Biologically-Motivated Machine Learning of Natural Language and Ontology

A Computational Cognitive Model

David M W Powers

School of Informatics and Engineering
Flinders University of South Australia
PO Box 2100, Adelaide 5001, South Australia

David.Powers@flinders.edu.au

Abstract

The individual cognitive science disciplines all have contributions to make to the understanding and modelling of human learning. Our previous research has explored unsupervised learning of phonology, morphology and low-level syntax, as well as basic noun, verb and preposition ontology and semantics, plus musical and speech prosody. Successful applications using a mix of supervised and unsupervised techniques include speech control of equipment, deep web search, confused word spelling correction, multi-lingual semantic models and audio-visual speech recognition.

Our current research is focused on doing simultaneous learning of ontology, syntax and semantics by embedding the learner in realistic situations and by developing low-level biologically-plausible models of perceptual and cognitive processing.

Keywords: Ontology, Ontogenesis, Stereology, Stereophony, Stereopsis, Stereognosis, Stereosemy, Cognitive Linguistics, Computational Linguistics, Sensor Fusion, Data Fusion, Neurological Models, Neural Nets.

1 Introduction

1.1 The Total Turing Test

The dream of an intelligent, thinking, learning machine is older than the computer itself, both in science and science fiction. While some AI texts will go back beyond the Turing Test, it was turing who formally recognized that our understanding of the terms learning and thinking is intrinsically defined by our subjective experience of our own thought processes and our interactions with other people. Turing (1950) not only defined an objective test to capture the intention behind these subjectively defined terms, but predicted that to achieve human equivalent performance it would be necessary to arm a robot with "the best sensors money could buy" and it would then learn to understand and interact with the world in the same way a person does – this thus introduces the Symbol Grounding Assumption, that language needs to be *grounded* in the real world.

Turing also predicted that by 2000 a computer would be able to fool people into thinking it was human for 5 minutes or more at least 30% of the time – this was actually achieved by a Loebner Prize (see loebner.net) competition bronze-medalist in 1998 (with 10 'judges' of whom 5 were experts in some aspect of Cognitive Science and 5 were selected to be representative of the Australian population generally, as were the 'confederates' the programs were compared with). More informally, Weizenbaum's 1966 Eliza program was already so convincing that people who were not expecting a program believed it was human and would not believe it was a computer when they were told. Weizenbaum's Eliza was so successful that he felt obliged to write a book (Weizenbaum, 1976) decrying the idea of using it as a surrogate psychiatrist. This behaviour is a manifestation of 'the Charitable Assumption' – our social and linguistic default is that other people are like us and have a similar background understanding and beliefs of the world, society and language as us, often leading to surprise and/or repairs when violated.

When a Natural Language or Artificial Intelligence expert looks at Loebner Prize or Eliza scripts, it is clear that they are a long way from what we would expect of a human-human conversation. This is why Loebner is not happy with just a 50% pen-pal success rate but wants evidence of understanding, of grounding, and has therefore required that there be an Audio-Visual component to the competition – the Gold Medal winner needs to demonstrate an ability to talk about the world, needs to demonstrate that it is grounded, and a proposed 'Total Turing Test' or 'Gold Medal Test' is based on the kindergarten paradigm 'Show and Tell' (Powers, 1998).

1.2 Ontology and Ontogenesis

For centuries philosophers have discussed the nature of existence and the term 'Ontology' has been applied to this area. Since 1980, I have coerced this term to refer to the 'theory of existence' that a child develops implicitly as it learns both about the world and to communicate linguistically, as the basis for the grounding of language, and in particular of both syntax and semantics. From around 1990 it has come into increasingly common usage in computational linguistics to refer to formal handcrafted representations of the relationships between concepts, however in this paper I continue to use it to refer to an individual human's unconscious 'theory of existence' that is perceptually grounded in his social and physical environment. I want to distinguish clearly the pseudo-semantics and pseudo-ontologies, that are mere relationships between symbols, from the true semantics based on *grounded* ontologies whose relationships with percepts derive directly from the real world, whether in the human brain or a robotic brain or a brain model.

Biologists have traditionally used the world 'Ontogenesis' to refer to the unfolding of the structure of an organism, and more generally a biological unfolding of events. Whether this term in its true biological sense rightly applies to language depends on both the nature of the structure and events envisaged in the traditional definition, as well as the theoretical paradigm of language development. In the 70s, language was assumed to be innate, and mediated by a specific Language Acquisition Device, and thus the development of language fit the definition of unfolding of events and/or structures. For those that reject this viewpoint and recognize the role of self-organization, learning and environment in the ontogeny of language, the use of the term can be understood as a metaphoric extension and is not intended to have any implications as to whether language is innate or acquired, a product of nature or nurture.

More recently the recognition of the inadequacy of the genome to determine biological and concomitant structure has led to the use of the term 'Epigenesis' to recognize the role of self-organization and environmental influence, whilst 'Phylogenesis' emphasizes the genetically determined distinctions between species, giving rise to a three-dimensional characterization with Phylogenesis, Ontogenesis and Epigenesis as the axes (Sipper et al., 1997), which is still habitually subsumed into Linguistic Ontogenesis (Slobin, 2004).

1.3 Stereology and Stereosemy

The term 'Stereology' refers to the development of a mensurable understanding of the 3-dimensional objects and relationships based on lower dimensional sensors and representations. We apply this term advisedly to the 3D interpretation of the multiplicity of sensory-motor percepts that are interpreted by the brain. The terms stereophony and stereopsis, as well as the less familiar stereognosis, refer to the limited stereology that occurs within the single modalities of hearing, sight and touch respectively. A multimodal analysis can provide information that is not available within a single modality and thus the fusion of information from multiple modalities can provide a more accurate and detailed Stereology. Moreover, there is evidence that motor control, feedback and proprioception, as well as intention, functionality and purpose, all play a major role in the development and nature of our Ontology, and a complete understanding of the sensory-motor extends to the nervous, circulatory and immune systems, so we extend our definition of perception to include the full sensory-motor range (Powers and Turk, 1989).

At the level of cortical structures, it is well known that the sensory and motor cortices evidence 2-dimensional homuncular representations of the human body – distorted homunculus pictures representing the relative areas associated with different parts of the body are ubiquitous in basic Psychology or Neurology textbooks. More controversial is our hypothesized backprojection of the entire homunculus onto a specific organ – this underlies the homunculus theory basis for natural medicine techniques such as iridology and reflexology, but also makes strong theoretical predictions about mechanisms for learning associations.

Whereas Stereology refers to the abstraction of 3-dimensional information from 2-dimensional input, there is an additional interpretive step in associating diverse percepts and components with a single object or event – this we will call 'Stereosemy' and it is 4-dimensional in nature as time is an essential additional dimension used to associate percepts that belong together. Stereology refers to an objective reconstruction of a *mensurable* 3-dimensional representation, whilst Stereosemy refers to the subjective imputation of a *meaningful* 4-dimensional representation of event and object integrity in space-time. For parts, aspects or attributes of imputed events or objects to be associated in the Stereosemy requires that they be proximate in both space and time, with biases towards contiguity, convexity and connectedness – that is there should be a trajectory or continuity connecting the parts or else there is an expectation, triggering surprise, that something is hiding the connections. In a spreading activation model (Anderson, 1983; Deane, 1992) the relationship between object parts, cores and wholes is dependent on the integrity of these connections.

1.4 Synchrony and Assembly

Adding the dimension of time introduces sequence and simultaneity as aspects of sensory-motor fusion. This leads to a 4-dimensional correlation of sensory-motor percepts. To be related in the Ontology, however, does not require the continuity biases that characterize Stereosemy, but rather the Ontology is characterized by a conceptual or taxonomic representation mediated by similarity and without any requirement of space-time proximity. Thus at this point we are focussed at a much lower level of abstraction than that of Ontology.

Events, internal, external or reflexive, logically and chronologically correlate across all pertinent percepts, and objects we treat as a special case of events made salient by their appearance, disappearance, alteration or motion. Furthermore, there is evidence that entire networks of neurons that represent information about an event, whether direct percepts or higher level concepts, fire synchronously forming (cortical) Hebbian cell assemblies that tend to fire with the same pattern and in a fixed phase relationship. The different frequencies seen in EEG are hypothesized to relate to the round trip time for such cell assemblies, the mechanism is hypothesized to be direct or indirect recurrence, whilst the function is hypothesized to be the binding of the diverse cell clusters to represent an event (Lutzenberger et al. 1994, Fingelkurts et al. 2003).

It is clear that, at the perceptual level, the correlation of information can serve to identify a single event or object that is perceived in multiple ways or across diverse portions of the sensorium. This might be a clap perceived in terms of motor activity, touch, pain, generated sound energy (at a specific range of frequencies with a specific temporal pattern) and changes in the intensity of reflected light (at a specific range of frequencies in specific parts of the field of view), or it might be a dog that barks (at a specific range of frequencies with a specific temporal pattern), is seen to have particular colour, texture and motion (again perceived visually in a variety of ways), is smelt, and felt, with associated motor actions or responses (active patting or reflexive withdrawal of the hand). The same mechanisms that correctly identify intrinsic aspects of an event or object will also identify extrinsic (e.g. linguistic, metaphoric or superstitious) aspects of an event of object since there is no way of distinguishing perceptually that the colour of one person's hair is intrinsic (natural) and another's is extrinsic (dyed). Children have to learn that the names and conventions surrounding objects and events are arbitrary rather than tightly bound to the object. The field of Cognitive Linguistics (see e.g. Deane, 1992) is built on the hypothesis that all of language, and in particular phonology, syntax and semantics, is built on this kind of mechanism, which identifies a variety of relationships that we may in different contexts refer to as emic, analogous, metaphorical or metonomous.

2 Applications versus Models

2.1 Artificial Intelligence/Cognitive Science

In the earliest days of Artificial Intelligence, the field was dominated by Computer Scientists with significant representation by Psychologists who saw the computer as an opportunity to model their theories. Linguistics, however, also had its connections with Computer Science, but these tended to be more closely allied to Compiler Technology and Abstract Machines than to Artificial Intelligence – as illustrated by the cross-fertilization that lead to the understanding of the Chomsky hierarchy of languages in terms of both syntactic constraints on languages and memory constraints on machines. Similarly there was more of a connection between Neurology and Computer Architecture than with Artificial Intelligence.

The advent of Cognitive Science was born in large measure by the inconsistency and incompatability of theories that had developed in one field but had implications for another. In particular, Chomskian Linguistics made assertions or predictions that impinged directly on Biology, Psychology, Psycholinguistics and Computer Science. In my case, I was concerned that Natural Language Processing was largely proceeding in ignorance of Linguistics and Psycholinguistics, and that Biologically-motivated Neural Networks were at first denigrated as useless and then largely replaced by Artificial Neural Network models that explicitly rejected the constraints of Biology. Furthermore, there were constraints that were recognized in Psychology that were not being applied in either Linguistics or Artificial Intelligence.

It is not possible to go into any detail on these points, and indeed I have previously done so at book length (Powers and Turk, 1989), but rather I will seek to elucidate briefly some of the key insights that have been missed and regained. But first my focus is on the general premise that biological mechanisms of language, learning and thought are worth exploring computationally. There are two main reasons for this:

a. Biological, linguistic, neurological, psychological and sociological models all need to be computationally feasible and verifiable. That is a model that is provably impossible on computational grounds or whose predictions cannot be verified from a computer model, should be rejected, following a Popperian paradigm.

b. Computational and engineering goals that correspond to lowest common denominator human capabilities can usefully adopt as their starting point the known mechanisms, constraints and load characteristics of their biological counterparts. It is not necessary to labour the point by claiming that these are optimal in the sense that they are God-given or Evolution-optimized. Rather, when computers can't achieve anything like the performance of humans, achieving that level as a minimum is a reasonable first step before attempting to do better.

2.2 Poverty, Connectedness and Chunking

2.2.1 The Poverty of the Gold

One of the first results that heralded the age of Cognitive Science was Gold's (1967) proof that it was impossible to learn a superfinite language in the sense of 'identification in the limit' without negative feedback, and the corresponding evidence from Psycholinguistics that children neither received nor responded to negative feedback.

The fallacies in this argument are spelled out in Powers and Turk (1989), but in particular:

a. Gold (1967) himself proved that with 'anomalous text' where there were constraints on the ordering of the examples, overt negative information was required. Others very quickly showed similar positive results for different models, and in particular probabilistic grammars.

b. There is little evidence that language is superfinite in the sense of Gold's assumption – that would imply that clauses or phrases could exceed the length of our lifetime (conjunction and concatenation are not an issue). There is even less evidence that language must be at least context free as has often been claimed by Chomskian linguists – this would imply an infinite stack and a head to accommodate it.

c. There is evidence that children have available positive examples of all the constructs they learn and *can* take into account positive and negative evidence, providing that the input is not too far outside their level of competence. Our theory of anticipated correction is based on evidence that both adults and children autocorrect when what they say 'doesn't sound right'. Also the evidence is that children do not learn their parents' language(s), but develop their own unique idiolects and dialects.

d. The model does not take into account grounding.

2.2.2 The Locality of the Net

Another of the early results in Cognitive Science was Minsky and Papert's (1969) proof that simple biologically-motivated Perceptrons were not able to determine concepts like 'odd' or 'connected', but were able to handle abstract mathematical concepts like 'convexity'. This led to the suppression of neural nets for almost 15 years, on the basis that if a Perceptron couldn't deal with such simple concepts it wasn't much use.

The fallacies here include:

a. Far from being abstract and mathematical, recognizing convexity corresponds to the facility of human Perception to see that there is a hole in your shirt, a dent in your car, or a piece our of your apple. This is achieved by *low-*

level subconscious parallel processing at the level of Perception, the level the Perceptron was designed to reflect.

b. Connectedness requires tracing a path and parity determination requires counting. The Perceptron result should thus not be surprising as the child doing the puzzles in the Sunday paper *consciously* traces the path from the rabbit to the carrot in the maze and *consciously* counts the number of dots on the clowns shirt to work out which is odd man out, suggestive of high-level *sequential conscious* processing.

c. The fundamental locality constraint, which observes that each neuron has a relatively small fan in and fan out of connections to other neurons, reduces the problem to manageable complexity whereas the subsequent fully-connected Multi-Layer Perceptrons have scalability issues as well as an overt assumption of supervision.

2.2.3 The Magic of the Seven

Of a somewhat different character is the drawing together of a variety of research on diverse cognitive constraints by Miller (1956). There are some fallacies circulating regarding this, in particular a lack of recognition that the paper addresses a number of quite independent phenomena (limitations on discrimination, subitized counting, working memory) that seem to have quite independent underlying causes. The first evidence of the significance of this paper is that it represents a concern for the underlying cognitive limitations that underlies the fallacies in relation to the Poverty and Perceptron results. Beyond that, are essential to guide development of biologically-plausible language and learning models, as in the development of Human-Computer Interfaces that optimize the way information is presented based on an understanding of cognitive limitations (Powers and Pfitzner, 2003).

2.3 Self-Organization of the Pudding

Now we come to the key insights as to how a child's language and ontology can self-organize given that the child and his social and physical environment, constitute a closed system from the perspective of learning – there is no teacher or supervisor outside the system. The learning paradigm that is relevant is one of self-organized unsupervised multimodal learning, and whilst in a supervised paradigm, there is an automatic mechanism for evaluation, in an unsupervised paradigm there is none – patterns are simply discovered according to programmed biases.

2.3.1 Unsupervised versus Supervised

The Poverty and Perceptron results we discussed above are representative of two distinct learning paradigms according to whether or not there is a teacher available to say if examples are right or wrong – classically both paradigms assume there is a source of examples, with the unsupervised paradigm having only positive examples and the supervised paradigm having labelled positive and negative examples. Other supervised paradigms may label more specifically – e.g. with POS tags for Parts of Speech. However, for theoretical purposes this may be regarded as separate binary positive-negative labellings for each tag. Paradigms that label structures or rules as being correct or not, or belonging to a specific POS or not, are however somewhat different – the critical observation is that, like the POS labels, the structures are inventions that have no documented relation to neurological reality.

Gold's proof, in an unsupervised paradigm, led to the conclusion that supervision was necessary in the sense that a source of labelled examples was required – it is necessary to know which sentences are grammatical and which are not, otherwise it cannot be learned. Minsky and Papert's proof, in a supervised paradigm, was totally independent of the learning – they showed that it was impossible to represent the correct answer with Perceptrons, so it was impossible to learn it.

In fact, the distinction between supervised and unsupervised is not clear cut. The fact that sufficient supervision for Gold's Identification in the Limit could be achieved using ordering or statistical information illustrates this. Indeed any supervised learning system can be used in an unsupervised mode to learn auto-associations – that is one part of the input (from the same or a different modality) can be used to predict another, and often this will lead to useful categories or rules being developed. Conversely, labels can be provided to an unsupervised system as an additional input and treated in the usual way, and often this will change the bias of the system to prefer classes that correlate with the provided labels.

2.3.2 Evaluation in Application

The obvious problem with unsupervised learning is that there is no way of evaluating it, since the data sets for supervised learning are all based on a pre-existing theory. Given that self-organization and unsupervised learning are used to discover new patterns and invent new theories, they cannot be evaluated against human theories that are known to be incorrect or incomplete. All a supervised learning paradigm can achieve is to take the expert out of the loop and substitute the computer in the application of the expert's theory to new data. It is not capable of improving the theory, and depending on the learning algorithm used it may not even be capable of disproving the theory since in general supervised algorithms can be trained to give perfect results and deal with any desired balance between rules and special cases. The best we can achieve is comparing competing theories and competing learning algorithms.

A better way is based on the aphorisms, 'The proof of the pudding is in the eating', and 'The exception that proves the rule'. Puddings are for eating not for dissecting and analysing – similarly Language is for communicating not for dissecting and analysing. Our approach has thus been to compare unsupervised and supervised algorithms in real applications and thus gain objective performance measures rather than the subjective feeling that one rule or class is better than another. Furthermore, the original meaning of proof is exemplified in the idea of heating silver to allow removing the dross and *im*proving the purity – it does act as a test, but beyond that it is a mechanism for improvement.

In Natural Language Processing, it has become conventional to throw out the exceptions (e.g. 'water' as a verb) to improve the statistics, because the model is so bad and dependent on the lexicon at distinguishing noun from verb that allowing this case just opens the door to more errors (Entwisle and Powers, 1998). These observations have led to the development (Powers, 2003) of unbiased measures for evaluating both supervised and unsupervised algorithms for both binary and arbitrary labellings.

In summary, our approach is to use an unsupervised paradigm across multiple modalities, to evaluate results in applications of commercial relevance, and to use unbiased measures of informedness. The remaining sections of the paper summarize our scientific goals and the models we are currently developing, but here we illustrate this principle with some application-oriented evaluation we have performed to date.

2.3.3 Examples of Applications

a. Spelling Correction/Chinese Transcription – statistical information and automatically (unsupervised) derived categories are used to correct 10,000 commonly confused English words and then applied to choosing the correct characters for a Chinese PinYin transliteration (Powers, 1997b; Huang and Powers, 2001; 2003).

b. Machine Translation/Summarization – derivation of holo-/meronym and hyper-/hyponym relations and new algorithms for characterization of word similarity and comparison of unsupervised approaches with Wordnet both directly, and in application to Machine Translation and Summarization, focussing currently on nouns and verbs. In this case the performance of our algorithm using WordNet already significantly exceed human performance expectations (Yang and Powers, 2005), so it is not necessarily expected that unsupervised learning will do better – doing so would be evidence against the model being biologically accurate to the extent that the task is a reasonable one. The machine translation, document summarization, spelling correction and Chinese transcription tasks are natural tasks performed in context, whilst comparing lists of words or doing Wordpower (Reader's Digest) or TEFL Examinations are tasks performed out of context and contrived purely as tests – and it is known that human native speakers can achieve better test results with training.

c. Information Retrieval User Interface – this area is rich with application not only for word similarity, semantic and syntactic classifications, but also for investigation of the role of cognitive limitations and context of human and machine performance in common tasks such as websearch. Applications have included user and context modelling (can we use unsupervised learning to improve queries, rankings and summaries), analysis of the way people use keywords to describe documents as opposed to search for documents, discovery or relationships between dynamic pages and the underlying databases (YourAmigo.Com) and development of interactive multidimensional search interfaces (Powers and Pfitzner, 2003).

d. Speech Recognition/Control – this area also allows the use of syntactic and semantic models to improve selection of the correct word, but is also a focal point for multimodal fusion, using lip-reading to improve performance (Lewis and Powers, 2005). Specific applications include voice control of home equipment (I2Net.Com.au, Clipsal Homespeak), use of speech recognition in a sports stadium or a bank (commercial applications under development for noisy environments).

e. Brain-Computer Interface – this area tests unsupervised algorithms, for the separation of noise, artefact and signal components of EEG signals, in applications ranging from comparing the conscious and subliminal processing of language (Powers, Clark, Dixon and Weber, 1996) to monitoring the skill acquisition and stress levels of a soldier, testing the predictions of unsupervised models in relation to functional and content words, and extending our home control interface to allow multimodal AV+biometric control (commercial-in-confidence and military applications under development).

3 Biologically-Plausible Unsupervised Learning

3.1 Unimodal Models

Unsupervised learning and self-organization as biologically-plausible models have a history that extends back to Turing (1952) and von der Malsburg (1973), were generalized and popularized in the Kohonen Net, and are also broadly used in other guises, notably Independent Component Analysis (ICA). Notably, von der Malsburg (1973) demonstrated how the on-centre off-surround lateral distribution function provides sufficient constraint to model self-organization of the cortical hypercolumns sensitive to angles and Kohonen (1988) demonstrated the way the same kind of constraint produced an array of phonetically sensitive regions that resembled a typewriter, in which speech corresponds to trajectories across this surface, and even reduced speech produced recognizable trajectories. Powers (1983a, 1991) showed that the same approach could self-organize phrase- and clause-level structure from word-level input (1983) and character, phoneme or speech input (1991; Schifferdecker, 1994). Powers (1983b) also showed how the required lateral interaction function could be explained from first principles based on neuroanatomical considerations.

The connection between statistical models and biologically-motivated unsupervised learning or self-organization is very strong, but they are not equivalent classes – clearly not all statistical models have biological plausibility. Nonetheless it can be useful to consider the neural network models from a statistical perspective, and the lateral distribution function can be related (by a scale factor of normalization) to a probability distribution function (over distance), the area of network associated with a particular feature is monotonically related to the probability distribution function (across features), and the optimal sigmoid for ICA (Lee et al., 1999) is a linear function of the cumulative distribution function (of the sources). It should be noted that ‘sources’ refers to the underlying signals or causes that are perceived only in a mixed and convoluted way after modulation, transmission and perception within a medium/modality. The components detected by ICA correspond in many cases to features (e.g. edges or voicing), as well as to canonical sources (e.g. speaker location and identification).

3.2 Multimodal Models

Moving from a single modality to multiple modalities gives us not only the opportunity to learn more from a richer array of sources, but allows us to expand the range of paradigms available to us for learning. The obvious approach to multimodal unsupervised learning is not necessarily the best – whilst it is possible to simply throw everything at the learning system in an undifferentiated fashion, this does not correspond either to the differentiated structure of biological systems or the exigencies of achieving efficient learning and processing. Even in a supervised context, the undifferentiated approach tends to lead to lesser performance (efficiency and efficacy), and even catastrophic fusion – that is the results for multimodal learning are worse than can be achieved in one of the individual modalities alone.

3.2.1 Early Fusion

Early fusion of raw attributes across the modalities can usually be improved by a ‘horses for courses’ approach – the individual modalities are used to identify features they are good at identifying, and a late fusion of attributes and features is performed, possible making use of information or estimates about the noise, error and reliability characteristics of each mode in the current context (Lewis and Powers, 2005).

Although the terms early and late fusion are usually used in relation to a supervised learning paradigm, it is clear that the concepts can be adapted to unsupervised learning, and indeed there are a number of ways we can see this as a natural consequence of use of structured multimodal learning paradigm.

3.2.2 Multimodal Self-Supervision

The supervised-unsupervised dichotomy breaks down in a multimodal context as we can arrange to predict features or events in one modality based on input from another – that is one modality can be used to supervise another. Thus we have the full range of learning algorithms available to us, and we highlight again our earlier point that it is the paradigm that is supervised or unsupervised, and specific algorithms may be used in either mode notwithstanding their design for or close association with one paradigm. The extension of Kohonen Nets to Linear Vector Quantization, and the self-organization of the hidden layers in a Backpropagation network are well known examples.

3.2.3 Unsupervised Emic Fusion

Once unimodal unsupervised learning has been performed, either using an unsupervised paradigm or a multimodal self-supervision paradigm, feature information is automatically available for late fusion. In the case of the more powerful multimodal self-supervision paradigm, there will also be information about the noise, error and reliability of the unimodal features in terms derived directly from the predictability of attributes or features of the other modalities.

The raw attributes or inputs for each modality are intrinsically etic in nature – that is they are objective and the values are independent of the linguistic, behavioural or social characteristics or purposes of the perceiving individual or, in this case, system. However, once a learner has started to learn from examples from a particular linguistic, behavioural or social environment, the learned features reflect the probability distributions of that environment and hence aspects of the linguistic, behavioural or social characteristics and purposes that underlie and determine them. These features thus tend to be emic in nature and are increasingly subjective and dependent on the linguistic, behavioural or social context to which the learner has been exposed (Pike, 1954; Pike and Pike, 1977).

Adding to this model the possibility of recurrence leads to a Piagetian model of reflection and reflecting. The first level of features tend to be unimodal and are based on percepts alone, but successive levels of features increasingly build on mixtures of etic and emic attributes and features, facilitating the representation and learning of more complex concepts (Powers, 1997) as well as a model that closely reflects the blackboard models that are popular in Psychology, Speech Recognition and Artificial Intelligence (van der Velde and de Kamps, to appear; Powers, to appear).

4 A Biologically-Plausible Stereosemic Model

We now outline our low-level model. Whilst we and others, (e.g. Powers and Turk, 1989; Deane, 1992) have marshalled evidence concerning particular areas of the brain and their role in various aspects of language processing, our focus at present is to understand what kinds of interactions can explain stereosemy and we feel it is premature to devote much energy to hypotheses about higher levels or lower level of processing. We assume that the sensoria for each modality are projected across the brain hemispheres and that visual, auditory and vestibular projections of both eyes and ears are available as inputs to our stereosemic model, ignoring the senses of taste and smell.

The cortex is classically divided, on the basis of the gross neuroanatomy and characteristic densities of different classes of neurons, into six layers, I to VI, which alternate between white and grey, and in turn may be subdivided into finer sublayers distinguished by lower case letters. Generally, Layer IV is the primary input layer, and this accepts in particular sensory information relayed or echoed by the thalamus. Layer VI is the primary output layer that projects back to the thalamus, whilst layer V mainly projects to the striatum, brain stem and spinal cord. Layers II and III are hypothesized to be responsible for multimodal cortical association as they project to other areas of the cortex both in the same hemisphere and via the corpus callosum in the other hemisphere.

Whilst it is usual to think of feedforward and derived recurrent neural networks, this is at best an oversimplification and at worst nothing like what we see in the cortex. The classical biological model assumes that clusters of neurons, possibly of different classes, act together as a high level cell and make the apology that the models probably apply to such cells rather than neurons. To the extent that there is feedforward and recurrent activity, there would appear at least two such networks – one projecting inward and one outward from layer IV. In fact it looks more like projection from layer IV to III, II and I for processing and association, then reversed projection from those layers to layers V and VI for output. We should also allow for the possibility that (distinct and common) neurons may be part of separate virtual layered networks that are overlaid in the same cortical space.

4.1 Visual and Auditory Input/Output

Currently our focus is on the Visual and Auditory modalities, both the inputs and the outputs that control convergence and focus as well as compensating for overall intensity. We will assume simple RGB visual input from a pair of colour cameras as representative of the kind of input available from the eyes, and will encompass retina, ganglia and the relevant (neo)cortical layers and regions into our model. We will assume multiple microphones but at this stage will not seek to model the modulations introduced by the pinnae or bone conduction – rather we will use at least four microphones (tetrahedral array) and preferably more (parallelepiped array). Our robot baby was designed to include orientation and acceleration sensors as well as convergence, head orientation and limb locomotion motors (Powers, 2002), but at this stage we are confining ourselves to a simpler model with only cameras, microphones and wheeled locomotion, or even simulated world cameos.

In relation to the visual cortex, there is evidence that layers II/III are concerned with major disparity detection and feedback to control vergence (and thence focus), whilst layers V/VI are concerned with minor disparity detection indicated for stereopsis. This is our focus here, and we similarly will not be concerned with modelling edge detection and shade/texture filling but rather would look for evidence of self-organization in these respects. Similarly we have hypothesized that opponent colour relationships are self-organized and our work on face finding and lip tracking strongly suggests that the opponent colour system is essential to distinguishing mammalian foregrounds and features, from non-mammalian (face) and mammalian (feature) backgrounds. In particular, the tuning of the red cones to haemoglobin is remarkable, and the red-green opponents appear to be optimized for distinguishing mammal from vegetation, whilst the blue-yellow is useful for distinguishing both features within a face or animal as well as animals and vegetation against a water/sky blue background.

4.2 A Low-Level Laterally Recurrent Network

A number of factors are involved in stereopsis and stereophony, not least of which is the need to adjust convergence and focus to the appropriate distance, both attentionally and during saccade. From a multimodal perspective, there is also a need to reconcile the difference between the speed of light and the speed of sound – a served tennis ball has covered a quarter of its trajectory by the time we hear the sound, but we regard these as simultaneous, the window for simultaneity varying from about 3ms to 3 seconds depending on the modalities involved and the contextual feedback. Distance can be estimated visually, from the motor control of convergence and focus, from disparity, from vestibular information about head movement versus expectation of target size, from perceived versus remembered as well as interocular texture variation, and from interocular velocity differences determined in tracking the target. Aurally, we can use intensity and phase differences and visual-auditory delay information.

The model we are exploring assumes rapid bijective distribution of left and right visual and auditory fields to both hemispheres. There is some evidence to support a log polar representation of the opposite half of the visual field in V4 – the foveal area occupies the inner and the periphery the outer sides of the map, with a transition area between. This has advantages in terms of centricity and orientation invariant recognition, recalling that translation can be accommodated by saccade.

We propose that the recurrence between the cortex and the thalamus produces the characteristic synchronous labelling of the event even prior to Stereology being complete. We further propose a spreading activation that produces a match between labelled events from the same modality (stereopsis and stereophony) as well as across modalities (stereosemy) when firing patterns correlate at a specific location in the field of perceptual processing. The propagation time across the cortex is comparable with the interaural delay so that a sound can be automatically localized and the labelling interactions with the thalamus can trigger visual attention to the auditorally signalled event.

The requirement for vergence and focal adjustment based on coarse disparity information from Layers II/III naturally precedes the availability of the fine grain information from Layers V/VI, but it is not clear from where the visual motion/change triggered attention signalling occurs, but this may take place in the thalamus through comparison of recurrent visual information with incoming perception. The relative delays in the fusion of audio and visual information also contribute to automatic localization of the event in the depth dimension.

4.3 Work in Progress

The exploration of this model is experimental in nature, and we are seeking increased collaboration with neuroscientists to guide the precise parameterization and interpretation of the model. Our initial goals are relatively modest, being to explore different theories of multimodal binding, synchrony and stereopsis, whilst avoiding the complexity inherent in the total vision and audition problems. We are exploring variations on spatial and temporal, including spatiotemporal, encoding to study the role and nature of working memory and the binding of the individual percepts relating to an event. Our model is blackboard like, with the addition of a (bilateral or unilateral) spreading activation that provides very short term memory and the basis for stereology and stereosemy, with the coded recurrent pulse streams being directly initiated by the forwarding and recurrent echoing of information through the thalamus. We hope in this way to bridge the gap to our existing higher-level learning models.

5 References

Anderson, J. (1983): Spreading Activation Theory of Memory. *J. Verbal Learning and Verbal Behavior*, **22**:261-295.

Deane, Paul D. (1992): *Grammar in Mind and Brain: Explorations in Cognitive Syntax*, Berlin: Mouton de Gruyter.

Entwisle, J. and Powers, David M. W. (1998), The Present Use of Statistics in the Evaluation of NLP Parsers, *NeMLaP3/CoNLL98 Joint Conference*, ACL, 215-224.

Fingelkurts, Andrew A., Fingelkurts, Alexander A., Krause, Christina M., Möttönen, Riikka and Sams, Mikko (2003): Cortical Operational Synchrony during Audio-Visual Speech Integration, *Brain and Language* **85**#2:297-312

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

Huang, Jin Hu and Powers, David M. W. (2001): Large-scale Experiments on Correction of Confused Words, *Proc. Australian Computer Science Conference*, 77-82

Huang, Jin Hu and Powers, David (2003): Chinese Word Segmentation based on Contextual Entropy. *Pacific Asia Conference on Language, Information and Computation.*

Kohonen, T. (1988) The neural phonetic typewriter, Computer 21:11-22.

Lee, Te-Won, Girolami, Mark and Sejnowski, Terrence J. (1999): Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources, *Neural Computation*, 11, 417-41.

Lewis, Trent and Powers, David M. W. (2005) *Special Session on Audio-Video Signal Processing and its Applications*, *IEEE ISSPA*, Sydney

Lutzenberger, Werner, Preissl, Hubert, Pulvermueller, Friedemann, Pantev, Christo, Elbert, Thomas and Eulitz, Carsten (http://psycprints.ecs.soton.ac.uk/perl/search?year=19941994) Brain Rhythms, Cell Assemblies and Cognition: Evidence from the Processing of Words and Pseudowords, *Psycoloquy* 5#48

Miller, George A. (1956): The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 63:81-97.

Minsky, Marvin, and Papert, Seymour (1969): *Perceptrons*. MIT Press, Cambridge MA.

Pike, Kenneth (1954): *Language in Relation to a Unified Theory of the Structure of Human Behavior*, The Hague: Mouton

Pike, Kenneth and Pike, Evelyn (1977): *Grammatical Analysis*, University of Texas Austin and SIL.

Powers, David M. W. (1983a): Neurolinguistics and Psycholinguistics as a Basis for Computer Acquisition of Natural Language, *SIGART* **84**:29-34.

Powers, David M. W. (1983b): Lateral Interaction Behaviour Derived from Neural Packing Considerations, *DCS Report* 8317, University of NSW, Australia.

Powers, David M. W. and Turk, Christopher C. R. (1989): *Machine Learning of Natural Language*. New York/Berlin: Springer Verlag.

Powers, David M. W. (1991) How far can self-organization go? Results in unsupervised language learning. *Proc. .AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*, Stanford CA, 131-136.

Powers**,** David M. W. (1997a): Perceptual Foundations for Cognitive Linguistics, *International Conference on Cognitive Linguistics* (Poster Paper), Amsterdam. http://www.infoeng.flinders.edu.au/papers/19970001.pdf Accessed 28 September 2005.

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

Powers, David M. W. (1998): The Total Turing Test and the Loebner Prize. 279-280, *NeMLaP3/CoNLL98 Human Computer Conversation Workshop*, Sydney: ACL.

Powers, David M. W. (2002): Robot babies: what can they teach us about language acquisition? In J. Leather and J. Van Dam, eds The Ecology of Language Acquisition, Kluwer Academic 160-182.

Powers, David M. W. (2003): Recall and Precision versus the Bookmaker. *International Conference on Cognitive Science*, University of New South Wales, July 2003.

Powers, David M. W. (2005), On the unproductiveness of language and linguisticss, commentary on van der Velde, Frank, and de Kamps, Marc. To appear in *Behavioural and Brain Sciences.*

Powers**,** David M. W. and Darius Pfitzner (2003): The Magic Science of Visualization. *International Conference on Cognitive Science*, University of New South Wales, July.

Schifferdecker, Georg (1994): "Finding Structure in Language", Diplom Thesis, University of Karlsruhe FRG

Sipper, M., Sanchez, E., Mange, D., Tomassini, M., Pérez-Uribe, A. and Stauffer, A. (1997): A Phylogenetic, Ontogenetic, and Epigenetic View of Bio-Inspired Hardware Systems**.** *IEEE Transactions on Evolutionary Computation* , **1**#1:83-97.

Dan I. Slobin. (2004): From ontogenesis to phylogenesis: What can child language tell us about language evolution? To appear in *Biology and Knowledge Revisited*. J. Langer, S. T. Parker, & C. Milbrath (Eds.) Mahwah, NJ: Lawrence Erlbaum Associates.

Turing, Alan M. (1950): Computing machinery and intelligence. *Mind*, **59**: 433 – 460 (http://loebner.net/Prizef/TuringArticle.html).

Turing, Alan M. (1952): The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London*, 237.

van der Velde, Frank, and de Kamps, Marc (2005), Neural blackboard architectures of combinatorial structures in cognition. To appear in *Behavioural and Brain Sciences*

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

Weizenbaum Joseph (1966): ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* **9**#1: 36-35.

Weizenbaum Joseph (1976): *Computer Power and Human Reason*, San Francisco, CA: W. H. Freeman

Yang, Dongqiang and Powers, David (2005): Measuring Semantic Similarity in the Taxonomy of WordNet, *Proc. of the Australian Computer Science Conference*.