

Linguistics as Science and Engineering

or

The Science and Engineering of Linguistic Behavior: The Robot Baby

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

Our hard-science linguistics has been focussing on the high level structure of an assemblage of communicating people and their surroundings. We have been developing a high-level understanding of the relationships between the members of a group as well as their physical environment, discovering the significant properties and constituents and exploring their representation in terms of linkages. This high level is the level we think in, and reflects the way we characterize the situation we find ourselves in. However there is a considerable distance from the recognition of objects and people, their relationships and their roles, to the basic perceptual input and motor output which is our interface with the world.

This low level extreme, the interpretation and structuring of the sensory-motor information our brain receives and generates, is the logical endpoint of our top-down human linguistic analysis. Of course there are no people, no surroundings, no relationships available to us other than what is built up from this sensory-motor level. Even our own self-awareness is mediated through our proprioceptive system; even our thoughts are eventually grounded in our experience. This level is the point where linguistic theory gives way to neuroscience. Indeed there is a considerably body of knowledge concerning the transduction from physical properties of our body and our environment to the neural impulses that enter the brain. There is also considerable understanding of how the outer layers of the brain, the cortex, process this information and provide detectors for a large variety of recognizable features of our sensory-motor world.

The last two decades have seen the emergence of a Cognitive Science that seeks to mesh the work of many component disciplines, from Neurology to Psychology to Linguistics. One of the key elements in this has been the increasing role of the computer model as a tangible expression of a theory. A theory of human cognition or human linguistics is faced with a bewildering array of detail at a multitude of levels. There are limitations to how well we can work out the consequences of a theory in our heads or on paper, and the computer provides a means to implement the theory in such a way that the predictions of the theory emerge as properties of an implemented computer model. Moreover, the requirement to specify a theory tightly and completely enough for computer implementation forces the development of a far more stringent theory that deals with every last detail.

Theoretically and rather simplistically - as a 'thought experiment' this goes back to Turing's 1950 paper on 'Can a Computer Think?' - we could replace the brain by a computer connected up to all its afferent and efferent neurons, and see how our computer-implemented theory of cognition fares. More practically, we can use what we know of the nature of the information that crosses this brain/nervous-system and emulate them too, using cameras and microphones, touch, orientation and acceleration sensors, motors and relays, etc. The android, or artificial

human, is a well known concept in science fiction, but is also some way from being science fact, even apart from our impoverished theories of human cognition and human linguistics. But we can make a start, and explore our first order approximations to a theory of human linguistics and cognition in the context of a first order approximation to a human body: in our case a neonate-sized doll.

Assumptions

The assumptions we need to make in a hard-science linguistics are very different from those we make in a traditional linguistics. We do not assume the existence of words and sentences, let alone constituent grammars and parse trees: these should emerge, if valid, as consequences of lower level assumptions. We will also not assume that linguistics is divided into phonology, morphology, syntax and semantics. As a consequence we will not assume that we have a language acquisition device with modules for each of these, but rather will examine the basic communicative ecology that provides our objective evidence, and develop hypotheses on an 'as needed' basis. Hypotheses will eventually have to boil down to first principles mechanisms, and eventually linguistics should reduce to biology, chemistry and physics. Parsimony tells us to make the minimum of assumptions, and to make each hypothesized mechanism do the maximum amount of work. Even once we have a good-looking theory, we need to examine our hypotheses to see if they themselves may be emergent properties that need to be explained with lower level theories. We do not even assume the existence of languages as concrete artifacts shared by a community of speakers. Rather we focus on the community as a group and as individuals, and examine the properties of their interactions with each other and with their environment.

In this work we will extend the basic scientific assumptions of a hard-science with some specific empirical observations and some tentative hypotheses.

Big heads

One of the basic empirical observations we can make is that the agent's head is of bounded size and thus the memory capacity limited. This basic observation leads directly to the insights behind the Depth Hypothesis [Yngve, 1960; Miller, 1956] although without a solid basis for a theory of constituent structure, specific proposals about the depth of parse trees are essentially useless as they are untestable [Yngve, 1996, Chapter 5]. However, without such a restriction any theory that assumes the existence of a grammar with at least context-free power is clearly false.

This immediately tells us that we are *not* looking for a traditional context-free grammar, although with the addition of constraints like the Depth Hypothesis, derivative models are possible. Nonetheless finite depth-limited recursion can be emulated simply by making the appropriate number of copies of the mechanism. But of course there is no need to constrain these to be identical copies, and we should therefore assume a simpler and more general mechanism.

Looking at what we understand about the neural process which evidently are responsible for our cognitive and linguistic processing, we don't see a push-down automaton - we see a network of very simple processors. It certainly is possible to implement a depth-restricted grammar using neurons - indeed it is theoretically possible to implement any electronic circuit using out of simple perceptrons, a model that assumes neurons fire when the sum of their inputs exceeds some threshold - but our starting point should be what we already know about the neural mechanisms in the brain and the type of processing they perform.

Artificial cochleae

Another basic empirical observation we can make is that the auditory world is presented to the brain in a particular way. The basilar membrane performs a transduction of the acoustic sound into frequency space - indeed the inner and outer surfaces provide high and low temporal resolution, and low and high frequency resolution, respectively. A computer model based on time resolution of around 22ms and frequency resolution of around 45Hz provides sufficient information to do basic speech recognition. The success of this kind of transduction model, varying the precise parameters and embellishing in various ways, is attested not only by a range of commercial speech recognition products, but by the success of the cochlea implant in which an artificial cochlea based on a relatively crude model of the human cochlea can successfully be implanted, connected up and permit a deaf patient to hear and understand speech!

Similarly we understand the 3-cone 1-rod structure of the retina reasonably well, and the RGB 3-color theory is sufficient to form the basis for color television. Again, artificial retina implants are being researched and look promising. So standard technology is easily able to provide the raw input for connecting our linguistic theory to the visual world. We do not reformulate this information overtly to coincide with the known properties of the organization of the visual cortex, but we will make comparisons between these known properties and the structures that emerge in our model. For some experiments may for convenience reorganize this information to provide features that are evidenced in neurophysiological studies.

Thus, although we are aware that our audio and video devices do not correspond exactly to the way the ear and eye are transduced, they are a reasonable approximation. Similarly, pressure, acceleration, motion, temperature and touch sensors can model a variety of human perceptual capabilities sufficient to commence exploring the connection of our linguistic theory with these aspects of our reality. The motor side, is similarly important for a grounded model of human linguistics, and although the similarities between motors and relays and human muscle-driven motor behavior are relatively slight, they also would seem to suffice for a first approximation.

Neural mechanisms

The basic neural mechanisms of perception seem to operate in a bottom up fashion. We are aware that the raw input from the retinal ganglia is processed layer by layer in the visual cortex giving us successively more and more complex features. The features of one layer are composed out of those of the immediately preceding layers, and self-organizing neural models can automatically discover or invent useful features [Malsburg, 1973].

There is clear evidence that neurons perform an associative function - recognizing patterns of inputs and outputs from other neurons. In addition neurons introduce delays, and thus a temporal element which is reflected in the depth of the neuron. The layout of the neurons typically form layers that capture spatial information related to a 2 visual field, to an exaggerated homunculus mapped according to enervation of different parts of the body, to the frequency space formed on the basilar membrane. A bottom-up approach where neurons recognize associations amongst a large collection of inputs, which may be direct sensory inputs or outputs from other neurons performing associative or other processing functions. These may reconstruct a lost spatial dimension (in the case of stereo vision and audition), they may detect features such as curvature or motion. Motion and trajectory are applicable not only in vision, but in auditory processing - e.g. diphthongs seem to be complex motions in a multidimensional frequency domain, whilst speech can appear as a trajectory in a "phonetic space" [Ritter and Kohonen, 1990].

Recursion alternatives

One of the consequences of assuming finite memory and a neural implementation is that recursion does not fit the model neatly - this means we are forced to abandon one of the favorite mechanisms of traditional linguistics, which begs the question: is this a problem?

Probably not! There is clear evidence that language is not recursive. Fries [] catalogued the structure of the noun phrase and showed that there was a distinct order in which elements of different type must occur - you can't arbitrarily turn "small green box" into "green small box". The ordering of the adjectives seems to reflect the degree of entrenchment - how fixed and inalienable the property is [Deane, Silverstein]. This kind of model fits the neural network well. A standard feedforward network has neurons that take inputs from only the immediately preceding layer, but higher order networks can take inputs from many other layers. The number of such layers corresponds to an automatic limitation on the degree of apparent recursion. Given that the layers are similar in nature, the resulting behavior will resemble depth-limited recursion.

Other constructs that are frequently cited as demanding recursive grammars are nested relative clauses ("the old woman that swallowed the cow that swallowed the ...") and prepositional phrases ("the wife of the son of the best friend of ..."). These are precisely the constructs that lead to the development of the depth hypothesis, which allowed unlimited right branching (like these examples) while restricting left branching. But perhaps what we have is just a list.

One notable characteristic of these unbounded right-branching constructs is that they are a challenge for memory and tend to involve long term memory (memorization as a song, as with the first example) and a logical sequence of connections (which the speaker has to already know, and the hearer has difficulty in constructing). However, syntactically these examples are completely undemanding because the ultimate length of the clause or phrase does not change the fact that it is a clause or phrase, and does not affect its syntactic relationships with anything outside the construct. This is what tells us that we have nothing more complicated than iteration or the kind of simple tail recursion permitted by a regular grammar. Viz. we can see that

```
S <- A B
B <- C B
B <- D
```

is equivalent to

```
S <- A {C} D
```

where the braces indicate a list of 0 or more instances of C.

Thus one possibility is that the right-branching, 'house that Jack built', use of 'that' really corresponds to a list, and at each iteration, we are adding another stroke to the painting, but there is no additional grammar-processing memory overhead. Our syntactic state before and after the additional clause 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 previous states. The repetition displayed in numbers and conjunction also has this immutable character.

A recurrent neural net also works like a painting: unlike a feed-forward network in which outputs of neurons from one layer always connect forward, or deeper into the network, it is known that cortical neurons include lateral connections and backward (or recurrent) connections. A recurrent layer is one that includes neurons that accept backward connections - it builds up a representation that gets modified each time round, with the new information and

the old information being combined to determine its new outputs. It also has a memory function as it presents current and stored information simultaneously.

To return to our painting analogy: there is a constant surface area, and new strokes (representing either new information, or a repainting revising old information) simply overlay what ever was there before.

Questions about sentences and worries about words

One of the features of our hard-science linguistics is that we must throw out, or at least question, all our traditional assumptions and the terminology that entrenches them. It is very difficult to discuss language without using words like "phrase", "phonetic", "word" or "sentence" because these are the labels which we use to discuss various chunks of "language". But all of these are suspect, and need to be reinvented from first principles using only the assumptions and reasoning tools of hard science.

Let us illustrate this by asking what is so special about sentences anyway? We can replace the full stops in any set of sentences by 'and' and have a new legal sentence. Similarly we can replace any multiclausal sentences involving conjunctions or relatives by a set of single clause sentences conveying the same essential meaning. Perhaps it is clauses that have the strongest grammatical reality, and the different ways of combining "clauses" tend to involve either extreme limitations of depth, or unconstrained iteration. The paragraph, and higher units, are not normally regarded as subjects for grammatical analysis, but they are no more or less real than the sentence, as pointed out by Pike[.]. The sentence has developed a special role in language largely since the invention of the printing press entrenched prescriptive grammars and rules of style, although it is still being casually murdered in everyday informal speech. Even then, most of us haven't mastered the correct use of punctuation because of the lack of correspondence between the written conventions and the prosodic cues.

Similarly linguistics has tacitly assumed that words are the starting point for our analysis and have a concrete existence. However there is no definition of word that adequately covers our usage in English, let alone carries over to other languages. What of "out of" and "put up with"? Are they or are they not words? Since nothing tends to intersperse or disrupt the order (unless you happen to be Winston Churchill dealing with something up with which you will not put), there is no reason to deny them the status of words. And comparable constructs in other languages, e.g. the separable verbs in German, are routinely treated as words even though interspersions and reordering are actually required for certain tenses. What of "today"? Or is it "to-day" or "to day" or "this day"?

The assumption that words are the primary atomic unit of syntax is one that certainly deserves to be questioned, along with the distinction between clitic and affix. Ideally the existence and definition of words would emerge naturally in a theory based on lower level assumptions and non-linguistic organizational principles. As with punctuation, the lack of correlation between words and phonological words is another symptom of this issue.

Similarly, at a lower level, we must consider the correlation between morpheme and syllable and whether either or both of these emerge as significant within our theory. At a lower level still, we can consider whether phones and phonemes emerge as consequences of a general theory.

The reality of rules and the fuzziness of performance

Why should we make errors in performance? This applies to my tennis serve as much as my speech. There's many a slip between cup and lip... 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, and 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.

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 makes 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. Many of these shortcuts then get enshrined in the language, as frozen lexicalizations or clichés, as changes to the conventional usage of a word (e.g. compare US/UK 'write me/write to me'), as telegraphic forms, acronyms and portmanteaus and catchy slogans, progressing from 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. 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 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 rules, and the apparently simpler form is belied when we move to a morphological level.

The brain seems to be based on neurons, which are often viewed as operating probabilistically. 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, and neural network simulations are able to self-organize efficiently in this same sense. We also saw that probabilistic approaches to grammar acquisition avoid the negative results about learnability. Our choice of words and 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.

On the other hand, there do seem to be stable points, inviolable rules in one language that are totally irrelevant in another – at all levels of structural analysis e.g. whether 'kn' is pronounceable or whether word order is rigid in a phrase. In a connectionist model based on neurons, it is relatively easy to see how probabilities come into play, and relatively difficult to ensure this kind of absolute rigidity. The rigidity seems to come about from interaction with other subsystems, such as the articulation or semantic decoding, where the rigidity is integral to the control model. If there is no other way to indicate a relationship, then word order is essential, especially when cliticization and prosodic processes are involved. If an articulation model for one language sees two modes as independent (e.g. position of back and front of tongue) whilst another sees only one mode (viz. position of tongue), then the system is incapable of translating the incompatible juxtaposition to co-articulation.

The frequency and frozenness of phrases and the entrenchment of concepts seems to have an effect on the complexity of the constructions in which they can be embedded, presumably because rather than relying on an arbitrary subsumption parameter, the complexity is really limited by memory and computational factors. The concepts of frequency, frozenness and entrenchment are all closely allied to probability, information and efficiency. The Zipfian

principle of least effort and the Shannon information theory both predict that more frequent information should be more accessible and impose lower computational loads. They will occur with greater frequency, and be accessed and produced with greater probability and greater rigidity in the sense that deviation from the set phrase will be increasingly unlikely as we proceed to utter it. This can lead to production errors as we slip into a well-worn pathway because a chance juxtaposition of units resemble a frequent sequence. Given a frequent and a less frequent confusion pair we might expect that we will make more substitutions of the frequent item by the infrequent in proportion to their raw frequency, consistent with a null hypothesis that errors occur independently of the probability of occurrence of the item. Conversely, there is some evidence (based on analysis of typos like "are/our", "here/hear" [Powe96]) that we make relatively more errors in the direction of the more frequent item as suggested by our well-worn pathway concept.

Without attempting to crystallize an actual model at this point, but merely looking at the commonalities among known neurophysiological mechanisms, connectionist models and statistical learning paradigms, we see that there is considerable scope for development of a model in which rules are probabilistic rather than absolute.

With such a model, performance errors would occur probabilistically precisely because our competence is based on probabilistic connectionist representations. Nonetheless, although probabilistic models (based on a Markovian assumption of limited context) are used frequently in speech and language, they tend to be employed within a single modality and not reflect the cross-modal interactions that seem to be necessary for even correct phonemic, morphemic or syntagmatic identification. Thus I am not suggesting that the Markov assumption, or the employment of probabilistic grammars, is a panacea, but rather that probabilities and conditional probabilities play a complex role in normal language competence due to the plexus of relational associations and the probabilistic nature of the connectionist wetware that forms them.

So let's consider what predictions could be made from the claim of absolute grammaticality. The inviolability of grammatical rules should be automatically determined in any appropriate probabilistic framework – the probabilities on the connections corresponding to rules should all be high (unity, once conditioned appropriately), and associations which correspond to no rule would be only at chance levels. We could define a model which used a threshold to distinguish the two cases and would not reach the threshold for incorporation as rules, and associations that correspond will reflect deviation from 100% obedience only due to performance errors. Moving from a model without thresholds to a model with thresholds, where probabilities approaching unity were forced to 1.0, should produce better, indeed perfect, performance.

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 since the absolute framework is a special case, but it is important to understand in what sense and at what level probability and fuzziness comes into play. Probability theory allows us to deal with limited sized samples of a large or even unbounded corpus.

If we assume that we are trying to learn a specific language, but we only have a sample of the set of possible sentences, probability theory could tell us what to expect given we knew the rules of the language and made appropriate assumptions about how the sample was derived from the full corpus. A typical assumption might assume that the language consisted of sentences and sentences of a particular length were equally likely. Empirical observations could then be used to formulate a model of the distribution of sentences with length as well as theories of what factors and mechanisms might lead to this particular distribution holding [Zipf,1956]. Probabilities thus arise as an artifact of our exposure to a sample of the language, with the selection of the sample being determined by such factors as the cognitive

mechanisms involved and the nature of our environment. In fact there is nothing special about language in this respect except that there is no evidence of there being a specific target language, which was our starting point. If we replace 'language' by 'ontology' and 'corpus' by 'ecology', we come to a similar conclusion that our model of the world is probabilistically in the sense that our experience is a sample of the full range offered by our environment.

Thus our individual idiolect (as opposed to some target language) and ontology (as opposed to some target thesaurus) are built up from our experience and necessarily have the probabilistic character appropriate to a model derived from a sample. The precise form of our internal models is dependent on the nature of our environment, including the nature of our social and communicative interactions - and it makes little sense to distinguish between linguistic and other models - all of them are part of one multifaceted ontological model (which need not be internally or externally consistent in any objective sense). The language or dialect that provides the means of common communication between members of a group may have no real existence - members who communicate frequently with each others will negotiate an overlap of their communicative. Languages, sublanguages and dialects are convenient labels, but their reality may be more political than linguistic in the sense that no two speakers use language in the same way (although prescriptive linguistics in the age of the printing press has gone some way to defining canonical languages).

Our internal ontological models are also dependent on the precise nature of the mechanisms we use to make sense of this environment. These ontological models are further complicated by a kind of recurrence: the mechanisms we use to make sense of our environment (develop an ontology) impose a particular structure and a particular selectivity to all our experience, whether we regarded it as being subject to linguistic, social or physical laws - all of these are instances of ecological laws and are subject to complex interactions: the environment influences our ontology (and our language or idiolect is part of this); our ontology influences our communication (which is heard by ourselves and others); our communication influences our environment (which includes both our communications and others'). Furthermore the mechanisms we use to comprehend our environment help determine our ontology and our role in the ecology, whilst the nature of the ecology influences which mechanisms are most appropriate to deal with it.

Probabilistic and absolute models

The robot baby project uses algorithms based on neurological, statistical and information theoretic insights. The initial models are designed to deal with the physical world through sensory-motor interaction, and are not specifically designed to deal with social or communicative interactions. Basically they are looking for patterns - ubiquitous patterns involve relationships that are overwhelmingly consistent, and can be construed as rules or laws. Empirical rules and laws may not however appear to hold 100% of the time for a variety of reasons: e.g. we may not have seen all possible instances of a class (a rule that all members of a class act a certain way will not in general be strictly verifiable); there may be overlaps in class membership (homonyms); there may be extraneous influences which we haven't taken into account (illusionists seem to defy the natural laws); there may be performance errors (due to tiredness, stress, lack of time); the speaker may not be a native speaker (*viz.* does not obey the conventions that constitute the *lingua franca* of the group); there may be constraints on the communication medium or the communicative participants (telegraphic speech, telephone speech, lip-reading, motherese).

By the use of thresholds or inhibition/selection we can ignore the occasional or systematic exceptions or errors, and our statistical mechanisms can be made to give absolute judgements, but the more natural usage is for them to give relative indications. Simple thresholds deal well with simple error models describable in terms of low error rates, but are or relatively

limited utility. Inhibition allows more complex exceptions to be dealt with, and permits active identification of elements that license an alternate system of rules or associations).

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 or the order of exposure to different constructs 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. Similarly, changes in environment can lead to changes of register or sublanguage, and more flexible mechanisms will allow adaption to different environments and automatically give rise to such phenomena.

Our model is not designed to lead to the development of formal phrase-structure-based grammars, but the mechanisms of thresholding and inhibition/selection along with additional assumptions and heuristics can be used to extract phrase-structure (and functional) grammars. These turn out to be broadly consistent with our linguistic insights, and where there are surprises they do provide additional insights.

We now make a final connection. Our statistical model can be implemented using neurons and neurophysiologically attested mechanisms, and this further enhances its plausibility. Neural learning is all essentially probabilistic in nature.

Learnability and supervision

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 a target language to criterion. What is important is the paradigm, the context or environment in which learning takes place. What kind of information is available to the learner? How clearly does it indicate what the correct form or rule is, or how much is left to be induced or deduced. Is there clear feed back that a particular form, rule or utterance is wrong, or is there clear information from the context about the role of a particular word in a sentence? Machine learning reduces this to a very simplistic concept, that of supervision. The well-known learnability results of Gold assume a sequence of sentences without supervision, whilst the normal level of supervision envisaged in learning theory would simply 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 machine learning. 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. A simple approach could assign each word the appropriate part of speech labels, and then collect a set of one-level sentence-rewriting rules, one for each pattern that occurred in the corpus. A more sophisticated approach seeks to find recurring fragments and define separate classes and rules for them. But in either case, this paradigm would seem to be 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 be 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 let us discuss what kind of information is available to the child. There is some evidence [Turk] that the child has a repository of recognized sentences or fragments that 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. In such a context, we can now consider mechanisms and algorithms that can make use of the additional information.

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 will seek to reflect the total environment in which the child finds himself.

This traditional view of supervision is, however, usually predicated on the idea that we are trying to learn a target language or concept, and this is the context in which we have considered it here. However, we have earlier argued against the existence of such a target language in the traditional linguistic sense, but if we consider mechanisms designed to learn about the world, and consider feedback from the environment as a form of supervision, we can see that this concept is relevant to the learning of an ontology given that there is an objective world out there - the target language or concept. Furthermore once we have such mechanisms and algorithms, designed to learn about the world, and we expand our concept from the physical world to the full ecology, which includes social and communicative interaction with other agents, we see that communicative behavior that fits the structural interpretation we place on the world should be learnable through the same mechanisms, and feedback about social and communicative effect should constitute feedback in the same way as success or failure dealing with the physical world provides positive or negative feedback.

This leads us again to a holistic view of language and ontology - language being a name we give to particular communicative commonalities in our sensory-motor environment. Language learners - which strictly speaking includes everyone, not just infants and newcomers - apply their ontology learning mechanisms to the totality of their experience of the ecology, including the perceptual and social effects of their own behavior. What produces, directly or indirectly, interpretable percepts and desirable outcomes can be regarded as being positively reinforced - and structures that are similar to previously encountered and analyzed communicative behavior will also have an advantage. That which doesn't produce positive reinforcement, or indeed gives rise to overtly unpleasant or disadvantageous effects, can be regarded as negative reinforcement. Furthermore, we are under no compulsion to assume a binary dichotomy of positive versus negative reinforcement - a spectrum of reinforcement of different degrees and different kinds is far more likely, and probably more powerful.

[Bod; Turk, Clark; ASA; Popper]

The Robot Baby



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.*

Building a Baby

The archetypal language learner is the baby, and hence our project focuses on reproducing the capabilities of a baby, hypothesizing likely cognitive and perceptual learning mechanisms, and exploring what kind of linguistic and ontological structures emerge. We have argued that a statistical or connectionist basis is appropriate to be consistent with the known neural substrate for cognitive and linguistic processing.

A baby needs to have basic auditory and visual capabilities, and should also know when it is touched, dropped, lifted, turned, fed, etc. It should also have the capability of interacting with the world, making noises, moving head and limbs, directing its visual attention, etc. We have seen that current audio and video technology is a good model for human audition and vision, even to the extent that prosthetic devices have been developed for each, and we have incorporated these a variety of other sensory-motor capabilities into our system. However, they are not a perfect match, and therefore it is important to evaluate to what extent the imperfect model will impact on the results of our learning model.

The robot baby project has been pursued by the author for over twenty years, although only in the last year has the full range of sensory-motor interaction with the real world been available - and even now, the cameras and some of the microphones are externally located for technical reasons which are beyond the scope of this chapter. Our first robots with the capability of interacting with its environment were simulations (in common with those of most other researchers [Winograd, Feldman]). A 3D computer graphics simulation [Hume,Homes] allowed for multiple actors (controlled by program or from the keyboard) to interact in a world in which the basic physical laws were simulated - e.g. objects couldn't pass through each other. Some objects were fixed (landmarks) whilst others were movable (motile) by the actors. Complex systems of joints were also possible (viz. objects or parts could be partly fixed). This simulated world provided considerable flexibility - objects could be dealt with at the level of the basic geometric shapes/features, in a choice of absolute or relative coordinate systems - or could be dealt with at higher levels with concept names directly returned when the object was sighted. The simulation had its own world model which was manipulated to

generated the visual display (for the person at the keyboard/screen) as well as a tree-structured representation from the perspective of any of the simulated eyes.

In using the original model, both language and ontology were constrained to fit the traditional phrase structure mould (usually a context-free grammar although occasionally a unification/feature grammar was used). Semantic learning thus consisted of looking for relationships between the parse-trees corresponding to sentences and those that corresponded to ontological information. Simple concrete nouns, action verbs and prepositions were learned [Powers:1989; Hume; Chan; Homes] using paradigms with a fairly high degree of supervision (although with the work on prepositions, we explored whether the correct word, landmark and trajector could be found automatically rather than identified by the teacher).

The work on learning prepositions [Homes:1998] is of broader relevance and is a touchstone task for the L0 project too [Feldman]: in our experiments, an object is moved around and a subject labeled the scene either with a single word or a phrasal preposition (like 'to the right of' versus 'beside') or a full sentence. The point of this complex formulation is to move away from the supervised paradigms where the 'correct' word is associated with the scene and the whole issue of how the learner focuses on the correct parts of the utterance and the scene is overlooked. The algorithm needs to be able to decide whether to attach meaning to an individual word or to a phrase, and it also has to deduce which object in the scene was the landmark and which the trajector (if it selects something else it would not find any useful correlations, but if it reverses them it would learn inverse relationships – left for right, etc.) The program then has to make hypothesis about the various Cartesian relationships that hold and see which are consistent for a particular preposition (word or phrase). In this supervised context it is easy to learn the teacher's concept for each preposition – which varies somewhat between individuals, and very significantly across languages (we compared English and Ukrainian in this experiment). In the full sentence version, an additional complication is present: it is now necessary to focus on the correct word. This was not reliably achieved in the absence of 'knowledge' of other words (open or closed class) in the sentence - it would seem that syntactic information should be used to tease this out.

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 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, and these have been pursued at an early stage as a means of evaluating the sufficiency of the hardware to the learning application.

As part of verifying the suitability of our sensors for the ontological learning task, we considered how well we could do speech recognition with the available sensory information, and sought to ensure the robot baby had the capability to locate a speaker aurally. We also wanted to see if there was sufficient information in the visual stream to lip-read or at least improve the speech recognition processes. 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, the environment will tend to be noisy and the doll's electromechanical components will also introduce noise, as will the rustling of clothes if the doll is being held or rocked.

We have therefore performed experiments with a tetrahedral array of microphones (located at the ears, nose and crown) which is sufficient to localize a sound in three dimensions, and are using conventional blind signal separation (BSS) techniques to locate and isolate up to four

sounds - in theory. In practice BSS is not very effective on sounds recorded in real rooms with lots of reverberation and irregular obstacles. Thus we are looking to explore a composite approach in which learning of ontologically significant events is done using full input, albeit at relatively low quality.

Another set of experiments involved using stereo cameras with associated microphones, and here our focus has been on whether we can use visual cues to improve on speech recognition and speaker location. At this stage we have some promising but preliminary results (e.g. 38% improvement over pure auditory phoneme recognition of stops and nasals). We can track the lips well, and for some phonological features (e.g. bilabial) we can provide useful additional cues that considerably enhance the baby's phoneme recognition capability. For these experiments the best contrast enhancement function we discovered, which we call red exclusion [Lewis,Powers:2000], takes into account that red is the predominant color in skin tones of all races and therefore seeks to use the other colors to maximize discrimination. This probably relates to the yellow-blue discrimination in the component color theory, but we have yet to explore to determine which precise shade in the yellow-green range gives optimum performance - we used the standard RGB green as provided by the camera. Our statistical or connectionist algorithms should be able to automatically determine an optimum opposition in order to distinguish faces optimally, and we expect that this may be more related to the color of blood which gives the skin its red flush, than to the peak response wavelength of the red cone.

Finally, we hope to that an auditory speaker identification capability will contribute to the ability to form a model of the people interacting with the baby, and that this model will also be usable to attend to a particular speaker, helping to localize her and isolate her speech.

As with all of the experiments discussed here, this is work in progress. Although it may seem less relevant to the modeling of language acquisition, we see it 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.

Hypotheses underlying our model of a learning individual

We have made a number of specific definitions, hypotheses and implementation decisions that are the basis for our robot baby model and our grounding of hard-science linguistics, all of which relate to the modeling of a learning individual. These hypotheses concern firstly the interface with the world, which gives rise to percepts, being the inputs at the bottom level of our model of communicating individual, and secondly the nature of the mechanisms which determines the structuring and processing of these percepts. The mechanisms are intended to reflect potential connectionist processes, but are not always expressed or implemented as neural networks. The details of how such non-connectionist models would be mapped on to neural nets has not been worked out for all models, although some models have been formulated using neural nets, and all have been designed with the intention that such a mapping is feasible.

We outline one specific model here for which a significant degree of structuring has been achieved in preliminary experiments, although it is to an extent idealized as experiments at different times and by different students/collaborators have explored different variations. Many alternatives have also been explored, differing both in the nature of the mechanisms or the precise parameters used (around 6000 variant definitions of similarity were explored in

[Powers,1997]). The results will be outlined for the general model presented here, which is the one that has been analyzed most fully, but some variant will also be presented.

Percepts

1. Our perceptual model is implemented by means of cameras, microphones and sensors and standardized preprocessing. This model is supposed to be adequate to the task of learning to communicate as a human, but while these decisions have the character of hypotheses the precise form they take is partly determined by the hardware and parameter settings. Separate experiments have been performed to verify that appropriate discriminations can be made using this perceptual model.
2. We define a basic level of percept that reflects the kind of sensory information presented to the brain by each kind of sensory organ.
3. For vision, we approximate the human percepts with a stereo pair of 352x288 24-bit 20 fps video signals (either located on a monitor or in the doll).
4. For audition, we approximate the human percepts with a 16-bit 22 kbs audio signal transformed to frequency space with a resolution of 45-50Hz and 20-22fps (22fps is more standard, but 20fps synchronizes with the video signal. 10 microphones have been used for our experiments (wiring both the room and the doll).
5. For acceleration/shock the robot has a huge dynamic range that provides a signal of the order of 20 mV for gentle movements and 100 V when dropped a meter or so, but it is unable to determine direction of the force and is more sensitive to a vertical force.
6. To determine body orientation the robot has a crude sensor that can distinguish 8 orientations.
7. 8 simple switches act as touch sensors distributed around the body and limbs.
8. The robot can turn its head through $\pm 30^\circ$ and detect the orientation to within approximately 1° .
9. Currently each limb is a standard doll's limb shaped appropriately for crawling and is controlled by a single motor at the shoulder or thigh.
10. For some experiments, structuring is performed assuming phonemes, characters or words as inputs, the most comprehensive experiments having been performed with character based input. The most thorough experiments have been carried out with character-based input in English and Dutch. In general, good results from text have also been achieved even after preprocessing to eliminate spaces, capitalization and other cues that may not be available in raw speech.

Mechanisms

1. Structuring involves segmenting or grouping inputs as well as identifying classes of similar segments.
2. Segmentation is achieved implicitly by forming multiple hypotheses as to the appropriate segmentation. Different versions of the model may or may not require a unique segmentation, allow segments from multiple levels to participate, or allow a class from the same level to participate recursively or recurrently. However, the most impressive

structures seen have been achieved under the condition that unique segmentations are used and only classes from the immediately preceding level are available.

3. Classification is achieved by identifying sets of contexts that admit the same set of segments and considering them as a putative class.
4. The more contexts in which a putative class can occur, the stronger a candidate it is. The strongest candidate is denoted as a class and identified with a new non-terminal symbol which becomes an input to the next level of processing. The input sequences of symbols/units forming the segments constituting the class may then be rewritten using the new non-terminal symbol, either replacing the original segments or as alternatives to them. The most impressive structures have been achieved under the condition that the rewritten segments are no longer available.
5. Initially only inputs deriving from a single modality, e.g. vision or audition, are considered. The intention is to derive such structure as may be identified in each modality separately without recourse to other modalities.
6. Subsequently cross-modal associations are permitted that can also make use of the structure discovered in the individual modalities. This has the effect of reducing the combinatorial explosion that would occur if all combinations of raw percepts were considered as potential patterns.
7. Parameters include the size of the left and right contexts (optimal is 2:3), the maximum size of a segment (2 suffices), the maximum number of levels of feedforward (only 1 explored) and the maximum number of levels of recurrence (only 0 and 1 explored). No other form of memory or constraint is provided.

Phonology, Morphology and Syntax

The first experiments performed in this general model were performed with words as input, and both statistical and neural algorithms were employed. Statistics were collected about how often particular words seemed to group with words or higher level constructs to the left or to the right (giving a total of four possibilities to be investigated for each word) or neural networks were used to learn comparable associations. The most interesting aspect of these statistical and neural experiments, based on quite different approaches, although all incorporating some form of memory limitation, was that the closed class elements (including punctuation as well as words like articles) tend to emerge as the critical elements of the language - and experiments using telegraphic utterances in which these closed class items were omitted failed to discover useful structure. The closed classes tended to act as seeds, or heads, around which the rest of a phrase or clause grew like crystals.

The techniques used in these experiments are closely related to the CIE/CAE method of Contrast in Identical/Analogous Environment used by Pike [1950] in discovering phonological structure, although we first applied them at word level. However we also performed experiments to see what kind of structures emerged if we commenced with characters or phonemes - proceeding to discover closed classes and segmentation/phrasing from the phonological level through the morphological, word, phrase and clause levels [Powers:1991;1992]. Some preliminary experiments were also performed starting with speech code vectors (Percept 4) and demonstrated similar segmentation to the character experiments [Schifferdecker,1984].

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 tend to find at the level of words. A, B and N represent classes of articles, adjectives and nouns that are not shown, T and V represent transitive and intransitive 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 worse looking grammars (including totally degenerate grammars) compared with the standard version of the algorithm. While the grammars would be easier to understand if labeled with standard English non-terminals, the classes are discovered by the program and labeled 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 (CIE/CAE), 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 labeling the induced class.

Notice that this approach in one step groups sequences of units together as a filler simply by not restricting the filler to be a single unit, so that it brings simple and complex fillers together as a class. A related two step approach to unsupervised learning which has been developed independently by a number of researchers [Langley, Grunwald], involves alternate 'merge' and 'phrase' phases (to use Langley's terminology). In one phase terminals or non-terminals are merged into a new class/non-terminal so as to achieve an equivalent but more efficient representation. In the other, sequences of symbols that occur in multiple rules are labeled with a non-terminal and summarized with a new rule so as to achieve a shorter representation. This is based on an idea of parsimony known as minimum description length, and related to Shannon's information theory [Shannon] and less directly, to Zipf's principle of least effort [Zipf].

Experiments based on this two step approach seem to be able to develop reasonable grammars based on a smaller corpus and a restricted language, but the restriction to strict CIE means they are less able to deal with large open classes. Nonetheless we have succeeded in learning a reasonable (proto)grammar for a small corpus of child-directed caregiver speech using this technique [Ceglar:1999]. Both algorithms make the assumption that words belong to a single class, but with the relaxation to CAE, complete parses are still possible but without distinguishing separate noun, adjective and verb subsets of the open class words a higher degree of ambiguity results. Generally, we find the correct phrase structure, but do not distinguish between noun-like and verb-like classes.

Note that we have made no assumptions about the existence of phrase-structure grammars, although the above classes can be interpreted as context-free grammars. In fact, the 'grammars' produced tend to be highly ambiguous, and except to demonstrate that parses can be generated we do not do so. Our rules show us many different sensible constituents at many different levels. For the time being we just leave it at that. In the robot baby experiments there is another part to the story - the learning of the semantic relationships between structures that emerge in the speech or text stream and the structures that emerge in the ontological analysis.

[Miller; Langley, Grunwald, Kirby; Aaron/Oliver; Shannon, Zipf]

Prepositional syntax and semantics

In our discussion of the learning of prepositional relationships we faced the issue that prepositions may consist of one or more English words, and that we had a problem of matching up nouns or phrases in the sentences with complexes in the ontological representation. Our aim now is to see if the structure provided by our unsupervised structural analysis assists the process of learning semantics.

Prepositions and particles are particularly significant in terms of our unsupervised learning of syntactic classes and rules, as they tend to be preceded by verbs and followed by noun phrases (and in particular articles). They are thus fairly distinctive syntactically. Indeed, as a closed class they are relatively frequent compared with most nouns and verbs, and they provide more specific information than the more frequent generic nouns and verbs. In particular, a child attending to an object (possibly himself or a body part) that has fallen or moved out of reach, or changed state in some other way, will tend to hear something like 'fell *down*' or 'rolled *under the chair*' and will thus tend to associate the preposition or particle with the type of movement. The same applies to questions and prospective state changes ('Do you want to get down now?'). Common words like 'get', 'fell' and 'chair' with relative high frequency for open class words have comparable frequency and should be learned around the same time, but less frequent nouns and verbs will tend to be learned later but should have learning facilitated by association with a preposition/particle that had been learned earlier.

Clearly these predictions of the model require confirmation from both experiments on the child's comprehension of these words (at 0;1 to 1;0) and the model's learning of these words in a context of sentences being associated with scenes. This can most convincingly be performed with a doll- or baby-sized robot learner. Experiments in a simulated world are too easily influenced by the means used to make items salient in an already oversimplified visual world. On the other hand, the experiments with real vision face considerable challenges in the area of visual processing. 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.

Language Evolution

The Artificial Life community has studied many aspects of social evolution, including the conventionalization aspect of communication. Steels (1996,1997) has been working in this area from a linguistic perspective using both real and simulated robots.

To date the major results of this have been in the spontaneous emergence of a lexicon and studies of language change as different subcommunities of robots meet and interact. The robots make contact and adopt the role of speaker and hearer in a shared context, with the speaker choosing an object and drawing attention to it as the topic of conversation. Initially this restricted lexical acquisition to concrete nouns, although some successful experiments have been performed where this has been extended to other semantic relationships including the use of collocations. In an artificial world, arbitrary features can be used, and the experimenter directly controls which are sufficient for the kind of semantics s/he wishes to explore. When using real robots with real sensory data, similar simplicity has been achieved in the early experiments by using primitive sensory data – e.g. based on patterns of LED flashes. It is a real challenge to move to full video data, in which case the features can again be handcrafted or can be developed by self-organization and auto-association, and this kind of work is at an early stage.

Steels' speaker chooses a feature set that distinguished the target from other contenders and encodes this using the lexicon (and grammar in experiments that explore this aspect), inventing a new word from the repertoire of possible words. The hearer decodes the expression, and may have to decide between synonyms or deal with an unknown word, with feedback from its expectations based on the shared focus that had been established. Under this regime, vocabularies within an interacting group tend to stabilize very quickly and without experiencing deleterious effects due to the combinatorial explosion of possible lexical assignments in richer environments. Rather, as more words are learned, and the set of remaining possible words decreases, there is what Steels describes as a combinatorial implosion – it stabilizes faster. But the assumption of a fixed and small (and eventually exhausted) set of possible word types and meanings is a major consideration and is a variable that requires systematic examination.

In the physical robot world, an entire ecosystem has been built in which robots have to recharge themselves, deal with competition, and cooperate to survive. This is hoped to provide similar opportunities for language conventionalization without the explicit paradigm of gaze direction and description which is reminiscent only of certain very specific early language learning experiences.

Data Collection

At the moment the dolls we are building are still not in a state where we can place them with young children, and there are issues relating to weight and robustness that we have yet to deal

with. Our present experiments involving real vision have been very restricted and many of our ontological and semantic experiments have used a simulated world [Hume;Homes] rather than real sensory-motor input. Nonetheless, we expect soon to be able to place dolls with 3-6 year old children and produce a comprehensive corpus of both their interaction with the doll, as caregiver, and their interaction with their caregivers as infant. This is another way in which we believe we will learn a lot from the robot baby, as one of the biggest issues in understanding the language learning process is the need for large comprehensive corpora giving full sensory-motor records of a child's linguistic and ontological inputs and outputs.

A model of speech understanding and production

Percepts in human linguistics