Robot babies: what can they teach us about language acquisition?

David M. W. Powers
School of Informatics and Engineering
Flinders University of South Australia
powers@ieee.org

Introduction

The computer and the information age have made an indelible impression on how we study Linguistics and Psychology, and our increasing understanding of information processing by computers has led to both analogies and theories of human information processing. From the beginning of AI, computational models of human neural and information processing, of memory, learning and language, have been developed not only within AI but within Linguistics and Psychology. As the various communities started to interact with each other, Cognitive Science emerged as an umbrella field in which Linguistics, Psychologists, Computer Scientists and others work together to develop theories of human information processing.

At the core of Cognitive Science is the 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 or all of the cognitive and behavioural sciences.

Experiments in the unsupervised machine learning of syntax and semantics using a physical or simulated 'baby' robot, call into question a number of widely held assumptions, and in particular the following: 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 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; and 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 we see than an adequate identification of the interactional 'context' of learning is fundamental to formal analysis of learnability.

What is Language?

The starting point for a discussion of language acquisition must be a definition of what we mean by language and what we mean by acquisition. A robot baby is adopted as our model so that we can explore the totality of the language and ontology learning experience, and suggests a broad scope for what we understand as language acquisition. Our goal is to see what can be learned using the same mechanisms both for the different individual modalities and for the learning of cross-modality associations. Rather than arbitrarily assuming that different mechanisms are involved, we adopt a parsimonious approach where we explore what can be achieved with a minimum of assumptions, and no specifically linguistic assumptions.

In other words, we are seeking to devise a situation in which language, and language learning mechanisms, can emerge. We avoid words like ' evolution' and ' development' as much as possible because we distinguish between the evolution of a species and the development of an individual. When we use these words in relation to language our choice of terminology brings

with us the baggage associated with the original framework. However, if we talk about the emergence and conventionalization of language, we can avoid this confusion: the social group-learning process is quite different from the evolution of specific organs or capabilities in the species or the maturational development of specific organs in the individual. The mechanisms we are examining are those involved in the self-organization of a perceptual/conceptual ontology and the learning of natural and social laws. Another chapter in this volume focuses specifically on the role of attention sharing as a basis for more efficient development of appropriate linguistic associations, as well as the role of shared ontological belief models as a basis for communication [Kozima and Ito, 2001].

In this chapter we will first make a quick tour of some of the implicit and explicit assumptions which we explicitly eschew in adopting our focus on self-organization and emergence, reviewing them both from a theoretical perspective and in terms of computational/robot baby experiments that throw further light on their validity.

Assumptions questioned in our robot baby paradigm

? phonology, morphology and syntax are distinct

Much of linguistics, and even more of computational linguistics, is focussed on the structure of language. Indeed, like the world out there, our written and spoken language is perceived through our senses. It is easy to examine the structure of a sound or an utterance in the same way as we examine the structure of an object or a visual scene. Indeed grammatical analysis has often been used as a metaphor in the description and analysis of non-linguistic phenomena. In all cases, there are raw percepts, which are evidently organized into features and other successively higher order constructs that we give different names to.

The robot baby experiments directly examine the hypothesis of distinct linguistic modules by exploring the same learning mechanisms in multiple modalities, across modalities, and at multiple levels of processing. This gives rise to what would normally be called phonological, morphological and syntactic rules and classes as part of a single uniform rule and class formation process. Speech code vectors form classes that correspond to phonemes and vowels, phonemes or letters are organized into syllabic/morphemic units, and these are organized into structures that we recognize such as cliticized words, phrases and clauses [Powers, 1992; Schifferdecker, 1994]. Note that word and sentence did not emerge as recognizable constructions. We will look closer at these traditional units later.

? syntax and phonology are independent of semantics and ontology

We now turn to a similar but typically implicit assumption that semantics and ontology are independent of syntax and phonology and employ different mechanisms. We employ our learning mechanisms both on the baby's sensory-motor inputs and outputs and to the learning of associations between modalities. The intermediate and higher level structures emergent within a modality are available to guide the learning of phonological and syntactic structure, as well as for the obvious ontological and semantic associations.

Clearly for the learning of meaning, associative connections need to be formed that link the aural (sensory) and oral (motor) forms of open-class words with multidimensional sensory-motor grounded concept, and the robot baby clearly provides scope for exploring this kind of learning. What is less obvious is whether such multimodal associations are necessary for, or even useful for, the development of syntax or for the interpretation and employment of functional forms. The default assumption has been that syntax and semantics are totally different kinds of beast, that need to be captured in totally different kinds of ways.

But one of the big issues we need to deal with is how to direct our focus to the constructs that are likely to form useful associations, since there are many more that won't [Hogan, Diederich and Finn, 1998; Kozima and Ito, 2001]. Prosodic cues also seem to be important in bring focus to particular words or morphemes (e.g. nouns), but syntactic pointers (e.g. articles) seem to perform a similar role. Here cues from ontology, syntax and phonology all contribute to the identification of a nuclear role, and conversely the learning of semantic associations involves learning the associations between the elements in focus in these different modalities.

When we give our robot baby both linguistic and non-linguistic input, we are providing something that is essential for semantic information to be learned, but perhaps these associations are also necessary for the learning of syntactic categories. Suppose that the outputs of different levels of processing of the various sensory systems were available to the module, along with the phonological, morphological and syntactic structures we discover. Now we have the right inputs to form a real grounded semantics that connects the linguistic inputs to the perceptual inputs – and indeed to the linguistic outputs and the motor outputs, both directly in a propriosensory form and indirectly through feedback through the external perceptual system. But this is an empirical conclusion on the basis that certain structures were self-organized from text or speech alone in our early experiments. However the concept of word and sentence did not emerge in these experiments - perhaps these are more semantic than syntactic in origin.

We should also note that our concept of semantics and ontology is not limited to the linguistic and visual domains. Our multidimensional sensory-motor captures everything from smelly nappies and sore bottoms to rumbling tummies and sticky faces. Our current robot baby doesn't have these particular problems and so doesn't have the appropriate sensors, but it does have the capability of sensing touch, motion and orientation, and it can kick its arms and legs and turn its head. It can make some appropriate non-linguistic responses too, and will turn its head to the side you touch, and will seek to maintain head orientation if you turn the body. All this sensory-motor data needs to be organized by recognizing ‘useful’ relationships, and the relationships between all of our sensory-motor inputs and outputs gives rise to our ontological understanding of the world. This gives us a Piagetian higher order model in which our inputs can be previously learned concepts and relationships as well as basic percepts. For example, the learning of motion verbs involves recognizing both spatial and temporal relationships.

The relationship between the complex patterns and associations we associate with a dog and the phonological word /dog/ gives rise to the associated ontology and semantics. The same simple mechanisms will also associate them with the corresponding phonetic form. In our preliminary semantic experiments we have experimented with learning with various degrees of explicitness of focus - to explore what is required to make appropriate associations. Given the word /dog/ has been heard in a significant proportion of the situations where a dog has been seen or heard, it is an integral part of the sensory-motor associations for dog and given a large enough corpus (which we are far short of at present) should be learnable on a purely statistical basis even with minimal direction of the attention. Note that even the visual and motor associations relating to the lip movements involved in saying/mimicking /dog/ are part of this total sensory-motor associative plexus. Young children find it difficult to believe that a rose by any other name would smell as sweet; rather the name of an object is regarded as an intrinsic part of the object along with its sensory-motor attributes and functional roles (e.g. that dogs get patted and can bite, that roses get sniffed and can prick).

But there are many other associations with /dog/ and /rose/, and some of these have a syntactic character: they will be preceded by /a/ or /the/ (or /this/ or /that/ or /your/ or /my/) a significant proportion of the time. This has actually defined a functional class. Other words that may precede /dog/ or /rose/ and intervene between the noun and its determiner/specifier, include /brown/ and /red/, /noisy/ and /beautiful/ - which themselves will be associated with

particular sets of scenes that build up an impression that clarifies their semantic scope. The part whole analysis we apply to a visual scene, we can equally well apply to the phonological input. The role of movement and change in identifying distinct elements of a visual scene carry over to the idea of contrast in identical environments in phonology (CIE) and the idea of paradigmatic variation in syntax. What is intrinsically different between recognizing the relationship between a person and her head and arms and legs, and recognizing the relationship between a clause and its verb and subject and object? They are all in the end relationships among percepts as organized hierarchically into higher order concepts.

The robot baby experiments are examining the question of whether distinct mechanisms are needed for semantics/ontology and syntax/morphology/phonology by applying the same algorithms to ontological/visual tasks involving static and dynamic concepts (concrete nouns, action verbs and prepositions), to the learning of syntactic constructs, and to the learning of semantic connections between words and pictures [Hume, 1984; Chan, 1988; Powers, 1989, Homes, 1998]. Many of our experiments resemble classical concept learning, but our aim is for a model that does not arbitrarily distinguish between syntax and semantics and ontology, but rather uses general purpose associative mechanisms bottom up, finding progressively more and more complex constructs and allowing elements from more and more distant sources to be combined at the higher levels.

? there is a universal grammar

So let us come back to the concept of a Universal Grammar. To the extent that we have genetically and functionally similar anatomies and physiologies that give rise to grammar, if grammar itself exists at all in its traditional sense, then there must be a universal grammar. But we have seen that we do not need to assume separate linguistic modules either in the sense of a language organ distinct from our mechanisms for understanding our sensory-motor world, or in the sense of specialized modules for syntax, morphology and phonology.

The assumption of Universal Grammar has two parts, the *universality* of language within the human species, and the claim that a particular kind of *grammar* has some species-specific organic realization. The idea that grammar is mediated by rules has a long and pervasive tradition, but is being increasingly questioned. In essence, rules assume some kind of classification of words and then some kind of restriction on the relationships between these resulting classes. But this kind of classification and recognition of relationships is exactly what our neural models are good at, and they don't usually involve explicit rules and classes. Is it possible that the relationships are stored implicitly and the cascade of associations leads to the implicit selection of appropriate choices and rejection of inappropriate choices?

The generation of utterances seems quite different from the comprehension of utterances, but linguistics finds it far easier to autopsy a corpus of utterances than to come to grips with the process of language production. But just as grammar rules can work forwards and backwards, so can associative networks of neurons. Language is not the only form of generated behaviour, and again we can ask where the difference is between the mechanisms responsible for the different kinds of generative behaviour, linguistic and non-linguistic e.g. giving a recipe versus baking a cake.

One difficulty lies in the area of intentionality and conscious control of the generation process. We believe in free will and don't like to think we are not responsible for our behaviour, or that it is controlled by probabilistic forces beyond our control. But whatever form our goals and intentions take, or our unconscious habits and drives, all of these are inputs into the process and influence the final form. In any case, the problem of which sentence to generate from a lexicon and a grammar is even more of a problem for an unadorned grammar. The robot baby is designed to have the ability to generate both language and behaviour, and we can program in needs and drives, allowing us to explore this issue too.

? grammar is recursive/context-free

An even more specific assumption that we should question is whether we truly make use of recursive rules. Many issues simply disappear when we abandon recursion. There is also an important distinction to make between recursion and recurrence. When we apply recursion, we have to remember where to come back to, and the more times we apply it, the more of these return points we need to store. The evidence is clear that if recursion is used it must be depth limited. Conversely, recurrence involves modifying a fixed memory, and iteration implies a simple historyless processing loop. Thus we can loop while a condition is or is not satisfied, and as a special case we can count the iterations within some range. In a recurrent neural net, new inputs and recycled processed material are processed on each cycle, with powerful results demonstrated in explaining various aspects of visual processing.

The robot baby experiments make no grammatical or linguistic assumptions, but are based on neural, statistical and information theoretic models whose assumptions take the form of choice of algorithm, parameters and other constraints, such as the size of networks, hidden layers and context/concept windows [Miller, 1956; Yngve, 1961; Entwisle and Groves, 1994].

? while lexicon and grammar are finite language is infinite

At any given time our personal lexicon and grammar are finite. They fit in our fixed capacity heads and are based on our finite experience. But our language is growing, new nouns and verbs are being invented continuously, and innovation is a characteristic of language behaviour. Thus, the noun and verb classes really are open classes. On the other hand the grammar and the morphology of the language are the real characteristics of a particular language, and the closed classes are closed almost by definition, because changing them means we have a different language.

The question of whether a finite lexicon and grammar embodied in a finite head can generate an infinite language has a simple mathematical answer. If we assume that the grammar is at least context free we are implicitly assuming that once a recursive operation is complete processing will return to a precise point in the previous context. In a computer implementation an explicit return address or backpointer would be involved. If return addresses are stored in a finite memory then the language is depth limited and finite, and sentence length is bounded. If we assume that no return addresses are stored, then we limit ourselves to regular grammars, which are essentially tail recursive and are equivalent to iteration, and whilst language and sentence can be unbounded we lose our ability to explain various apparent constraints. On the other hand bounding our depth of recursion with a reasonable magic number [Yngve, 1961; Miller, 1956], allows us to use even more complex grammars, whilst providing an explanation of various constraints, with the bound obviating the need for special purpose kluges like subjacency. We thus seem to have a choice between the context-free or context-sensitive grammars that seem to be necessary to cover the language parsimoniously, and a regular grammar that is really only capable of producing lists, but can produce arbitrarily long ones. In saying that these are lists, we mean that it doesn't matter how much we extend the construct, we still end up with the same kind of unit and do not need to keep track of the internal structure for syntactic purposes.

In fact, lists may suffice us. Certainly they satisfy the common examples of constructs that allow construction of unbounded sentences: conjunctions like ‘and’ clearly introduce lists, and the ‘house that Jack built’ use of ‘that’ without centre embedding also can be viewed as a list: at each iteration, we are adding another stroke to the painting, but there is no additional grammar-processing memory overhead – in both cases our syntactic state before and after the additional clause (or other construct) is the same, and each member of the list is constrained to be of the same class (in the sense of being substitutable by any other member of the class). There is no need to remember a whole history of backpointers.

This is what a recurrent neural net does, its recurrent layer builds up a picture, which gets modified each time round, with new information and old information being processed together to produce the outputs and the next recurrent layer. The other favorite example, numbers, is just another example of a list, either a list of digits or a list of words where once we run out of new words for bigger numbers we have to resort to repetition (one million million million …) – the very simplest form of iteration.

But what is so special about sentences anyway? We can replace any of the full stops in a set of sentences by 'and' and end up with new legal sentences. Similarly we can replace any multiclause sentences involving conjunctions or relatives by a set of single clause sentences conveying the same essential meaning. Ultimately it is clauses that appear to have the strongest grammatical reality, and the combination of clauses tends to involve either extreme limitations on depth, if not simple iteration/lists. The paragraph, and higher units, are not normally regarded as subjects for grammatical analysis, as we move to another level of processing for which we employ new terms like style, narrative and discourse in our explanations of their structure, although there is no *a priori* reason to stop our grammatical analysis at the sentence either.

Similarly, one could examine the other end of our traditional stamping ground for grammars, the word. What is so special about words anyway? Isn't it the morphemes that carry the meaning, aren't the rules for combining morphemes describable using rules? Is the position of an inflection in a word really more restricted than the position of an article in a phrase or a vowel in a root? We don't know where to write the space half the time, and use or omission of hyphens depending on whether the same phonemic/semantic form is used adjectivally or nominally. We can even intersperse whole words or phrases in the middle of so-called words! (Un-bloody-likely you say! Das shreibe ich nicht um!)

The robot baby experiments assume that open class classes are finite but unbounded, and assign no special role to sentence or word - class membership can include constructs of different lengths. Utterance length is not explicitly limited, but the length of learned phrases/clauses is effectively self-limited by the number of layers of processing used before a uniform set of non-terminals is reached, although recurrent variants of the algorithms do theoretically allow arbitrarily long phrases (like numbers) to be parsed but will be limited by memory in practice [Powers, 1992; Entwisle and Groves, 1994]

? semantics and the lexicon can be acquired by learning but grammar cannot

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 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.

The theoretical results about not being able to learn grammars hinge on several assumptions: that language is recursive and non-finite (which we questioned earlier) and that no sources of supervision or distribution information are available [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 based on another assumption), and whole families of grammars that allow more complex 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 assumptions. It could involve any form of reinforcement, not just overt correction. This can include being understood or having the correct form used in a reflection, response or clarification or just in similar contexts. It could also involve comparing a generated form with a remembered form.

But there is another implicit assumption alluded to parenthetically just now: 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. 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, some would say evolving, a language. But here is another reason why neither learning nor evolution captures the process, and conventionalizing does. Both the mother 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, sometimes are never acquired. People don't use nominative/accusative pronouns 'correctly' in conjunctions, and prescriptive correction has lead 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).

The robot baby project has succeeded in learning both recursive and non-recursive grammars, but there is no target grammar and there is no requirement of identification in the limit. The grammar can and does change if the language environment changes. The classes and rules in our model only serve to identify units that seem to act similarly in relation to their context, both syntactic and semantic. The purpose of syntax is 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.

? grammaticality is absolute and probabilities pertain to performance only

Why should we make errors in performance? This would seem more relevant to our tennis serves than our speech production. If we are using absolute rules, why doesn't our performance demonstrate absolute perfection? And why haven't we been able to work out what these absolute rules are and implement a successful natural language parser. Although we see accuracy claims of 97% or 99% for parsers and taggers, these are based on the number of *words* correctly tagged, but the number of unrestricted sentences handled correctly is typically more like 50%. This seems to be the state of the art at the end of the 20th century. Many of these systems use probabilistic tricks to improve performance: e.g. if you can' t tell with more than chance probability when ' dog' is a noun or a verb, just assume it is a noun and you will increase your *average* performance figure because mostly it is - even if this means you are choosing the alternative that seems less likely in the *current* sentence.

If we question the rigid application of rules, we allow for issues like efficiency and computational load to play a role not just in an accidental sense in relation to performance, but in an active sense in relation to competence. Competence in language is really about effective performance appropriate to the specific context. Humans are basically lazy and like to cut corners, and language is no exception here. If we can convey the message with less effort we will do so, even if this means breaking or bending the rules. If we have conveyed our message accurately, in what sense have we made an error? Many of these shortcuts then get enshrined in the language, as frozen lexicalizations or cliches, as changes to the conventional usage of a word (e.g. compare US/UK 'write me/write to me'), as telegraphic forms, acronyms, portmanteaus and catchy slogans, progressing from now meaningless conventions ('How do you do?) to lexicalized corruptions ('Howdy!').

Class-membership is clearly fuzzy. Some words seem to belong mainly in one class, and a little bit in others. Most open-class words can be pressed into service as any part of speech in the absence of a standard derivation or suppletive for a particular role. The more common words tend to break the more common rules – having irregular forms, or suppletives, or requiring omission or alteration of the normal particles or inflections. Rules are possibly fuzzy too. In fact, we have here another implicit assumption, that rules and class membership are different things. Basically rules tell you how to make a member of a particular class, while class memberships can be written as a set of lexical. But these ideas of fuzzy class membership or multiple class memberships are different from the idea of probabilistic performance. A parser can identify ambiguity without necessarily assigning probabilities.

The brain is evidently based on neurons, which seem to operate probabilistically - at least during learning. Information theory tells us that efficiency is related to probability, and there is evidence that the organization of the brain is relatively efficient in an information-theoretic sense [Zipf, 1949; Shannon and Weaver, 1949] and neural network simulations are able to self-organize efficiently in this same sense [Malbsurg, 1973; Hertz et al. 1991; Kohonen, 1982; Ritter and Kohonen, 1990],. Furthermore probabilistic and information-theoretic approaches to grammar acquisition can avoid the negative results about learnability (breaking the assumption that we don' t have or use information about distribution). Our choice of words and our grouping of words into classes all involve similarity of words and/or concepts, which ultimately boil down to comparing usage patterns (linguistic and ecological contexts) on a continuous rather than discrete basis – some things are more similar than others. Nonetheless syntax does *not* seem to be probabilistic or arbitrary in nature.

The real question here is what we gain from the assumption of absolute grammaticality. Clearly probabilistic frameworks are the correct ones to investigate to test this assumption. The robot baby project uses algorithms based on neurological, statistical and information-theoretic insights. Where the training data warrant it, these systems can give absolute judgements, but the more typical operation is for them to give relative indications. The fact that we make deliberate and accidental puns and misinterpretations illustrates that absolute judgements are not always appropriate. Also small changes in the algorithms can give rise to different grammars for the same language, or to slightly more or less general languages - or to languages which make different judgements or parses for the same sentences. But the vast majority of grammars proposed are broadly consistent with our linguistic insights - and the surprises that remain provide additional insights.

Figure 1. *The first physical implementation of the robot baby has microphones in its ears, crude switches for touch, independent control of the head and each limb, and internal sensors for orientation and acceleration/shock. The electronics is controlled by a 6809HC11 microcontroller. Philips USB videocams are currently used externally for vision, and also provide additional microphones, but the next version will incorporate the cameras internally.*

One area where we have been examining the fusion of data from different sources is speech-reading. Our AV cameras (Figure 1) pick up speech and images and independently look for phonetic and visual features that can be used to identify phonemes. In a relatively noisy environment with microphones and cameras located on a computer monitor, and different speakers at different distances and heights, recognizing phonemes from the auditory signal is quite difficult - but close to 50% improvement in recall can be obtained by taking into account the visual features [Lewis, 2000; Movellan and Mineiro, 1998]. In this case the networks take evidence from the different sources and weight appropriately to give the final judgement. We are also developing more sophisticated fusion techniques that estimate the error for an individual instance (as opposed to using expected error based on probabilities) before combining the features. Humans appear to be very good at assessing the reliability of different sources of information and compensating appropriately, and assessing reliability and noise conditions is an obvious first step.

? closed classes/functional forms are learned late

Another fundamental bias arises from our interpretation of children's speech and the relative difficulty of assessing a child's comprehension of adult speech. 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 in some ways 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 first syllabic sounds (/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. Similarly /da/ accompanied pointing related to interest and attention-directing behaviour. /nana/ represented food, particularly

perhaps her mashed banana favorite. /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 visitor named John was also Ann! It is well known that homing in on the right level on generality is one of the hardest things a child has to do.

Although we are 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 it seems that 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]. 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.

Why have we neglected these words in our models of language learning? There is some evidence that they are recognized early, and there is room for considerable further exploration. Indeed, it is very easy to pull out these characteristic closed words and inflections, and it is worth considering whether how useful they are to the language learner.

So the question is why these words are produced so late. Or are they?

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 non-native speakers to master. 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 all cases 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 pseudodeictic 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' v(oir/regarder', 'gucken') will also be associated with the deictic function and become frozen into attention-drawing phrases ('Look at that!', 'Guck mal!', 'Voilà!').

So the assumption that open class words are learned before functional forms could be biased by our focus on production, and the difficulty of assessing comprehension, or recognizing exactly what was intended or just what was understood.

The robot baby project uses two classes of algorithms in terms of 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]

Paradigms and Algorithms

The total context.

Before going on to discuss our computational experiments in more detail, it is important to make a distinction that is fundamental to experiments on learning.

The formal results about learning do not relate to any particular theory or algorithm about learning – they are independent of mechanism. They say whether *any* mechanism is capable of learning the target to criterion under particular conditions. What is important is the ecological paradigm: the context or environment in which learning takes place, including the relationship between the learner and other agents. Mathematical formalisms reduce this to a very simplistic concept, that of supervision. The well-known learnability results of Gold[1967] assume a sequence of sentences without supervision, whilst the normal level of supervision envisaged in learning theory would simple indicate whether a sentence was correct or not.

The supervisory arrangements used in most corpus-based or data-oriented language learning are even stronger than those normally used in learning theory [Bod,1995]: the complete set of tags for each word of the sentence is provided, if not the complete parse. This doesn't so much tell you whether a sentence is correct or not - all are assumed correct - but quite explicitly tells you the classes and/or rules. This paradigm is clearly unrelated to the one the child is faced with since there is no direct source of information available to him about rules or parse trees, and neither is the interlocutor able to supply such information as it is not known to her either.

Poverty of the Stimulus denies that even basic supervisory information is available – the child does not get told that his sentence is grammatical or ungrammatical, nor is he supplied with a set of starred sentences along with the unstarred ones. Even when correction occurs, it tends to be unfocussed and implicit, and even when explicit focussed correction is supplied, the anecdotes are rife about how the child appears to unable to make use of it. This may be a simple as not being ready to learn the corresponding details, or being more focussed on some other aspect of learning or communication at the time.

So what kind of supervisory information is available to the child?

There is some evidence [Turk,1984] that the child has a repository of recognized utterances or fragments which can be used to repair their errors – they make the error, recognize that it doesn't sound right, and repair it. It is even possible that anticipated correction takes place – that is the sentence is repaired, or competing choices selected, based on what the sentence should sound like. With a significant memorized corpus, preference for chunks that are similar to remembered chunks could play a significant role. Anticipated correction does not technically constitute supervision, but it does provide distribution information that can serve a similar role.

The child also has available contextual and semantic information that can help the choice between different possible interpretations of a sentence, and hence different possible structures and rules. This raises the question of the relationship between the development of ontology and semantics on the one hand, and syntax on the other. Indeed the focus of psycholinguistics, as we saw above, has been on syntax and assumes that this is the hard part and that semantics is easier. But we saw earlier the possibility that simple surface grammatical/morphological cues may focus attention on the words that correspond to the features of the scene that are in focus, which suggests that such early syntactic awareness may assist in the development of semantic associations.

A proper model of language learning should associate models of ontology and semantic learning with grammar learning, and seeks to reflect the total environment in which the child finds himself.

The available mechanisms

It is difficult for us to know how much a chimpanzee understands about its world, because it cannot tell us. We can see some evidence of memory and reasoning, of understanding of principles of cause and effect and of simple physics, but it is tricky to tease out the effects of instinct from the effects of learning until we place the animal in an artificially manipulated environment. Much the same can be said about a child in his first year, but when contrasting the child and the chimp, we are unable to distinguish which factors that separate them after that first year are due to purely linguistic development, and which are due to more general cognitive development. But even leaving language aside, our ability to control and manipulate our environment would seem to be vastly superior. Indeed, tool-making competes with language as we try to characterize what capabilities distinguish the species.

At the lowest level, it would appear that there is a repertoire of mechanisms that is available to both chimp and child. The mechanisms that lead to recognizing visual features were first discerned in experiments on chimps and cats rather than human subjects. Basically these can be characterized in probabilistic terms, associating co-occuring characteristics, recognizing that there is no such things as identical stimuli, and that similarity will have to suffice.

This principle of quantifying similarity, classifying together percepts that are relatively similar and classifying apart percepts that are relatively dissimilar, is fundamental. It is also indicated in auditory processing, and indeed at many levels of sensory-motor processing of all kinds. It also seems that it is necessary to discard information that is of lesser significance, that we cannot retain every bit of sensory-information we are hit with, and we realize an advantage by discarding that which is less obviously relevant.

In general it would seem that we discard uncorrelated data – it conveys much information in a technical sense, but in the absence of patterns there is little we can do with it, and in the absence of correlation with basic survival drives there is little relevance. When there is patterning in the data, it means that some parts correlate highly, and we can represent these more efficiently as features precisely because of this predictability. Furthermore the prediction of significant events has considerable survival value. Once we have re-represented the data to abstract out the obvious correlative features in the local modality, we also impose some structure on the remainder of the data, and correlations across modalities can take into account both the features recognized in each, as well as correlations amongst the unrecognized portions of the data in each modality. This will automatically produce multi-modal concept representations. The more frequent and obvious features in each will act as locators helping us to associate the less frequent features, irrespective of their intrinsic salience.

Thus we expect to see frequent features, like edges, providing information about where to find the less frequent more widely varying kinds of information. The correlate in speech is that the frequent morphological features, like affixes and articles, provide information about where to find the more highly variable content words, and tend to identify their nature – including their part of speech and grammatical role. They further allow for the cross-correlation of what is sandwiched between functional forms in our auditory stream and what is bounded by edges in our visual stream. In fact there is psychological and psychoacoustic evidence that different features of the same object, even in different modalities, are grouped together using a kind of frequency coding of the recognizing neurons. This is apparently achieved by correlating according to ‘common fate’ that is collections of features that appear, move and disappear together, are coded together [ASA].

Looking back over our discussion of assumptions, there is a common theme: we should treat them as refutable hypotheses, noting that there is never total confirmation - only refutation is certain. What we have described so far in this section are mechanisms of very general applicability across different types of perceptual and cognitive processing, and even across different species. The question is how far can they go in explaining the acquisition of language and ontology. Given that other species do not seem to have the higher level linguistic and inventive skills, it is necessary to consider two possible reasons for this: they simply do not allow the complexity of associations that we are permitted – some kinds of information may never come in contact with each other – or that there are one or more specifically human, and possibly specifically linguistic, mechanisms. This argument has always underpinned Universal Grammar, neglecting the possibility that it is structure rather than mechanism that distinguishes us.

In this project we are not particularly concerned with whether mechanisms are innate or learned, linguistic or generic, but rather at coming up with the simplest explanation of the child's developmental and acquisitive processes. This bottom-up approach, starting with known, obvious mechanisms – and milking them for all they are worth – will tell us the kinds of things that can be learned using this simple model, allowing us to focus on what is left to see whether or not those functions correspond to capabilities that we have and other species don't.

The Robot Baby

Building a Baby

Language learning experiments with robot babies, either in thought, computer simulation or mechanical implementation, go back at least three decades, and the idea that language will need to be learned by a robot in a real environment rather than by an isolated computer was considered in the paper that originated the famous Turing Test half a century ago [Turing 1950]. Our own experiments go back over 20 years, but the majority have involved simulated rather than actual robots.

In some ways, these early experiments were premature, as the computational power required was unavailable and underestimated, and our understanding of Machine Learning, Neural Nets was not nearly so well-developed. Nonetheless useful principles emerge, including the idea that there should be a strong correspondence between the sensory-motor capabilities of the robot and the language learning mechanism. 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. Winograd[1973]'s famous language understanding robot arm, SHRDLU, was one of the first such simulations. Even where 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 robot babies are being built and some initial 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 robot-animal toys that claim to learn, to life-size or doll-size

babies. The smaller robots and animal- or 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 they 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 very interesting 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' .

Whereas previous experiments have been operated under artificial and restrictive experimental conditions, our concept of a robot baby extends to the idea of placing a robotic doll with a young child and using it to collect a comprehensive corpus of audio, visual and sensory-motor data from a perspective very close to that of its human owner/sibling, as the two of them experience the world together and learn together, or as a slightly older child mothers the robot baby. As the project matures, we expect that the doll will be able to interact with the child and his parents in an increasingly natural way, responding in appropriate ways, both linguistic and non-linguistic.

The Language Modality

The first stage in our robot baby language acquisition model involves separate correlative processing in the individual modalities. The connections between modalities are assumed to take place at a higher level – and in this case by high-level we mean the level of morphology (or in vision, the basic-level categories, that correspond to balls and dogs).

Since vision is not our focus here, we will say little about it. Certainly self-organization up to the level of features is straightforward, with edge-detection and colour-constancy correction being important factors. Unfortunately the self-organized grandmother cells do not recognize grandmothers too well, but certainly features like eyes and mouths, and hence heads and faces can easily be recognized by the same self-organization processes that produce blob and corner detectors. So the level at which interaction is proposed is the level where self-organization peters out, at the point where we have the features available to recognize eyes and mouths and heads, but need to intermodal correlation to attach significance to them. Visual learning is computationally expensive so for some of our experiments (e.g. on lip/speech-reading) we have explicitly programmed rather than self-organized the appropriate visual feature recognizers and work with a reduced set of selected attributes.

In the language modality, self-organization from speech-code vectors into phones and phonemes, morphemes and syllables, words and word clusters, phrases and simple clauses seems to occur straightforwardly, although no system has yet gone the whole way in one experiment. Moreover, it seems that around 10 levels are involved, and my students and I have separately self-organized directly from speech-code-vectors to phones, and from phones to phonemes, from phonemes or letters to CV/syllable structure and all the way up through to unnested phrases and clauses, at which point we end up with a sequence of NP or VP like contructs. The levels at which intermodal correlation is proposed are the top three or four.

Simple experiments in semantic learning across modalities have been performed, but not in connection with these self-organized hierarchies. As discussed above, it is more efficient to explore the different 'modules' separately, even when essentially the same algorithms are used.

We will now proceed to examine the different types of experiments and the prospects and hinderances in relation to bringing everything together.

Grammar Acquisition

Our initial focus in designing a language learning model has been the unsupervised acquisition of structure, since our aim is to learn with the simplest mechanisms and the minimum of assumptions, and in particular to see what can be achieved without supervision and to characterize what kind of things cannot be learned this way.

Our earliest experiments [Powers,1983-9] were based on extension of a basic phrase structure grammar, based on the insight that words either had to group to the left or the right, and they could either group with another word, or with a larger group such as a phrase. The first version was supervised, and explicit feedback was provided about whether the grouping was appropriate or not. This is a very strong form of supervision and can be done interactively or by making use of a pre-parsed treebank. An unsupervised approach was also developed in which we counted the number of times different grouping rules proved useful, which provided a form of implicit but delayed voting for the different alternatives.

In developing the unsupervised version, we also removed the initial grammar and forced the program to start from scratch, making hypotheses about word classes and grouping rules. We increased the number of levels that could be considered for hypothesizing rules from two to a nominal seven – the seven most likely candidates for the sequence of words seen so far were maintained. Whenever a new parse tree was needed to incorporate the next word, the least likely of the seven stored putative 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.

This model succeeded in learning to parse small phrases/clauses hierarchically, but proved to be extremely limited and quite unreliable as utterance length increased. But what was interesting was that 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 words as the keys, 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 words that were the seeds around which the crystalline parse structures grew. This meshes in with the perennial suggestions that articles may be the head of the noun phrase [e.g. Hewson,1991].

But there was another strand to this: a totally independent model [Powers,1984,1989] based on recurrent time-delay/decay self-organizing neural networks achieved almost identical results to the complex statistical model used in the previous model – and with only a page of code! There was clearly something significant about the closed class words. In addition preliminary experiments were undertaken with a simulated ontology in parallel with these early grammar learning experiments, and a simulated robot world was built to facilitate this work on semantics [Hume,1985].

The results highlighting closed class elements were influential on another series of experiments [Powers, 1991-2] inspired by Pike[1949]'s method of phonological analysis, and focussed at the character/phoneme/speech level. Rather than trying to work statistically, the

idea was that a particular combination either was or was not possible. The basic idea was that of contrast in identical or analogous contexts (CIE/CAE). By collecting together all occurrences of a particular context – the sequence of two or three units on either side of a target unit or sequence – we collect a filler class of two to seven fillers that contrast in a single set of identical or analogous contexts. The number of such slots in which the class occurs is used as an indicator of the significance of the class, and the most significant class is labelled with a non-terminal symbol.

We successfully predicted that at character/phoneme level, the vowels would be the first class to emerge, and that at word level the articles and punctuation would again be the most important classes. The members of the discovered class were recoded with the new non-terminal, and the entire process repeated until we ended up with an iterated sequence of a single non-terminal. This non-terminal essentially represented alternately noun phrases, prepositional phrases and verb phrases. A variant of the process was used which allowed a forming class to have its non-terminal added before finalizing the class, thus allowing the formation of hyperclasses involving recursive rules.

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:

```
A <- a
A <- e
A <- i
A <- o
A <- u

B <- rA
B <- Ar
B <- lA
B <- Al
B <- A
```

Normally, with this method we have started from characters, but the following grammar illustrates the kind of rules we might expect if we applied it starting from words. A, B and N represent classes of articles, adjectives and nouns that are not shown, T and V represent transitive and bitransitive verbs, and R corresponds to a noun phrase.

```
P <- at
P <- in
P <- into
P <- on
P <- onto
P <- out
P <- out of

Q <- N
Q <- B Q

R <- A Q
R <- Q

S <- T R
S <- V P R
```

In fact the grammars found are never this simple, and indeed allowing recursion (Q is recursive in this constructed grammar) tends to produce much worse looking grammars (including totally degenerate grammars) compared with the standard version of the algorithm. While the grammars would be easier to understand if labelled with standard English non-terminals, the classes are discovered by the program and labelled with successive letters of the alphabet.

The point of this example is simply to illustrate that when you group together sequences of one or more units that occur in essentially the same set of contexts, the resulting classes are not just simple lexical classes, but permit more complex entries, and indeed whole hyperclasses of context-free rules. This does not illustrate what any particular algorithm finds, but rather what the paradigm permits – the formal learnability results are, as mentioned above, not about any particular algorithm, but about what is possible or representable in the paradigm. This learning paradigm allows the representation of arbitrary context-free grammars. It cannot represent, and thus cannot learn, indexed or other context-sensitive grammars, because the left hand side is restricted to a single non-terminal labelling the induced class.

A related two step approach to unsupervised learning which has been developed independently by a number of researchers [Langley, Grunwald], involves merging words or constructed units into classes or phrases, with the aim being to achieve an equivalent but more compact representation. This is based on an idea of parsimony known as minimum description length, and related to Shannon's information theory [Shannon and Weaver, 1949] and Zipf's principle of least effort [Zipf, 1949].

Automatic Segmentation

The child's task is to make sense of his linguistic input, but we cannot assume that he has pre-segmented words or morphemes available to him, nor can we assume that he has the memory capacity to have whole sentences available for dissection. Rather, his first task is segmentation, identifying what the useful pieces are. Thus our bottom up approach seems a more appropriate model since it does not impose as much of a memory or computational burden. The simple noun and verb phrases that emerge correspond to a level that seems appropriate for crossmodal correlation and association, and indeed this model makes an interesting prediction, that these phrases are essentially of a similar kind.

We are currently looking at retaining the more frequent/functional/closed-class parts of the information as potential features, and propose that there is a finiteness feature that switches between verb-phrase/clause-like and noun-phrase/noun-like structures – for example 'to' marks a verb phrase as infinitival and noun-like (e.g. an infinitive can be a subject or an object), 'that' turns a clause into a noun, '-ing' makes a verb or a verb-phrase noun-like. These features are relatively easy to obtain using a variety of unsupervised techniques, as well as emerging automatically in our model. Semi-supervised learning techniques are already capable of producing a very competitive constraint parser given just these closed-class words and functional segments simply by noting which collocations are licensed in the corpus (an unsupervised algorithm) although at present manual decisions are made about words that are not derived or inflected forms whose range of roles is automatically determined by the morphology (the supervised aspect).

Proposals for automatic segmentation go back at least 50 years, and indeed many approaches use ideas very similar to those used by Harris [1960] and Pike [1949]. Generally speaking, the perplexity increases at a morpheme or word boundary – that is the number of choice for the following character/phoneme increases dramatically at these boundaries. However, as discussed earlier, the word is ill defined, and is not a true segment. For other levels of analysis, segmentation principles based on information theory are fairly effective and any impreciseness or error in these segmentations does not seem to preclude effective use at higher levels of analysis[Finch,1993; Brent,1997; Witten,2000]. Indeed it is possible to make multiple hypotheses available for higher level analysis in the same way that we deal with homonymy, homophony, polysemy and polyclassy through lattice parsers or Hidden Markov Models, and some associative models can deal with such fuzzy information without any significant increase in processing time or resources.

Ontological Learning

Work on recognizing objects in visual scenes and in learning simple concepts in a block world are areas of research in their own right within Artificial Intelligence, Neural Networks and Cognitive Science. Generally speaking however, supervised approaches are more common within AI and NN, whilst bottom up approaches tend to have more of a cognitive flavour.

Working bottom up, the lines and edges and blob-like features self-organize easily and early, and it is around this point that we want to look at how we can attach names to concepts, learning to recognize them in terms of these basic features. Most concept learning or ontological learning experiments would however assume that these edge-based features were provided directly by a simulated robot world. Our simulated world, Magrathea, was built in 1984 and used wire-frame models with full 3-D perspective, elementary physics (e.g. to ensure that objects bump into each other rather than pass through each other) and a variety of fixed, mobile and motile objects. The motile objects are agents that move around under the control of their own program, and were either simple behavioural scripts (the dog was on the look out for the postman and chased him when she saw him) or keyboard controlled (the teacher entered the world as a participant by controlling her persona). Similarly the learner could be a motile agent under the control of a learning script, but most often was represented only by an eye that provided a particular perspective. An eye could receive information in one of three forms: edge information in the field of view, geometric information about each shape that was seen, or ontological information identifying it as 'Fred's leg'. The different levels were used according to the aims of the current learning experiment – obviously you don't tell it you've seen part of Fred if the point is for it to learn to recognize Fred. It proved easy to learn simple geometric objects and noun-like basic categories. We also showed it was capable of categorizing different kinds of activities, and we learned some simple verbs of motion (1988).

Another major project on learning in a simple simulated environment is the ICSI L0 project research [Feldman et al., 1990, Hogan et al., 1998], where again a number of interesting concepts have been learned. Within this kind of simulation, it is not only possible to learn simple noun and verb concepts, but it is possible to learn more closely just how we define certain complex relations, and the preposition has been an important focus both for the L0 project and our own.

By setting up an experiment in which an object moved around and a subject labeled the scene either with a single preposition or phrase ('to the right of' versus 'beside') or a full sentence. The point of this more complex formulation is to move away from the supervised paradigms where the 'correct' word is associated with the scene. In our 1998 experiments [Homes, 1998], the learner had to deduce which was the landmark and which the trajector (in fact it only had to be consistent – if it selected the wrong one it would learn inverse relationships – left for right, etc.), it then had to make hypotheses about the various Cartesian relationships that held and see which were consistent for a particular preposition. In the full sentence version, an additional complication is present: it is now necessary to attach focus to the correct word. This was not reliably achieved in the absence of 'knowledge' of other words (open or closed class) in the sentence, and in a simulated world learning success is strongly influenced making items salient in an already oversimplified visual world. Although prepositional relationships can be learned easily once both the word and the trajector and landmark are salient, the real question is whether it can be learned using real sensory data based on plausible models of self-organization of structure, control of attention and assignment of salience. This is what we are providing with the physical robot baby.

Speaker Identification, Location and Separation and Speech Reading

The audio, visual and other sensors were added to the robot baby primarily with the intent of allowing the development of an ontology that permitted exploring the learning this kind of syntax and semantics in a rich environment with natural feedback of various kinds. However a number of other possibilities have been opened up for exploration by this step.

As part of verifying the suitability of our sensors for the ontological learning task, we wanted to see how well we could do speech reading. We wanted to ensure the baby had the capability to locate a speaker aurally and visually, and then to see if there was sufficient information in the visual stream to lip-read enough to improve the speech recognition process. Commercial speech recognition currently depends either on using a very small vocabulary (as used for simple phone menus) or on using a headset microphone in a quiet environment, and a statistical model to predict the likely path of the transcribed sentence. The baby is not usually going to be at the ideal distance for speech recognition and the environment will tend to be noisy. In addition the doll's microphones may also pickup sounds such as the child's heartbeat or the rustling of clothes if the doll is being held to the chest or rocked.

As with all of the experiments discussed here, this is work in progress although we have encouraging preliminary results. Although it may seem less relevant to the modelling of language acquisition, we see this AV processing as integral to the experimental program. Just as we have actually been making it harder for the computer by asking it to do parsing without semantic references and ontological grounding, similarly we are making speech recognition harder without the visual and directional cues that assist us in attending to and understanding a speaker. Also our recognition of a speaker's characteristics (both auditory and stylistic) is a key part of our ability to tune into and understand a speaker. All of this forms part of what we mean by developing a complete ontological model, although eventually we will be exploring emergent capabilities in this area whereas presently we are using a supervised training regime.

What the robot baby has taught us!

At the start of the robot baby project, 20 years ago, I naively assumed that grammar was about rules and expected to be able to learn cut and dried rules. I assumed that grammar was intrinsically different from phonology although I recognized that we usefully parse visual scenes and thus similar techniques should be applicable to language and semantic learning. The robot baby project has successfully used such perceptually-motivated mechanisms to discover patterns in simple simulated and real visual data, to analyze (auditory and visual) speech data, and to find classes and rules in phonetic and word to clause level data. Whilst the learned parsers and analyzers fall somewhat short of the best speech recognizers and parsers, we have been able to adapt the learned classes and rules to produce a commercially competitive grammar checker [Powers,1997] and a competitive constraint parser [Entwisle and Groves,1994].

The single most important discovery in this research program has been the role of the closed class - and the extension of the concept from a set of function words to the analogous classes at every level from the phoneme to the phrase. The second most important discovery is that segmentation comes for free if we simply allow the system work out what size phrasal units belong to a filler class filling a particular set of contextual slots, and the filler class thus becomes a hyperclass of context-free rules. Third, and this is something we are still currently exploring, it seems that the most frequent most closed-class elements in a phrasal unit lends their character to the unit and is responsible for cohesive interaction and syntactic constraints - we can regard this as self-organization of features. Our parsing strategy is essentially bottom up, so these constraints can bite early and influence the formation of the phrase structure or parse for the sentence.

Perhaps the most surprising discovery is that our learning algorithms produce a set of simple noun and verb phrases rather than full sentence parses and the possibility of these distinguished by a finiteness feature. Indeed, it appears that these phrases may actually be the most critical level for intermodal associations because they correspond more directly to the ontological associations.

Another interesting consequence of our focus on segmentation and classification is that rules turn out to be an emergent property. Moreover, while very consistent and accurate segments emerge at the various levels, there is often ambiguity as to how they may be composed from lower level units/segments. For the sake of drawing a parse-tree we can arbitrarily assume a greedy left-to-right heuristic that gobbles up as much as possible as early as possible, but this obscures the basic insight [Langacker,1997]. We don't need or have strict deterministic parses in the traditional sense, but rather our segmentation and classification processes allows extraction of constituency as an emergent artifact of the process. The syntactic constraints our systems learn need not force a unique parse tree, and we generate parse-trees only because they are expected by our peers and are necessary for quantitative evaluation of our parses against other approaches [Entwisle and Groves, 1994;Powers, 1992].

Bibliography

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

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

Brent, M. R. (1997). *A unified model of lexial 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.

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.

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.

Hertz, J. A., R. G. Palmer and A. S. Krogh (1991). **Introduction to the Theory of Neural Computation**. Addison Wesley.

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

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

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.

Kuczaj, S. A. (1983). **Crib Speech and Language Play**, Sprinter-Verlag.

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

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

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.

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

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.

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. (1997). Unsupervised learning of linguistic structure: an empirical evaluation. **International Journal of Corpus Linguistics 2**(1): 91-131.

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.

Shannon, C. E. and W. Weaver (1949). **The mathematical theory of communication**. University of Illinois Press.

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

Skehan, P. (1998). **A cognitive approach to language learning**. Oxford University Press

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

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. (1950). *Computing Machinery and Intelligence.* **Mind 59**: 433-460.

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

Winograd, T. (1973). **Understanding Natural Language**. Academic Press.

Zipf, G. K. (1949). Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley

Yngve, V. H. (1961). *The depth hypothesis.* **Symposia in Applied Mathematics XII**: 130-138, American Mathematical Society

Yngve, V. H. (1996). **From grammar to science: new foundations for general linguistics**. John Benjamins