played’ vs ‘playing’).

One further noteworthy property of this algorithm is that even when trained only on a dictionary (a list of unrepeated headwords), a sensible word-to-phrase/clause level constituent structure is still found (using a greedy algorithm that always rewrites the longest possible sequence in the event of ambiguity). After typically eight levels of agglomeration, the corpus is reduced to repetitions of a single symbol each of which tends to correspond to a Noun Phrase or a Verb Phrase or a Simple Clause [Powers, 1992]. Note that the use of the greedy algorithm is arbitrary, and that different structures may be appropriate in different ecological contexts, and that all ambiguity is actually preserved by our learned 'grammars'.

An entire family of closely related algorithms has also been investigated in which the yes/no character of CAE is replaced by a statistical distance measure and classes are formed in a vector space [Finch, 1993; Powers, 1997].

The first interesting thing to note about this set of experiments is that it vindicates the intuition that the vowel class, and other phonological classes (including the liquids, the nasals, etc.) have a similar characteristic to the closed classes we discovered at the word level, and form a similar substrate of language.

The following is typical of the first two classes found, starting from normal English text – note that it is discovering syllables from the inside out in order of decreasing sonority:

A <- a	B <- rA
A <- e	B <- Ar
A <- i	B <- lA
A <- o	B <- Al
A <- u	B <- A

The success of the hierarchical structure discovery, and the transfer between word and character levels, suggests that the distinctions between the phonological, morphological and syntactic levels are unfounded – the same basic algorithm works at both phoneme/character level and at morpheme/word level. Moreover, the generated character-level structure follows morpheme/syllable structure, groups clitics and affixes with the expected targets, and recovers omitted prosodic, segmentation and punctuation information in experiments in which spaces and punctuation are removed from a text before the algorithm is applied [Schifferdecker, 1994].

Another version of the system was run in such a way that it was possible to learn recursive rules and overlapping classes by keeping both the original segments and the new non-terminals and allowing the possibility of recursive segments involving a proposed new non-terminal. Generally speaking, the structures that emerged from these experiments were less convincing than those that emerged when recursion was prohibited. It is possible that this is due to the precise method that was used to allow recursion, but generally recursion introduces considerable difficulty and raises the spectre of formal learnability results that prove that a superfinite language cannot be learned.

Previously we called into question the rigid concepts of word and sentence as the boundaries of grammar, and pointed out that there was an overlap in the way

sentences and paragraphs were used to express complex concepts. If we think of a paragraph or a monologue as a sequence of sentences, with the odd bit of connective glue thrown in, why not think of a sentence as a sequence of clauses, with the odd bit of connective glue thrown in.

This is the way our non-recursive grammars tend to parse a sentence: we get a sequence of basic NPs, VPs and Clauses that all end up with the same non-terminal symbol. It has been observed that certain closed class elements can be interpreted as switching a kind of finiteness feature on and off: thus 'to' can switch something verb-like into something noun-like, and 'that' and '-ing' have similar effects. Conversely, a finite form (e.g. copular or verb) attaches to a noun-like object and makes it into a verb-like phrase, and apposition is essentially the default in this model. This is the kind of switching that looks like it might be happening in some of the grammars generated.

This leads to an iterative model of language rather than a recursive model. Various phrase or clause like entities are strung together with a variety of connectives including verbs, particles and prepositions. We envisage turning our structural parse into a cohesive grammar by allowing the closed class elements to act as features whose occurrence is restricted by higher level constraints. Our constraint parser [Entwisle and Groves, 1994; Entwisle, 1997] operates in precisely this way, inducing constraints based purely on the closed class information that can be derived automatically using the CAE/CD techniques described here.

This research therefore calls into question the entire modular tradition of linguistics. Phonology, Morphology and Syntax emerge seamlessly and demonstrate amenability to the same methodology. Even the distinction between Syntax and Semantics is called into question as we treat language as an emergent property of basic cognitive or connectionist mechanisms that apply equally to all sensory-motor processing. Even the idea that language is recursive and sentences unbounded – an infinite set of sentences generated by a finite grammar and lexicon – is now seen as highly unlikely due to the better performance of the non-recursive model, which is consistent with the fact that true recursion (viz. context-free complexity or above as opposed to regularity which corresponds to iteration and tail-recursion in power) requires an infinite stack and thus an infinite head.

Learnability results: the impossibility of learning a grammar

Our previous discussion suggests that learning grammar is no more difficult than learning semantics, and could even be the easier part. Certainly relatively little progress has been made in learning true semantics, as opposed to the pseudo-semantics whereby words that occur together are grouped into so-called semantic classes. Segmentation into words is still one of the most difficult things for speech recognition systems – and not surprising either, given the arbitrary and inconsistent definitions of words as we noted above. It is clear that a grounded semantics can only be learned using multimodal information, and it may be that this information is also necessary for complete learning of morphology, lexicon and grammar as discussed earlier. However, the work discussed above demonstrates that a basic phonological, morphological and syntactic structure can be learned effortlessly by

simple unsupervised algorithms. The question that this begs is, is it the right structure? This is particularly pertinent given it seems to force us to abandon the idea of language being recursive. In fact, it may also be expedient to abandon the idea that there is a particular target language to learn.

It is beyond the scope of this chapter to display myriad self-organized grammars learned from various corpora for various languages [Powers, 1991, 1992, 1997]. Displaying them only encourages people to compare them with their favorite theory, and the focus of this chapter is the ecological aspects and not the syntactic. The proof of the pudding is not how much it looks like Linguist X's Grammar of Y, but how it works in practice, in an ecological context. We have already referred to the success of the classes discovered in a text correction application and a constraint parser, and the fact that it 'works' on input from speech code vectors, to phonetic transcriptions, to standard character- or word-based text, but the real question is whether the structure discovered lends itself to learning the semanatic relationships necessary for comprehension. In fact, one of the conclusions we feel forced to draw from this work is that it is *not* appropriate to target some kind of standard grammar for a standardized language.

An unwarranted assumption: poverty of the stimulus, lack of negative information.

There are a number of theoretical results about not being able to learn grammars that hinge on a set of implicit and explicit assumptions: that language is recursive and non-finite, that no sources of supervision or distribution information are available, and implicitly that we are seeking to learn a specific language based on the traditional word to sentence model of grammar without taking semantics or any other ontological or ecological factors into consideration [Gold, 1967]. Under these assumptions, there are always multiple grammars that can generate any corpus – e.g. the one that consists only of rules rewriting S as one of the sentences in the corpus, a reduced finite grammar, the correct one (whose very existence is itself an implicit assumption), as well as whole families of grammars that allow more complex (and unevidenced) recursive constructs. The negative results are no deeper than this. However, dropping any one of these assumptions destroys the theorems.

Supervision can involve simply a constraint on the order of presentation or some guarantee that every construction/rule will be used in a fixed amount of time, or various other probabilistic or distributional assumptions. The proof that a probablistic grammar could be learned came hot on the heels of Gold's result [Horner, 1969], but is far less well known. Even Gold's original paper [1967] included a little known/understood exception for 'anomalous text' – he showed that the entire class of recursive languages *could* be learned if distributional assumptions could be made about the order of presentation. Also supervision could involve any form of reinforcement, not just overt correction. This can include simply being understood as well as having the correct form used in a reflection, response, augmentation or clarification, or just in similar contexts. It could also involve comparing a generated form with a remembered form [Turk, 1984].

Supervision may thus be explicit or implicit, but the poverty of the stimulus evidence suggests that children receive very little explicit correction and that such

correction has very little effect. On the other hand implicit supervision arises directly from ecological context and is not necessarily linguistic in nature. The child's sounds may or may not achieve the desired effect – he may get food when he wants a cuddle. The child's attempt to push open a door may fail, whilst an attempt to pull succeeds – or vice-versa. Also, there is no reason to expect that the child's logical or linguistic powers are sufficient to make correct use of the negative information when first received, or after a single experience.

Turk [1984] proposed a theory of anticipated correction where the memory of previous experiences, positive and negative, provided implicit supervision. Thus the child's initial or even planned attempt is compared with remembered examples of successful behaviour (most likely on the part of the parent) or unsuccessful behaviour (most likely on the part of the child). There is also an effect whereby large advances from the present state of mastery seem not to be possible in a single giant leap, but a series of small steps interspersed with periods of consolidation is observed [Powers and Turk, 1989]. In our algorithms, if there are too many differences between two states, then we just say they are different and can do nothing about it. However, where there are very few differences (another magical parameter defines this limit as a difference in at most two parameters) we can see contexts as similar and analyze the differences both in the linguistic domain and the broader ecological context.

An unwarranted assumption: there is a target language a child is aiming to learn.

Let us now focus on another implicit assumption we keep alluding to. What makes us think there is a specific correct target language? Whose would that be? The mother's or the father's or the babysitter's or the teacher's?

Each individual has their own idiolect shaped by their own experience and even twins develop differences in their language (although they often first develop a private language that differs markedly from their parents'). So the assumptions underlying the language learnability results have two further potential holes: that we may have no particular target language, and that we are not so much learning as negotiating, or evolving, a language. But there is another reason why neither learning nor development captures the process, and an ecological process does: both the parent and the child adapt. The conventions adopted are not just those of the mother. The family picks up and uses expressions the children coin, and develops its unique family conventions, others develop in the peer groups at kindergarten, Sunday School, in the park. Word play and other games also play a role in this conventionalization process, and appear to be an integral part of the child's learning process [Kuczaj, 1983].

The assumptions about learning to criterion also fall down in that even the conscious targets of the accepted common language, including spelling and pronunciation, are often never acquired. People don't use nominative/accusative pronouns 'correctly' in conjunctions, and prescriptive correction has led to error-inversion (the unschooled say "me and my sister saw …", the schooled say "… saw my wife and I"). Certain mispronunciations/misreadings persist notwithstanding

recognition that they are incorrect (e.g. 'misled' read to rhyme with 'whistled' – even as an adult).

An unwarranted assumption: language is recursive and non-finite.

The robot baby project has succeeded in unsupervised learning of both recursive and non-recursive grammars, but there is no target grammar and there is no requirement of identification in the limit. Our algorithms require only very minor modifications to allow for learning recursive grammars: we simply allow a recognized class to be used as a recognizer (as Piaget proposed) without or prior to freezing it as a class. Thus sequences can be rewritten using a proposed class and then recognized as being a segment of a sequence that fits the same slot as other members of the class. However, generally the recursive grammars produced tend to be less intuitively appealing, if not actually unstable or degenerate. Of course modifications to the algorithm to produce more stable grammars are probably possible, and we along with a number of other researchers, have used techniques based on Minimum Description Length, as a formalization of Parsimony, that can learn a variety of recursive target languages [e.g. Grünwald, 1990].

Nonetheless, our point here is that it is not necessary to assume that language is recursive and indeed our algorithm performs better if we don' t. Furthermore, context-free languages (or worse) require an infinite stack to parse an arbitrary sentence, and since we don' t have infinite heads it is formally impossible that a human language is strictly context free. The aspects of language that seem to be truly unbounded recursive seem to be restricted to the iterative/regular class illustrated by the formation of numbers, simple noun phrases with arbitrary sequences of adjectives, and sequences of clauses, sentences or phrases - all of which sequences have the same grammatical relationship to their context as their atomic variants. Formally only such iterative constructs or their equivalent regular or tail-recursive grammars can give rise to unbounded utterances because this stateless iteration of a unit means that no stack is needed to store state. Ironically, although a context-free grammar does not depend on context on the left-hand side, it does have to store state concerning the right-hand side context so that it knows where to continue from on return from the recursion.

An unwarranted assumption: linguistic structure is independent of semantics.

In our experiments the grammar can and does change if the context changes. The classes and rules in our model only serve to identify units that seem to act similarly in relation to their context, whether syntactic or semantic. The purpose of syntax is patently to provide a framework for semantic interpretation, and it seems to be the cues that are important rather than the precise form of rules. Different runs and different algorithms can produce slightly different grammars, but this does not necessarily affect the utility from the perspective of semantic interpretation. However experiments to demonstrate this convincingly require us to build up a much larger corpus of speech in sensory-motor context. But the clear implication of our research to date is that ecological context is necessary not just for learning semantics, but for learning morphology, phonology and syntax.

Why do we need an ecological context and a concurrently learned ontology and semantics to learn linguistic structure?

Pike [1949] implicitly recognized this need for an ecological perspective in his formalization of the phonological analysis procedures. CIE and CAE require recognizing an ecological contrast as a basis for establishing if there should be a phonemic contrast. If there is no ecological contrast, the decision is that there is only phonetic contrast and allophonic variation. The ecological perspective was later recognized more explicitly [Pike, 1967] as he extended the analysis technique to the study of human behaviour.

Our experiments learn certain kinds of structure by making use of distributional information. However this structure is uninterpreted and currently peters out at the level of simple phrases and clauses, and we end up with a sequence of units that we have parsed but for which we have not determined the interrelationships. We are currently investigating automatically treating the higher frequency more closed class component of a constituent as a feature and looking to form associations, and for heavily declined languages there are easily learned associations. However there are some basic relationships that cannot be determined statistically in an ecological vacuum, e.g. which noun phrase is the subject and which is the object, or which is the landmark and which is the trajector; or worse still, which sequences correspond to the same object possibly in different relationships, and which correspond to different objects possibly in the same relationship. Pike's basic assumption was that we need to know when words are interpreted differently and when they are interpreted identically.

We have found no way past this ultimate dependence on ecological context, and rather see it as applying at every level, not just at the level of words. Furthermore, we see no advantage in trying to force the system to learn more detailed structural and cohesive relationships in an ecological vacuum, when to understand the use of language and to develop a true semantics we need to take account of the sensory-motor environment and the interaction of the learner with his world. Indeed even before coming to the conclusion that the independence of structure and semantics was an unwarranted assumption, we had been exploring the learning of syntax and semantics in the context of a simulated robot world [Hume,1984; Powers,1989].

2.3 Learning to move and feel and see!

Language learning experiments with robot babies, either in thought, computer simulation or mechanical implementation, go back at least three decades, along with the idea that for a computer to learn language it will need to be learned by a robot in a real environment rather than by an isolated computer. Indeed this was considered in the paper that originated the famous Turing Test half a century ago, and Turing believed that a computer that could pass the Turing Test would have to have access to “the best sensors money could buy” [Turing 1950]. Our own experiments go back over 20 years, but the majority have involved simulated robot worlds rather than actual robots.

In some ways, these early robotic experiments were premature, as the computational power required was unavailable and underestimated, and our understanding of Machine Learning and Artificial Neural Nets was not nearly so well developed. Nonetheless useful principles emerged, including the idea that there should be a strong correspondence between the sensory-motor capabilities of the robot and the language learning mechanism [Moulton and Robinson, 1981; Block et al., 1975; McCarthy et al. 1968]. In fact, Turing himself played with self-organizing processes very similar in character to those discussed above as well as playing an important role in defining the family of computational machines that correspond to the various members of the formal language hierarchy [1952,1936].

Until very recently, the robot in Natural Language experiments was usually a graphical simulation, if not a figment of the researcher's imagination (a thought experiment, or a simulation in which commands and representations were generated and parsed). Winograd[1973]'s famous language understanding robot arm, SHRDLU, was the first AI program to successfully explore natural language interaction with a graphically simulated robot. Even when a real robot existed it was often more convenient to carry out the more complex experiments with simulations. Even today, it is usually much more appropriate to run small modular experiments assuming particular kinds of inputs and examining the outputs, than to try for the supercomputer level of performance required to do everything at once. At the moment we have to use our imaginations to envisage how a total system would operate. But nonetheless some robots are being built for ecological experiments related to language and learning and occasional attempts are being made at full integration – though still not in real-time.

The robot 'babies' that have been built, range from a 2 metre giant and a disembodied head at MIT [Brooks et al. 1998], to a commercial explosion of robot-animal toys that have an elementary ability to learn or adapt. The smaller robots, including most of these animal- and baby-like robots, have the advantage that they can be brought up like a real baby and exposed to the same inputs as a real baby, to the extent that the perceptual system is up to it – and only now are such robots becoming feasible. Ideally, these robot babies will respond in a way that encourages and directs attention and interaction ('supervision'), in terms of gestures, expressions or words.

Another kind of language learning robots is more like cars or trucks or bulldozers [Steels, 1996-7]. These are interesting in a different way in that the goal is to study social evolution and in particular the invention of a communication system - rather than the learning of ours! In this case, the ecology is set up so that cooperation and communication are necessary for the robots to ' survive' [Brooks and Steels, 1994; Steels and Brooks, 1995].

Simulated language learning robots

Our first attempts to allow for the simultaneous learning of syntactic and semantic information involved contriving sequences of scenes that represented the key elements and meaning of a target sentence. The scene was represented in a constituent structure form that closely resembled a parse tree, and naturally this

biased towards learning phrase structure that corresponded closely to this representation. The visual representation included features such as RGB colour components, physical dimensions, and Cartesian coordinates of the centroid. The experiments were designed so that subtrees of the parse were distinguished from subtrees of the ontological representation only by the use of a modality label such as 'hear' (for the textual language input) or 'see' (for the simulated visual input). Relationships could now be learned between 'hear' subtrees (cohesive associations), between 'see' subtrees (ontological associations), and between 'hear' and 'see' subtrees (semantic associations), as well as between the associations themselves (which also permitted role associations). For example, the meaning of 'jumps' was abstracted as an association that related the word to a set of arithmetic constraints on the Cartesian coordinates of the associated subject (passive forms were not explored) and essentially was captured with a requirement of motion through three states with a linear relationship between the X and Y coordinates, and a constraint that the second Z coordinate was greater (higher) than the third [Powers, 1983; 1984;1989].

The business of handcrafting even simple scenes in the required representation was extremely onerous, and for that reason the Magrathea world creation system was developed [Hume, 1985]. This system allowed the depiction of objects using line drawings of geometric shapes, presented graphically with 3D perspective and elementary laws and constraints (e.g. objects could not pass through each other, and collisions were detected). The world designer could specify fixed and mobile objects, as well as motile actors controlled by a separate program. The system was originally implemented using communicating Prolog processes on a Unix system and a Tektronics storage scope, and was originally used to explore learning of nouns and verbs [Chan, 1988]. Once the IBM PC arrived, it provided a more convenient output device; and later an X-windows version was developed to take advantage of the more general capabilities of X-terminals – whilst still retaining the underlying. The latest incarnation makes use of Java-3D to provide fully rendered graphics and has been used to explore learning of prepositions, nouns and adjectives simultaneously [Homes, 1999].

With Magrathea, it is possible to arrange for both the learner and the teacher/parent to be simulated robots, and for other entities to be programmed and/or controlled from a terminal. For example, in one scenario, the dog would look out for the postman and chase him, while a bee buzzed around his head. This allowed for sequences of complex scenes to be generated quickly, and the terminal interface with the terminal-controlled actors permitted the postman to be directed with simple commands like 'north 100' (which it executed at its standard walking speed), while the learner was given English sentences or commands and also received a representation of the world, in relative coordinates based on the geometric elements that could be viewed from its location in the simulated world.

Magrathea worlds have been used for standard concept learning experiments involving nouns (e.g. learning to recognize an arch), as well as for more innovative experiments involving verbs (e.g. distinguishing verbs of motion, such as fly, walk, run [Chan, 1988]) and prepositions (e.g. mapping the range of situations where different prepositional relationships apply – including comparative linguistic

analysis of English and Ukrainian [Homes, 1999]). But nonetheless it only represents the rudiments of an ecology, and the experimenter still determines exactly what the ecological, semantic and syntactic relationships are that then become the basis for the learning experiments. The semantic and ontological structures learned thus have more to do with the limitations of the system and the preconceptions of the designer than with the real world, and are drastic oversimplifications. A similar approach in a crude robot world has been adopted in the L0 project and a number of researchers are working in this paradigm [Feldman et al. 1990; Hogan et al. 1998].

Nonetheless, most of these worlds have been parameterized in such a way that assumptions about the level of information available to the learner can be controlled. For example location coordinates could be absolute or relative to the eye, objects could be "seen" in terms of basic geometric properties or identified to the learner as body parts [Home, 1984], sentences could have a target preposition and the associated landmark and trajector identified, or could be forced to deduce the relationships between the words and their referents at the same time as learning the relationships between the prepositions and the interrelationships of the various objects. The aim was to demonstrate that certain things could be learnt from first principals, and then to implement these capabilities at a higher level so that more complex relationships could be learnt - manually optimizing Piaget's process of turning recognized into recognizor. A key issue in determining the useful associations in an unsupervised way is handling the focus of attention, which is the goal of work described elsewhere in this volume [Kozima and Ito, 1998; this volume].

3. REAL WORLD ROBOT BABIES

Just as in the simulated world, not all of our real world experiments are carried out with the doll-based robots. In fact we have one robot baby, a secondary/experimental mock-up arrangement that lacks the cosmetics desirable for interaction with children, and an intelligent room that is wired up with multiple microphones and cameras.

Both for purposes of data collection and interactive learning, the robot baby needs to be supplied with a variety of sensor-motor capabilities. We want to go beyond what can be achieved by simply videotaping a formal session with a child, or even spontaneous interactions between a child and his family/environment. Ideally we want an audio-visual data stream from the child's perspective. Additionally it is useful to have an audio-visual data stream from an external perspective. Our simulated world originally provided for each object to have arbitrarily located eyes, and eyes were provided as standard in front and above the stage.

Our initial robot baby is about 30cm long and is designed with multiple (2 to 4) electret microphones, 12 touch sensors and 5 motors (one per limb plus one for the head), as well as acceleration, shock and orientation sensors - and a speaker located in its stomach! Originally this information was subject to extreme bandwidth limitations and we could collect only the sensory data and trivially compressed 8-bit

8kbs stereo audio through the internal 6809HC11 microcontroller. However, experiments with direct data collection by an array of Windows PCs use a tetrahedral array (designed to fit ears, nose and crown at 16-bit 22kbs analog) in a room wired with two ceiling mikes (16-bit 22kbs), two directional USB mikes (16-bit 44kbs) and two USB videocams located on a monitor 1 to 2m. in front of the subject (16-bit 44kbs 20fps 352x288).

A new power-hungry robot of similar size is currently being completed that also incorporates stereo cameras and a 500Mhz Pentium III running Linux, has an additional motor to control convergence of its new eyes, and uses a PIC-based interface to control the motors. This uses around 10 amps at 12 volts and is designed to operated at a body temperature of around 38-39º C - as a side-effect of the heat generated by the processor. The batteries are specified to last an hour or so per charge.

A further robot is on the drawing board that will use multiple Digital Signal Processors to distribute the processing load for the eyes, ears, sensors and motors.

3.1 Preliminary Experiments and Programmed Functions

At this stage we have not placed a robot with a child or tried to mother it as a child - but the robot that is currently being completed will be able to be used in such experiments. The initial robot has been programmed with behaviours appropriate to these experiments, for example it turns to look at the speaker (identifying location by means of its multiple microphones), or it turns to the side it is touched on. The robots are theoretically able to learn to crawl, and simple crawling motions have been programmed, but the initial robot is too heavy to crawl and the new robot with its more powerful motors is not yet complete.

For initial experiments in data collection, the motors and speaker will be used to provide feedback only - even though the second robot could be programmed to crawl, this is a behaviour we wish to emerge rather than program. The responses we have programmed are designed to give feedback to the human child and maintain interest. Gradually we hope that more and more sophisticated responses will be learned by the robot. However for data collection we believe that the programmed responses are adequate - and young girls will play with totally unresponsive dolls for hours anyway. Thus we expect that high quality data will be collected even in the early stages, and that we will be able to validate it ecologically with experiments with different degrees of responsiveness/feedback provided by the doll.

Speech Reading and Sensor Fusion

There are a huge number of engineering issues in designing a robot, and the hardware side we will ignore here. However the software side and the capabilities of the hardware are both directly germane to our ecological language learning goals. It was important for us to validate that the capabilities of our robot baby were adequate to the ecological goals we have for it. In particular, is the microphone array capable of locating a speaker, identifying a voice, recognizing speech,

recognizing a face and synchronising and integrating information from multiple sensors.

One preliminary demonstration of these abilities was the crude ability to orient to a voice as alluded to above. A more interesting and significant touchstone problem is the ability to recognize speech and to improve on pure acoustic recognition by making use of the available visual information, there being considerable psychological evidence that visual cues influence human speech recognition. Clearly this involves, as a first step, detecting appropriate features in the auditory and visual streams. Then, as a second step, we have the problem of combining the features appropriately.

This is a specific instance of a more general problem known as sensor fusion or still more generally, data fusion. Unfortunately, research in data/sensor fusion has been less successful than hoped due to a problem known as catastrophic fusion - often the fused result is less accurate/useful than the best that can be achieved with one of the components [Movellan and Mineiro, 1998]. Analyzing the phenomenon of catastrophic fusion shows us that this arises because the fusion system cannot distinguish which features are more helpful and which are less helpful but gives each a similar weight that is independent of the specific context. The ecological context is very important because, for example, in a high noise environment more weight should be given to vision whilst in dark conditions or with bearded subjects it is probably more appropriate to weight toward audition, even for features which would be better determined using other sensory input under ideal conditions. There is a large measure of redundancy in our linguistic processes, and it is also important for our fusion system to be able to take advantage of this when redundant information is present, but even more particularly when the usual information that would be used to make a decision is absent.

Our aim was to demonstrate that our hardware and our approach to feature detection and evaluation could be used to learn to track faces, facial features, voices, and speech features, as well as to combine them. In our pilot study we were prepared to adopt a ‘no holds barred’ approach, but ultimately we aim for the speech reading capability to be learned totally unsupervised.

Our technique was to use standard speech recognition techniques to discover auditory features such as voicing and point of articulation, and then to use visual techniques to identify visual features that were better indicators than some of these features. These we call visemes, and those we made use of in this study were the points of articulation: bilabial, dental and velar. These are more accurately seen than heard when good visual imaging is available. Our experiment was made even more complex in that we wanted to learn to speech read at normal speaking or video-conferencing distances of around 1.3 to 1.8 metres in a normal noisy office or home environment, and we also trained on only two instances of each phoneme for each of three different speakers, tackling the general problem of face identification, mouth tracking, and speech recognition for multiple speakers of different sexes, bearded and unbearded. Under these difficult conditions, auditory phoneme recognition was only 21% but this increased to 29% by using the visual cues [Lewis, 2000; Lewis and Powers, 2000]. We are currently working to expand our data sets and expect that this will improve dramatically as more training data becomes available. It must

be remembered that commercial speech recognition requires a close microphone around 2cm away from the mouth, is based on masses of training data, and is assisted by high level linguistic and statistical models. In addition we were only able to run our camera at 288x352 resolution and 20 frames a second for this experiment, and the height of the mouth is only of the order of 25 to 50 pixels at this resolution and distance. Furthermore the recordings were deliberately made under 'normal noise conditions' in an office environment containing half a dozen noisy PCs!

The issues involved here are very similar to those involved in learning ontological, semantic and syntactic relationships. One of the key ideas we are using in this work is the concept of statistical significance - we only wish to seek correlations when the features show a certain salience and where they show significance for the relationships we are learning and for the goals of the learner. The issue of goals and motivation is a very significant factor in ecological research and one we are only just starting to come to grips with. These concepts of significance and salience have also proven useful in our high accuracy grammar checkers [Powers, 1997b; Huang and Powers, 2001] and in our learning of semantic relationships in a simulated world [Homes, 1998]. However, the amount of data involved in real cameras and microphones, running even at these relatively low resolution and speeds, is enormous. We can only fit about half an hour of multi-camera multi-microphone data on a CD-ROM. We are thus looking at ways to extract the useful data and correlate it on the fly.

The techniques we used for visual processing of the images, location of the face and tracking of the mouth were also novel and psychophysically motivated. We found that tracking a particular shade of (blood) red was useful for discovering the face, independent of beards etc.; we found further that excluding red was useful in maximizing the contrast on facial features; and we are currently investigating connections between these heuristic techniques (which work better than the standard engineering techniques we tried) and the psychophysical opponent colour model. Furthermore we are exploring self-organizing algorithms that will automatically determine the appropriate optimal face discovery and mouth detection contrast enhancements - and these are again essentially the same kinds of techniques we have been discussing in application to the discovery of linguistic structure.

Auditory Signal Separation

Another area where we have been applying a variety of unsupervised neural network techniques to our multimodal data is the location, separation and deconvolution of the auditory data we are recording. Currently we have been focussing on simple stereo recordings as again the computational complexity is considerable, and it increases enormously as we increase the dimensionality of the problem. The location problem is fairly simple. We faced the speaker by simply turning the head to balance the sound coming in each ear. However, once we have located a speaker it is relatively simple to combine the signals in such a way that we enhance the speaker's voice while cancelling out noise. Furthermore, we are also exploring the use of vision to assist in locating a particular speaker in a noisy environment. This is once again indicative of our 'no holds barred' approach as far as combining

multiple information sources goes. On the other hand we are trying to use self-organizing and unsupervised techniques to achieve our goals here, and at present our successful results have been achieved as hand-programmed solutions.

The traditional approach to auditory signal separation is to use blind separation techniques where all you have is the signals - the stereo audio in this case. However it is a bit more complicated than this because in real world situations the signals arrive at different delays and by multiple paths as the bounce of walls and other surfaces - this means it is really a blind deconvolution problem. We have developed a number of second order and higher order techniques that have shown good accuracy in separating and deconvoluting signals recorded in our office environment [Li et al, 2001]. In this case we are using recurrent neural nets somewhat similar to those used in our very first language learning experiments to look for similarity between the different signals and the delayed versions that are captured by the recurrent network - which acts very much in the same way as our blackboard model, allowing us to compare things that occurred at different points in time.

Other promising approaches include Auditory Scene Analysis (ASA) whereby acoustic events or frequency components that occur at the same time are associated [Bregman, 1995]. In fact, this mechanism is supposed to match up facets of an event however recorded by our sensory-motor system, and to be mediated by synchronized firing of the associated neurons.

4. CONCLUSION

In this chapter we have emphasized the similarities of the approaches and techniques we have used in many different facets of the total ecological robot baby learning experience. We have shown that a single technique can discover linguistic structure in speech, text, phonetics and word level linguistic input, and aggregate it into reasonable-looking chunks at levels up to simple clauses and noun phrases. We have seen that these chunks may be correlated by essentially the same techniques with chunks that have been discovered in a simulated robot world. We have seen that we can throw away many of the traditional assumptions and achieve highly encouraging results in the unsupervised learning of linguistic structure across multiple levels from speech and phonology to syntax and semantics. All of these results are based on the simple idea of CIE/CAE, of syntax and semantics. This of course relates closely to the emphasis Cognitive Linguistics places on metaphor, indicating that metaphor has very deep roots in our cognitive learning processes, as recognized also some Cognitive Linguists [Deane, 1992].

We have also discussed our use of similar techniques - though in these cases not identical - that we have used to achieve a reasonable level of speech reading and signal separation and deconvolution. In these cases there are many technical details that are beyond the scope of this paper, but once again it all amounts to detecting and identifying similarity and contrast, and evaluating the statistical significances of the correlation discovered. Correlation, of course, like metaphor, is really just another term for similarity.

REFERENCES

Block, H. D., J. Moulton, and G. M. Robinson (1975). *Natural Language Acquisition by a Robot.* **International Journal of Man-Machine Studies 7**: 571-608.

Bod, R. (1995). **Enriching Linguistics with Statistics: Performance Models of Natural Language**. ILLC PhD Dissertation, University of Amsterdam, NL.

Boy, J. (1977). **Dechiffrierungsalgorithmen zur phonetischen Identifikation von Buchstaben.** Dissertation, Universität Bochum, Studienverlag Dr N. Brockmeyer

Bregman, A. (1990). **Auditory Scene Analysis: The Perceptual Organisation of Sound**, MIT Press

Brent, M. R. (1997). *A unified model of lexical acquisition and lexical access.* **Journal of Psycholinguistic Research 26**: 363-375.

Brooks, R. A., C. Breazeal, M. Marjanovic, Brian Scassellati and M. Williamson (1998) *The COG Project: building a humanoid robot.* In C. L. Nehaniv (ed.), **Computation for Metaphors, Analogy and Agents**, Springer-Verlag LNAI 1562.

Brooks, R. A. and L. Steels (1994). **The Artificial Life Route to Artificial Intelligence: Building Embodied Situated Agents**. Lawrence Erlbaum.

Cardie, C. and R. Mooney (eds) 1999. Special Issue on Natural Language Learning, **Machine Learning**, Kluwer.

Chan, R. (1988). **Concept learning by computer: simple movement**. Computer Science Honours Thesis, Macquarie University, AUS.

Deane, P. (1992). **Grammar in mind and brain: explorations in cognitive syntax**. Mouton

Entwisle, J. and Groves, M. (1994). *A method of parsing English based on sentence form***. New Methods in Language Processing (NeMLaP-1)**: 116-122.

Entwisle, J. (1997). **A constraint parser for Englsh.** Computer Science PhD Thesis, Flinders University of South Australia, AUS

Feldman, J. A., Lakoff, G., Stolcke, A., Hollback Weber, S. (1990). **Miniature Language Acquisition: A Touchstone for Cognitive Science.** TR-90-009. International Computer Science Institute, Berkeley, USA.

Finch, S. (1993). **Finding structure in language**. PhD Thesis, University of Edinburgh, UK.

Gold, E. M. (1967) *Language identification in the limit.* **Information and Control 10**: 447-474

Grünwald, P. (1996) *A Minimum Description Length approach to Grammar Inference.* In S. Wermter, E. Riloff, G. Scheler (eds), **Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing**. Springer-Verlag LNAI 1040

Harris, Z. (1960). **Structural Linguistics**. University of Chicago Press.

Hewson, J. (1991). *Determiners as heads*. **Cognitive Linguistics 2(4)**: 317-337.

Hogan, J. M., J. Diederich and G. D. Finn (1998). Selective Attention and the Acquisition of Spatial Semantics. In D.M.W.Powers (ed), **New Methods in Language Processing and Computational Natural Language Learning (NeMLaP-3/CoNLL-98)** 235-244, ACL

Homes, D. (1998). **Perceptually grounded language learning**. Computer Science Honours Thesis, Flinders University, AUS

Horning, J. J.. (1969). **A study of Grammatical Inference**. PhD thesis, Stanford (Computer Science).

Huang, J. and D. M. W. Powers (2001). *Large-scale Experiments on Correction of Confused Words.* **Ausrtalian Computer Science Conference, Bond University, Queensland AUS**

Hume, D. (1984). **Creating interactive worlds with multiple actors**. Computer Science Honours Thesis,. University of NSW, AUS.

Kozima, H and A. Ito, (1998). *Towards language acquisition by an attention-sharing robot*. In D.M.W.Powers (ed), **New Methods in Language Processing and Computational Natural Language Learning (NeMLaP-3/CoNLL-98)** 235-244, ACL

Kozima, H and A. Ito, (2001). *How infants learn to control others' behavior - a route from attention-sharing to language acquisition*. **This volume!**

Kohonen, T. (1982). *Analysis of a simple self-organizing process.* **Biological Cybernetics 44**: 135-140.

Langacker, R. W. (1997). *Constituency, dependency and conceptual grouping.* **Cognitive Lingustics 8(1)**: 1-32.

Lewis, T. W. (2000). **Audio-Visual Speech Recognition: Extraction, Recognition and Integration** Computer Science Honours Thesis, Flinders University, AUS

Lewis, T. W. and D. M. W. Powers (2000). *Audio-Visual Speech Recognition using Red Exclusion*. **Visual Information Processing VIP2000**

Lewis, T. W. and D. M. W. Powers (2001*). Lip Feature Extraction using Red Exclusion*. In P. Eades and J. Jin (eds), **CRPIT: Visualisation 2000, vol 2**.

Li, Y, D. M. W. Powers and P. Wen, *Separation and Deconvolution of Speech Using Recurrent Neural Networks*, pp. 1303--1309**, Vol. III, Proceedings of the International Conference on Artificial Intelligence (IC-AI'01)**, June 25-28, 2001, Las Vegas, Nevada, USA.

Malsburg, C. von der (1973) Self-organization of orientation selective cells in the striate cortex. **Kybernetik 14**: 85-100.

McCarthy, J., L. D. Earnest, D. R. Reddy and P. J. Vicens (1968). *A computer with hands, eyes and ears.* **AFIPS Conf. Proc. Fall JCC 33#1**:329-338.

Mehler, J., P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini and C. Amiel-Tison (1992). *A precursor of language acquisition in young infants.* **Cognition 29**: 143-178.

Miller, G. A. (1956). *The magical number seven, plus or minus two: some limits on our capacity for processing information*. **Psychological Review 63**: 81-97.

Moulton, J. and Robinson, G. (1981). **The Organization of Language**, Cambridge University Press.

Movellan, J. and Mineiro (1998). Robust sensor fusion: analysis and application to audio visual speech recognition. **Machine Learning 32:**85-100.

Piaget, J. (1955). **The Language and Thought of the Child**, University of Geneva Press.

Piaget, J. (1971). **Psychology and Epistemology: Towards a Theory of Knowledge**, Viking Press.

Pike, K. L. (1949). **Phonemics**. University of Michigan Press

Pike, K. L. and E. G. Pike (1977). **Grammatical Analysis.** Summer Institute of Lingusitics and University of Texas.

Pike, K. L. (1967). **Language in relation to a Unified Theory of the Structure of Human Behavior**, Mouton.

Powers, D. M. W. and C. C. R. Turk (1989). **Machine Learning of Natural Language.** Springer-Verlag.

Powers, D. M. W. (1983). Neurolinguistics and Psycholinguistics as a basis for computer acquisition of natural language. **SIGART 84**: 29-34

Powers, D. M. W. (1984). *Natural Language the Natural Way.* **Computer Compacts** 100-104.

Powers, D. M. W. (1991). *How far can self-organization go? Results in unsupervised language learning.* In D.M.W Powers and L. Reeker (eds), **AAAI Spring Symposium on Machine Learning of Natural Language and Ontology**: 131-137. Kaiserslautern: DFKI D-91-09

Powers, D. M. W. (1992). *On the significance of closed classes and boundary conditions: experiments in Machine Learning of Natural Language.* **SHOE Workshop on Extraction of Hierarchical Structure**: 245-266. Tilburg NL: ITK Proceedings 92/1.

Powers, D. M. W. (1997a). *Unsupervised learning of linguistic structure: an empirical evaluation*. **International Journal of Corpus Linguistics 2**(1): 91-131.

Powers, D. M. W. (1997b). *Learning and Application of Differential Grammars,* **CoNLL97: ACL Workshop on Computational Natural Language Learning**, Madrid, July 1997

Ritter, H. and T. Kohonen (1990) *Learning semantotopic maps from context.* **International Joint Conference on Neural Networks**.

Schifferdecker, G. (1994), **Finding Structure in Language**. Diplom Informatik Thesis, University of Karlsruhe.

Shi, R., J. Werker, and J. Morgan (1999), Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. **Cognition 72**:B11-21.

Silverstein, M. (1976). *Case marking and the nature of language,* **Australian Journal of Linguistics 1**:227-244.

Steels, L. and R. Brooks (eds) (1995). **Building Situated Embodied Agents: the Alife route to AI,** Lawrence Erlbaum.

Steels, L. (1996). *A self-organizing spatial vocabulary*. **Artificial Life Journal 3(2).**

Steels, L. (1997). *Constructing and Sharing Perceptual Distinctions*. **European Conference on Machine Learning**.

Turk. C. C. R. (1984). *A Correction Natural Language Mechanism.* **ECAI-84: Advances in Artificial Intelligence**: 225-226, Elsevier.

Turing, A. M. (1936/7). *On computable numbers, with an application to the Engscheidungsproblem.* **Proc. Lond. Math. Soc. Ser. 2, 42**: 230-265; 43: 433-546

Turing, A. M. (1950). *Computing Machinery and Intelligence*. **Mind 59**: 433-460.

Turing, A. M. (1952). *The chemical basis of morphogenesis.* **Phil. Trans. Of the Royal Society, v237**, p5-72; London
Winograd, T. (1973). **Understanding Natural Language**. Academic Press.