A Model of Language Acquisition: Foundations

Robert Worden

Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, United Kingdom

rpworden@me.com

DRAFT FOR COMMENT - June 2022

Abstract:

A working Bayesian computational model of language acquisition can be seen online at http://www.bayeslanguage.org/demo . The model has fast, robust learning of words – including their sound, meaning and syntax. Words and other constructions are learnt as composite feature structures. The working and performance of the model is described in a companion paper. This paper describes the cognitive, mathematical and computational foundations of the model.

The model uses tree-like feature structures as a computational implementation of the constructions of cognitive linguistics. The operation used by speakers to compose sentences from meanings is unification of feature structures. Unification is a Bayesian maximum-likelihood operation - a minimisation of Bayesian Free Energy. The operation used to learn words and other constructs is generalisation of feature structures. These operations support all language learning and use, from the child's first learning of words and canned phrases, up to productive adult language.

Keywords: Cognitive linguistics; Bayesian cognition; constructions; unification; generalisation; Bayesian theory of learning;.

1. Introduction

First language acquisition has been one of the outstanding unsolved problems of cognitive science. Children rapidly learn their native languages, learning thousands of words and complex syntax from unreliable inputs, without explicit instruction. There are very few working computational models which address the whole process of language acquisition, from first words up to productive adult capabilities.

This is the second of three linked papers describing a computational which does that, which can be seen at http://www.bayeslanguage.org/demo. The model learns a fragment of English from a starting point of no linguistic knowledge. The papers are:

1. 'A computational model of language learning': [Worden 2022a] describes the working and performance of the model, and compares it with other models of language learning.
2. 'A model of language acquisition: Foundations' [Worden 2022b; this paper] describes the mathematical and computational foundations of the model.
3. 'A theorem of language learning' [Worden 2022c]: derives a theorem which can be proved in this model of language learning, which has important consequences for the scope of the model, for language diversity and for language change.

In this model of language, the meanings of words, other constructions, and sentences are represented by composite feature structures – tree-like data structures with nodes representing parts of their meaning. Feature structures are a computational implementation of the constructions of cognitive linguistics [e.g Langacker 1987, 2008; Fillmore 1985; Goldberg 1995; Croft 2001; Kay 2002; Sag, Boas & Kay 2012; Bybee 2001; Kaplan & Bresnan 1981; Lakoff 1987; Slobin 1986; Talmy 2000; Hilpert 2014]. Feature structures and related structures such as scripts and frames have long been used to represent language meanings. This model borrows structures from several sources, such as Jackendoff [1983, 1991]. However, choosing the best semantic representation is not the main focus of the model; it can work equally well with different semantic representations.

The model has a semantic layer and a phonological layer, but has no intervening abstract syntax layer, as is used in generative grammars [Chomsky 1965, 1981, 1995].

The model uses **unification** of feature structures for comprehension and generation, as has been common in computational lingutistics. It uses a uniform procedure for language learning at all stages. This procedure is the **generalisation** of feature structures. Examples of language learning and use by means of these operations can be seen in the online demonstration.

The computational model is described at Marr's [1982] Level Two, of algorithms and data structures, without a neural implementation. It is not a connectionist neural net model.

This model is a Bayesian model of language. Each feature structure has an information content, which is a form of Bayesian Free Energy. Unification is a minimisation of free energy, as in Friston's [2006,2010] Free Energy Principle.

This paper describes the mathematical and computational properties of feature structures and the operations on them, relating them to optimal Bayesian cognition and learning. These are the computational foundations of the model described in these papers.

2. Cognitive Foundations of Language

This model is a cognitive linguistic model, in which language is closely related to other human cognitive faculties, and is not just an isolated or autonomous module in the brain. Before describing the model itself, it is worth describing the overall context in the brain, in which the model is situated – to give some feeling for the other cognitive faculties which language interacts with.

These other cognitive faculties have not been implemented as a computational model. But they could be implemented, and they will be described in those terms – at Marr's Level Two, as set of implementable algorithms and logical data structures. The details which follow may not be correct, but may be taken as illustrating a possible context for the model of language, even if they need to be modified in some places. The context for language is based on a simple architectural model of the primate brain, shown in figure 1:

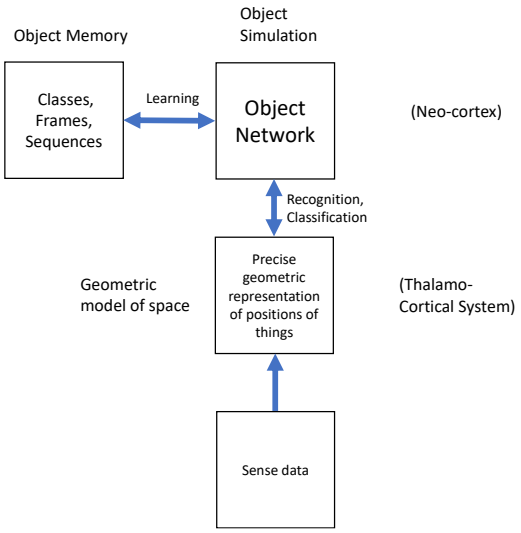

Figure 1: An outline architecture of the primate brain

Besides the incoming sense data (shown in the bottom box), there are three components to this model:

1. From incoming sense data, the brain constructs a fairly precise **geometric representation** of the locations the animal's body and the things around it.
2. There is also an **object representation** of current reality, which differs from geometric representation in ways described below.
3. The brain also has **memory** for the properties and dynamics of objects.

The main evidence for (1) - a precise geometric representation of local physical reality - is our conscious awareness, which consists largely of a faithful geometric representation of our bodies and our surroundings in the current moment, derived from sense data. Conscious awareness can only arise from some brain activity, involving fairly precise geometric representation of locations. In order to plan and execute complex physical movements, primates need to know the precise spatial dispositions of the things around them. The geometric representation supports the recognition of things and the planning and control of bodily movements.

Many aspects of this geometric representation are not yet known – not least of which is the unsolved problem of neural representation of the 3-D positions of many objects, with the high precision and high temporal resolution which primate brains achieve. It is not necessary to solve those problems here; but it is possible that the 3-D geometric representation of current reality is stored in the thalamo-cortical system.

The 'object network' representation of reality (2) overlaps strongly with the geometric representation (1), but has other capabilities to represent many kinds of object and their behaviour. In keeping with a description at Marr's Level two, the term **object** is used here in the sense of object-oriented programming (OOP) [Jacobsen et al 1992], which has been one of the most important developments in computing practice in recent years.

A computing 'object' may or may not represent a physical object. In OOP, the term 'object' denotes a composite package of information and behaviour, with defined interfaces to other objects. For instance, the primate brain must somehow represent a concept like 'insect'; this object is composite, in that it has linked parts (legs, head, wings, and so on), it holds information about each part, and it can simulate behaviours such as locomotion. Primate brains need to use these composite information objects to compute about insects or other things around them.

The object representation of reality (2) differs from the geometric representation (1) in the following ways:

- Many objects have a composite structure of wholes and parts.
- Objects may represent not only current physical reality, but other facts such as social reality (e.g. kinship links), or links to past and future events, such as recent history or potential for action
- Objects are linked to one another in a rich network of relationships, including physical relations and more abstract relations, many of which have a graded 'strength' such as a probability.
- Objects are classified into **classes** of objects that have been experienced before; the future behaviour of each object can be simulated using the learned behaviour of its class.
- Some classes of object inherit much of their information and behaviour from more general, superordinate classes, as in OOP.
- The object model represents reality as a symbolic network of nodes, each node having its own properties (here called 'slots'). Each composite object is itself a network of nodes, and the objects are linked in a larger network.
- While the symbolic object representation is more flexible than any geometric representation, it is more schematic and probably consumes fewer neural resources.

The key role of the object model of reality is to simulate the animal's body and surroundings in possible 'what if' scenarios for planning actions.

All these differences have obvious uses in the primate brain. They are the reasons to believe that the primate brain contains an object network representation of reality – because without it, primates could not do the complex things they do.

The object model relates to important concepts in cognitive linguistics, notably:

1. Langacker's [1987, 1991] detailed discussions of the semantics of language are descriptions of capabilities in the object layer, and how they interface to language.
2. The 'simulation semantics' of Embodied Cognitive Grammar (ECG) [Feldman 2006, Bergen & Chang 2013] recognises that the target of language understanding is not just a static symbolic structure, but an embodied simulation of reality.

Composite objects in the object network can be represented by composite feature structures – tree-like structures of nodes and slots, with cross-links between the trees, representing associations between the objects.

The object network can be envisaged as a network of computing objects (data and behaviour) on the surface of a pond. These objects have links not just to one another, but also to deeper layers of the pond. These layers are the more detailed neural representations in individual cognitive domains (such as smell or sound) many of which underlie conscious experience. The geometric model of space is one of these deeper layers.

Much of the power of the object network comes from the classifications of objects into classes with typical properties and expected behaviour – component (3) of the architecture. The creation of classes of object in component (3) depends on learning. Primates need to have a way of fast learning of composite objects – learning the class for a composite structure from experiencing only a small number of examples of the class. There is evidence for the fast learning of composite structures in primates.

Physically, components (2) the object network, and (3) the object memory are probably not distinct in the brain, but both reside in neo-cortex.

If this architecture is a high-level picture of our primate cognitive heritage, what extra ingredients are required for human language? I suggest that one extra capability is required. This is the ability of humans to recognise a 'common ground' between speakers [Tomasello 2003], as shown in figure 2:

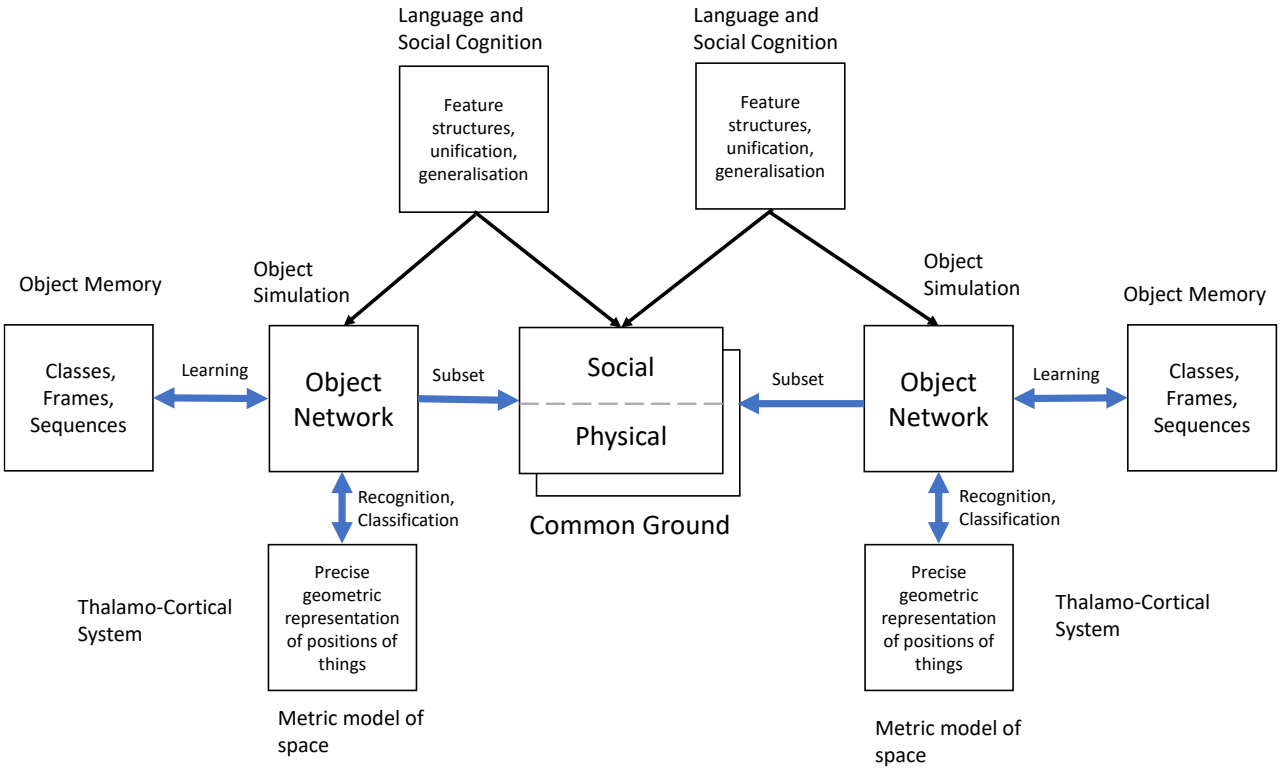

Figure 2: Architectural components supporting a conversation

The left and right sides of the diagram repeat the three previous components of primate cognition, for two people in conversation.

The new and uniquely human capability is the ability of each of the conversants to represent the **common ground** – (the central box) which is simply a subset of his or her own

object network, which is recognised as being shared with the other person.

The basic step for placing some object in the common ground is a pre-conscious inference such as: 'he can see that thing; so it is a part of our common ground.'

Because the common ground is a subset of each person's object network, it is itself an object network. It is a network of connected objects, each object being represented by a feature structure, and linked to other objects. The objects in the common ground can be roughly divided into objects representing physical things, and objects representing social reality.

The feature structures of language are based on the objects in the common ground, and more broadly in the object layer. Language consists of a set of operations on the common ground. Typical uses of language are to pick out some object in the common ground; to add to its description; to link it to other objects; or to add new objects.

The capabilities which are used to do these things are denoted in the top two boxes - feature structures, unification and generalisation.

Those capabilities exist already in the primate object network and memory. As will be described in section 6, primate brains need to store composite feature structures; to combine them by maximum-likelihood unification; and to learn them by generalisation.

3. Feature Structures, Unification and Generalisation

A **feature structure** is an information structure which has multiple nodes, connected in a graph (typically a tree or directed acyclic graph, or DAG) by links, or edges.

- Each node may hold several pieces of information (here called slots), of different modalities, depending on the domain. A typical slot is 'colour = @red', or 'size = @large', etc. Names for slot values are preceded by '@'.
- The edges carry information, which may be domain-specific – for instance 'time delay = 5 seconds ±3 seconds', or 'displacement = 5 cm left' or 'relationship = sibling' or 'node A is parent of node B'

The information held in nodes and edges has levels of uncertainty. Depending on the information in nodes and edges, and its uncertainty, each feature structure has an information content, denoted as B bits.

Feature structures have a model-theoretic semantics [Kay 2002], in that each feature structure represents a set of possible situations in the world. If a feature structure has information content B, then the set of situations it represents has probability approximately 2^{-B}. This underpins the relationship between feature structures and Bayesian models of cognition. In terms of the Free Energy Principle, [Friston 2006, 2010] the free energy of a feature structure is its information content B.

A primary relation between feature structures is the relation of **subsumption**:

A feature structure A *subsumes* another feature structure B, (written as A > B) if and only if all the information that is contained in A is also contained in. B.

Structurally, all the nodes, slots, and edges in A must also be in B; and B may have extra nodes, slots and edges. Any information in A must also be in B, so that B has equal or higher information content than A.

In terms of the model-theoretic semantics of feature structures, any situation in the world which is described by B is also described by A; but not necessarily the reverse. The set of situations described by B is a subset of the set described by A, and has smaller probability than the set described by A.

Subsumption can be used to define the main operations on feature structures, of unification and generalisation.

The **unification** C of two feature structures A and B (written as C = A **U** B) is the feature structure with smallest possible information content which satisfies both A > C and B > C.

The information content of C satisfies:

1. $I(C) \geq I(A)$
2. $I(C) \geq I(B)$
3. $I(C) < I(A) + I(B)$

There is an algorithm to compute C from A and B. This is to match pairs of nodes from A and B, trying to get the maximum match of information on the paired nodes, while respecting the constraints of the edges. The result retains all the shared nodes, and all the unmatched nodes which come from either A or B, allowing no contradictions. Hence the result (if it exists) contains all the information in A, and all the information in B. The best match of nodes maximises the amount of information in C which is not duplicated, coming from both A and B. So it minimises the information content of the result, and so minimises the free energy.

Unification may involve both discrete optimisation (choosing matching nodes) and optimisation of continuous varables (such as distances). For language, unification is mainly a discrete optimisation, of finding the best pairing between nodes of two structures with discrete slot values.

The model-theoretic interpretation of unification is as follows: Any situation described by C is also described by A, and is described by B. The set of situations described by C has the highest probability (lowest information content) of situations described by both A and B.

Thus C describes the maximum likelihood set of situations consistent with both A and B. Unification is an operation of maximum likelihood inference. Put another way, unification is an operation of scene construction, which constructs the most likely scene, or most likely model of the world, given A and B. Unification is the core operation of Bayesian scene construction [Mirza et al 2016], in any domain.

We can use this to describe how animals apply knowledge, once it has been learned. Suppose an animal has learned some cause-effect regularity, of the form (Cause => Effect). This can be paraphrased as 'if the current situation matches the cause, then the effect is likely to follow'. Both the cause and the effect can be expressed as feature structures, and can be combined as a single rule feature structure R – with a left-hand 'cause' branch and a right hand 'effect' branch.

Then the rule is applied by trying to unify the rule R with the current situation S. If the 'cause' branch of R matches S, then the effect branch of R predicts what will happen – and because of the Bayesian model semantics of feature structures, it is a maximum likelihood prediction of what will happen – which will give the animal greatest fitness. A rule R will not unify with the current sitation S if there is any contradiction between them. Then the rule will not apply.

Unifying the current situation S with learned rules R enables brains to construct a maximum-likelihood (and thus, maximum fitness), model of the world.

Subsumption is also used to define the operation of generalization, which plays a central role in learning new rules R.

The **generalisation** D of two feature structures A and B (written as $D = A \cap B$) is the feature structure with largest possible information content which satisfies both D >A and D > B.

The information content of D satisfies:

1. $I(D) \leq I(A)$
2. $I(D) \leq I(B)$

There is an algorithm to compute D from A and B, which is similar to that used to compute the unification C. To compute $D = A \cap B$, you again match pairs of nodes from A and B, trying to get the maximum match of information on paired nodes, while respecting the constraints of the nodes and edges. For generalisation, you retain only the matched nodes, slots and edges – and throw away any parts of A and B which are not matched. The best match of nodes maximises the amount of information in D.

There is a complementary relationship between the unification $C = A \cup B$ and the generalisation $D = A \cap B$. Because of this complementarity, there is an approximate relation between information contents:

$$I(C) + I(D) = I(A) + I(B)$$

The operations of unification and generalisation fit together in an algebraic structure, or **feature structure algebra**. Typical relations of this algebra are:

$$A \cap B = B \cap A$$

$$A \cup (A \cap B) = A$$

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Some of these relations are not exact, but can be used in practice. They underpin the self-consistency of feature structure operations. The relations mirror the relations of set theory, because each feature structures denotes a set of situations in the world.

A fundamental aspect of the feature structures used for language is the presence of **subsumption links** [Worden 1997] which are the key to the productivity of language. Learning and using subsumption links is what sets human language apart from most animal learning and conditioning.

Suppose that a feature structure A has nodes a_1, a_2, ... There is another feature structure B with nodes b_1, b_2 …, and A subsumes B; this is written as $A > B$.

The nodes of A and the nodes of B can be put in correspondence with one another: $(a_1 => b_1)$, $(a_2 => b_2)$ … $(a_n => b_n)$, for every node a_n, and the subsumption relation must apply separately to each node pair and their descendant subtrees: $a_n > b_n$. This implies that if a_n is a leaf node of A, b_n can have an arbitrary extra subtree of descendant nodes b_x, and still the relation $A > B$ will hold.

That describes the case where A has no subsumption links in it – the case we have been discussing so far. Now suppose that A has a subsumption link between two of its nodes. A subsumption link has a start node (say, a_n) and an end node (say, a_p). In the online model , these links are written as curved grey lines; here we write a subsumption link as $a_n \sim a_p$.

The link can only be only valid if a_n subsumes a_p; i.e. if $a_n > a_p$. So a_p must have the same or more information as a_n; Furthermore, the overall subsumption relation $A > B$ can only hold if, for the nodes b_n and b_p which correspond with a_n and a_p, $b_n > b_p$. If a_n and a_p are leaf nodes and have a subsumption link, the corresponding subtrees b_n and b_p cannot each be extended arbitrarily, independent of one another. Any extensions must obey the relation $b_n > b_p$.

That defines the effect of subsumption links like $a_n \sim a_p$ on the subsumption test $A > B$. As described above, the operations of unification and generalisation are defined in terms of the subsumption test (if $C = A \cup B$, then $A > C$ and $B > C$; and if $D = A \cap B$, then $D > A$ and $D > B$). The same definitions hold in the presence of subsumption links. The impact of subsumption links on these two operations is then:

- In forming the unification C = A U B, if there is a subsumption link $a_n \sim a_p$, this results in the sharing of information between the result nodes $c_n = a_n \cup b_n$ and $c_p = a_p \cup b_p$.
- In forming the generalisation D = A ∩ B, if the relations $a_n > a_p$ and $b_n > b_p$ both hold, then the result has a subsumption link $d_n \sim d_p$.

Generalisation discovers subsumption links, even when they are not in the inputs; and unification uses subsumption links to 'pipe' information between the linked nodes, in both directions. These properties are built into the algorithms used in the online model.

Learning the feature structures for language by generalisation discovers the subsumption links which make language productive.

Generalisation has some similarities to the operations of the Structure Mapping Engine (SME) [Falkenhainer Forbus & Gentner 1989; Forbus, Ferguson, Lovett & Gentner 2017] – although the main focus of SME is on analogical reasoning [Gentner 2003]. Operations similar to generalisation have been built into some models of language learning, such as the 'model merge' operation of [Bailey et al 1997], and the model of [Beekhuizen et al 2014].

4. Optimal Bayesian Learning Theory

The central result of Bayesian learning theory is simple and can be stated as follows: as soon as the evidence for some regularity in the environment is statistically significant (and no sooner) that regularity can be learned [Anderson 1990; Worden 1995,1997].

This implies that the number of training examples needed to learn some regularity can be very small; so that if animal brains are nearly Bayesian, they can learn regularities from small numbers of training examples. This result, which agrees with extensive data on associative conditioning and other forms of learning, differs from the much slower rates of learning of neural net models [Rumelhart & McLelland 1986; LeCun et al 2015], which typically require many thousands of training examples.

The speed of Bayesian learning can be illustrated by the example of a biased coin. Suppose a coin is biased, and gives heads on 80% of tosses. How many tosses will it take to learn that the coin is biased, if most coins are unbiased? When the coin has been tossed 20 times, it will have given heads approximately 16/20 times. Then, there is only a 1% chance (posterior probability) that it is unbiased; so, there is statistically significant evidence that the coin is biased. The bias can be learned from a small number of tosses.

To apply this to animal learning or language learning, we take the case of two events (called s and m) which may or may not be correlated with one another in the environment.

In a series of trials, each trial has one of four outcomes with probabilities:

$$a = P(s \,\&\, m)$$

$$b = P(s \,\&\, \text{not } m)$$

$$c = P(\text{not } s \,\&\, m)$$

$$d = P(\text{not } s \,\&\, \text{not } m).$$

where (a+b+c+d) = 1.

The frequencies of these outcomes will reveal any correlation, positive or negative, between s and m.

As the number N of trials grows larger, the probability P of any specific sequence such as acdbca.. grows exponentially smaller with N. The graph of ln(P) against N approximates to a descending straight line, with negative slope:

$$d\ln P/dN = a\ln(a) + b\ln(b) + c\ln(c) + d\ln(d)$$

We can re-express a..d in terms of conditional probabilities:

$$w = P(s)$$

$$x = P(m \mid s)$$

$$y = P(m \mid \text{not } s)$$

If there is no correlation between s and m, then x = y. In that case we can show:

$$d\ln P/dN = (a+b)\ln(a+b) + (a+c)\ln(a+c) + (c+b)\ln(c+b) + + (c+d)\ln(c+d)$$

This is the negative slope of lnP which is expected if there is no correlation. If there is any correlation between s and m (positive or negative), the negative slope of the correlated line is less than the uncorrelated slope, as shown in the figure:

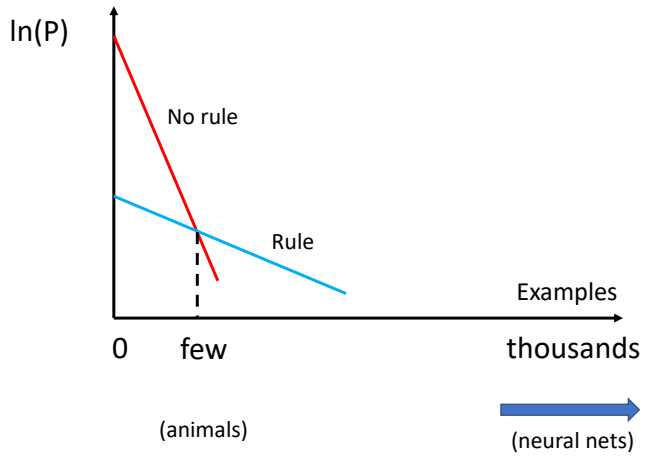

Figure 3: If there is a correlation between two events s and m, the evidence for that correlation becomes statistically significant after a small number of learning examples. The prior probability of any rule is small, so the blue line starts lower than the red 'no rule' line'. But

the blue correlated line gives a better account of the learning examples, so it has a smaller slope than the red line. When the blue line overtakes the red line, evidence for the rule is statistically significant. Slopes of the lines are given by elementary probability theory. The vertical axis is a log probability, or negative free energy

This shows the logarithm of the probability (negative free energy) of a sequence of examples, as the sequence extends. When there have been no examples, the prior probability of any correlation between s and m is small; thus the 'rule' line starts below the 'no rule' line.

After the two lines have intersected, the evidence for the rule is statistically significant- enough to believe the rule, in spite of its prior probability being lower than 'no rule'. At this point, a regularity can be learnt. The rule is the most likely (minimum free energy) account of the sequence of events.

It is in principle not possible to learn a correlation between events s and m any faster than this – because before the two lines intersect, any apparent correlation between s and m might just be a statistical fluke.

If there is only a weak correlation between the two events, the two lines have similar slopes, and it takes many examples before they intersect. More typically, the lines intersect rapidly, after a small number of examples.

This theory applies to the learning of word feature structures is as follows: The event **s** is the occurrence of the sound of a word in a learning example, and the event **m** is the occurrence of its meaning in the inferred meaning of the example. The word combines the sound and the meaning in a single feature structure, which has information content I_s bits in its sound part, and I_m bits in its meaning part.

The total information content of the feature structure is $I = (I_s+I_m)$, so the prior probability that it is a part of the language is approximately 2^{-I}. In the graph above of ln(P) against N, at the zero intercept when N = 0, the 'rule' line is lower than the 'no rule' line by approximately I bits.

However, the word sound **s** and its meaning **m** are strongly correlated in the learning signal, so the the two lines have different slopes. After a small number of learning examples which include the word, the two lines intersect, and the word can be learnt.

The application of feature structure generalisation to word learning is described fully in [Worden 2022c], and a brief summary is given here.

Suppose that an utterance is produced by unifying a number of word feature structures W, X, Y…, in some order of unification; and that utterance is used as a learning example L. Then L and its meaning part both obey

$$L = (W \sqcup (X \sqcup (Y \sqcup ..))))$$

This implies that the meaning of each word W is part of the meaning of L; so W subsumes L:

$$W > L.$$

Suppose that the same word W is used in a number of learning examples L_i , for i = 1,2,… This implies that

$$W > L_i \text{ for } i = 1,2,...$$

Now make the generalisation G of the learning examples:

$$G = ((L_1 \cap L_2) \cap L_3) \cap ..$$

The properties of generalisation and subsumption then imply that

$$W > G$$

This means that G contains all the information in W, and may contain other information (in defined slot values). The extra information in G arises from coincidental similarities between the learning examples L_i, and as more examples are added, this extra information soon falls away. So G is a good approximation to the feature structure of the word W.

This analysis can be extended to the learning of general rules (such as: 'all verb past tenses end in -ed) and special exception rules (such as: 'goed' is not a word). Then the diagram of figure 11 above would have three intersecting lines, for 'no rule' 'general rule' and 'special rule'. Learning the special rule depends on the **token frequency** of sentences containing 'go'; whereas learning a general rule depends on the **type frequency** of verbs like 'go' and others. The course of learning depends on the relationship between these frequencies, in complex ways.

Bayesian learning theory predicts that words can be learnt by generalising learning examples together; and that if this is done in an optimal Bayesian learning framework, any word can be learnt from a fairly small number of examples. However, the model of these papers does not yet implement an optimal Bayesian learner, and doing so might be difficult.

The model uses an approximation to optimal learning, called **permutation learning**. Suppose there is a set of learning examples L_i which all use some word W. The word is learnt by generalising those examples in some order:

$$G = ((L_1 \cap L_2) \cap L_3) \cap ..$$

On some occasions, the generalisation does not produce a result with significant meaning, because the learning example might be a distractor meaning, rather than the speaker's true meaning – the word sound occurs without the word meaning. These examples must be rejected from the generalisation, by some criterion of the resulting information content. Whatever the criterion, it is a tradeoff, and it may sometimes reject good learning examples. Hence the result of generalisation depends on the order of generalisation.

Serial generalisation, in the order in which the examples are encountered, is one possibility; but it does not always work well, as a word may sometimes get off to a bad start, and then not recover to the correct meaning.

Permutation learning works by making random permutations of the learning examples, and then forming candidate words by generalising the examples in the order of each permutation. The candidate words are put into groups with the same word meaning. For instance, for 40 random permutations of 20 learning examples, there are usually only 1 or 2 large groups of 10 or more candidate words with the same meaning. It is then possible simply to choose the largest group (the meaning which occurs most frequently), or to apply other criteria of quality – such as how many learning examples are matched successfully.

So given a set of learning examples, permutation learning tries out a number of different ways of combining the examples, and chooses the best resulting word. This is an approximation to Bayesian optimal learning from those examples. In the language learning model, it works well.

5. Feature Structures in the Primate Brain

There are several contexts where primate cognition requires the use of composite data objects:

1. Perceiving a complex multi-part object – sensing only some parts of the object, and filling in other parts from known properties of its class.
2. Planning a novel sequence of movements, using known constraints on movement, and making choices dynamically as the sequence unfolds
3. Doing the sequence of actions for some common daily activity, such as arranging bedding, or finding food, or tool use.
4. Negotiating a social situation, involving issues such as kin, rank and dominance, over extended timescales.
5. Navigation in a region of known territory.

Each of these requires computation with composite data objects – which may be represented (at Marr's level 2) as multiple connected nodes, each node having defined property values. That is, it requires computing with feature structures.

Primates have been under sustained selection pressure for many of millions of years, to do these computations very well. It is a reasonable hypothesis, and is supported by observation, that they do it almost as well as it can be done.

In all these situations there is potentially a large degree of uncertainty. Optimal computation under uncertainty requires a Bayesian estimation of the posterior probabilities of different outcomes. When a situation is partly described by several feature structures, the single feature structure which best describes the situation is found by a Bayesian optimal computation, which (as in section 3) is the unification of the feature structures which give the partial descriptions. So the primate brain needs to compute with feature structures by unification.

Some of the feature structures involved are long-lived feature structures stored in the object memory. They are created by learning, and there has been sustained selection pressure to make that learning as fast as possible. We know from associative conditioning experiments that many kinds of learning happen approximately as fast is they can happen, as defined by the Bayesian theory of learning [Anderson 1990].

So it is reasonable to suppose that in primates, the learning of composite feature structures is done at the speeds defined by Bayesian learning theory. This requires the learnt feature structures to be made by generalisation of feature structures for learning examples, because the generalisation of a set of feature structures contains the largest amount of common information which they share.

So there are sound empirical reasons to suppose that primate brains:

1. Use composite feature structures for cognition in several domains
2. Compute with feature structures by unification
3. Learn feature structures as fast as they can be learned, using generalisation.

This implies that the main computational capabilities needed for this model of language and language learning evolved in the primate brain, before human language.

6. Feature Structures in Language

While these papers are about language learning, rather than language use, we need to describe the computations supporting language use – to know what the child needs to learn, and how it is then used. The description of language use which follows can be illustrated by examples in the online model.

The feature structure for a typical word 'eats' is shown below, in the form displayed by the online model:

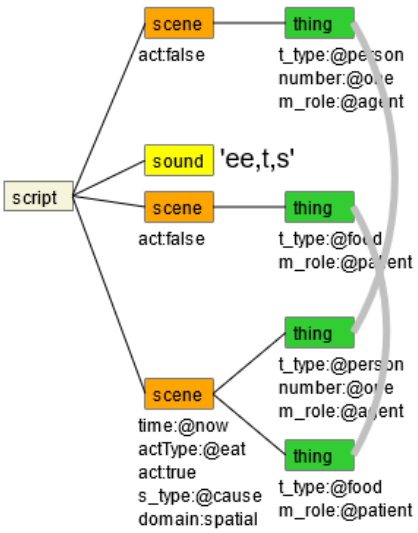

Figure 7: feature structure for the word 'eats' – which embodies its sound, its syntax and its semantics

The feature structure for a word embodies its sound, its syntax, and its meaning. The word sound is contained in the yellow 'sound' nodes (which in the figure, have been merged into one node for simplicity of display.)

The orange scene subtree at the bottom embodies the meaning of the word 'eats' – a scene in which an act of eating is happening in the present moment, with an agent (the eater) and a patient (the food). These are both shown as subordinate 'thing' nodes below the 'scene' node.

The syntactic constraints of the word 'eats' are shown in the two top orange 'scene' nodes, which identify the agent and the patient. When understanding a sentence, these scene nodes will have been made before 'eats' is unified, by unifying nouns and noun phrases. The slots define constraints on the agent and patient:

- The eater must be singular ('they eats' will not unify)
- The eater must be a person (maybe the language learner has never seen an animal eating)
- The eater must be able to be an agent ('him eats' will not unify)
- The patient must be of type 'food'

The curved grey lines in the diagram are subsumption links. These convey meanings (slot values and subordinate nodes) from the input 'agent' and 'patient' meaning scenes into the final 'act' meaning scene – which then contains all the information in the eating situation.

To understand the sentence 'Fred eats a banana', you first unify the word feature structures for 'Fred', 'banana' and 'a'. This creates the agent and patient meaning scenes, and implies that the word 'eats' cans now unify – because its input nodes (meanings and sounds) are all present, and their slots have values compatible with the feature structure for 'eats'. Unifying 'eats' produces the full final meaning scene – of a person called Fred eating a banana.

You can run through this example of language understanding, and other examples, step by step in the online model.

Language understanding starts with a feature structure consisting entirely of sound nodes, and successive unifications grow the structure by adding meaning nodes, until there is a meaning node giving the meaning of an entire sentence.

Language generation is the reverse of language understanding. Generation starts with a single orange meaning scene and subtree, containing the meaning to be expressed. Successive unifications add yellow sound scenes, until there are sounds of all the words to be said. In the final 'Hide' view of generation, only the sound nodes can be seen.

The same word feature structures are used for generation as for understanding. Unification with words can either add sounds to existing meanings (for language generation), or add meanings to existing sounds (for understanding).

In the online model, the sentence generation can be demonstrated step by step for any sentences or fragments of sentences in the 'Sentences' menu.

7. Prospects for Neural Implementation of Feature Structures and Language

This model of language and language learning has been defined at Marr's [1982] Level Two – the abstract level of algorithms and data structures which can be implemented on a digital computer. It has not been implemented at Marr's Level Three, of neural implementation.

We can ask about the potential for modelling a neural implementation of feature structures and their operations. I am not aware of any neural-level models of composite feature structures. We may list some of the things required of a neural implementation:

1. **Dynamic Representation of Feature Structures**: A feature structure is an open-ended multi-sensory data structure with connected nodes in a tree of unlimited depth, with an open-ended set of slots and values on any node, and some cross-links. It seems unlikely that the links between nodes in a feature structure can be represented by static neural connections; or that slots and their values could be represented by neural firing rates with fixed meanings.
2. **Dynamic Combination of Feature Structures**: It is often necessary to unify or generalise a feature structure with any other. So it seems unlikely that a feature structure could be neurally implemented at

any static location in the brain – without the ability to move it dynamically to where it can combine with other feature structures.

3. **Fast discrete node matching**: Unification and generalisation are both discrete optimisation processes, searching a discrete space of possible pairings between the nodes of two feature structures. This requires fast exploration of the different possible node pairings, to find the Bayesian maximum likelihood fit.
4. **Associative retrieval of feature structures**: When we hear a sentence, we effortlessly retrieve a few word feature structures from the many thousands of words that we have learned, using their sounds. To produce speech, we rapidly retrieve a few word feature structures based on their meanings. Both these require fast associative retrieval of feature structures.
5. **Fast Learning**: Any feature structure needs to be learned rapidly, from only a few examples of its use. In this respect, a neural net model of learning would be much too slow.

This is a very demanding set of requirements, and it is no surprise that no neural model of computation has yet come close to meeting them. Yet we know that as we have language, nature must be able to do it somehow. Most of these capabilities are required not only for language, but also for fast complex inference and learning, as seen in many primate species.

It appears that we need to abandon the traditional approach to the neural modelling of cognition, which is to say that a neural firing rate represents some fixed kind of information - some fixed type of information. This may work for sensory processing, where sense data has fixed type of information content – but it will not work for dynamic. multi-sensory feature structures. For those, it appears that neural firing rates must represent both data and metadata. Metadata is 'data about data' and can take a great variety of forms – such as 'what does a firing rate mean?' or 'what kind of processing is required?'.

There are very few models of neural information processing in which neural firing rates represent metadata as well as data. Unfortunately, even if such neural models can be devised theoretically, it may be very hard to test them experimentally. However, the rapid plasticity of cortical regions seems to imply that neural assemblies can be rapidly reconfigured to new tasks, possibly by changes of incoming metadata.

There are two lines of work which may be pointers towards a neural implementation of feature structures:

A. The Free Energy Principle [Friston 2006, 2010] has been the source of neural level implementations of Bayesian optimal processing, so may be a starting point. I am not aware of any applications of the FEP to complex tree-like feature structures; and the Bayesian optimisation done in FEP applications is usually a search in a space of continuous variables, rather than a discrete matching as is required for unification.
B. Work in Embodied Cognitive Grammar (ECG) [Feldman 2005; Bergen & Chang 2013] has explored neural implementations of models which, at Marr's Level Two, are similar to the feature structure model of this work. They have explored neural implementations of unification [] and learning operations similar to generalisation [Bailey et al 1997].

8. Evolution of the Capacity for Language

As described in section 6 of this paper, the computational capacities needed to support language – composite feature structures, with their operations of unification and generalisation - are needed for primate cognition, and so could have been present in the human brain before language. This is fortunate, because a speed limit for cognitive evolution [Worden 1995] implies that there can only have been a small amount of new genetic design information in our brains since we diverged from other great apes. There could not have been sufficient new brain design in that period to support fundamental new capabilities such as feature structures with unification and generalisation.

As described in section 2, the main innovation in the human brain may have been our tendency to cooperate, and our ability to recognise a 'common ground' of understanding with another human. These extra capacities are sufficient to support language learning and language use.

There remains a large question: why is the communication capability of language so remarkably powerful, compared with the capabilities of any other species? To list some aspects of its power:

- We have huge vocabularies, up to 50,000 words
- There is almost no limit to the range of topics we can talk about
- We can communicate and understand complex ideas within seconds
- We can pack unlimited meanings into a single sentence, using complex nested syntax
- We learn language in childhood, without effort or coaching
- We convey meanings quickly by many linguistic short-cuts

In these respects, language appears to be much more powerful and fast than would be needed to serve any purpose related to survival in our natural habitat. Language is over-engineered.

If our capability for language had evolved to help us in hunting, or gathering food, or finding shelter or caring for loved ones - all of which take place over timescales of minutes or more - then something much slower and less powerful than modern human language would have sufficed. We could speak in slow, simple sentences, using only a few hundred words, taking several minutes to express any meaning. We would not need all the short-cuts, speed, and expressive power of language. The benefits in fitness of these extra features are marginal.

Feldman [2006] has written (echoing a consensus) that: *'Everyone agrees that expressive language conveys very significant evolutionary advantages for groups that can use it'*. On the contrary, there is evidence that over long periods, language has not brought any significant evolutionary advantage.

Mankind has had the cognitive and physical traits needed for language for at least several hundred thousand years (perhaps for as long as a million years ago, when cooking was invented, enabling us to digest enough to support larger brains). For the great majority of that time, *homo sapiens* has been a marginal species in Africa – even passing through a single female 'genetic Eve' about 300,000 years ago; and possibly also passing through a geographically restricted 'aquatic phase'. So mankind nearly did not make it; language has not enabled us to dominate the planet. Our domination has only happened in the last 10,000 years, since the dawn of agriculture and complex societies.

On the basis of habitat-related selection pressures, one would not expect the human capacity for communication so far to exceed that of chimps, who have lived in similar habitats as mankind for millions of years. This presents a puzzle. Because language seems to be so massively over-engineered for any survival-related purpose, natural selection does not account for language.

Whenever a species has a trait which is unique to that species, and which is exaggerated beyond any natural need, it may be a sign that the trait evolved by sexual selection. In fact any species-specific trait is most likely to be the result of sexual selection, because sexual selection is the only selection pressure which differs markedly between closely related species in the same habitat. Think of birds' plumage, or flowering plants.

Humans differ from other primate species not only in language, but also in our larger brains and greater general intelligence. This suggests that the whole package of greater intelligence and language may have evolved together, by sexual selection.

The theory of sexual selection [Lande 1981; Maynard Smith 1982] shows how sexual selection leads to traits which:

1. Evolve rapidly, by runaway positive feedback between the sexes
2. Are arbitrary and species-specific
3. Are exaggerated beyond any habitat-related need, to a point where they decrease overall fitness
4. Are a real handicap to the individual (acting as an 'honest signal' of fitness to potential mates)

Miller [2000] has given extensive evidence that human intelligence evolved by sexual selection. Our large brains show all the hallmarks of sexual selection - such as rapid evolution, species-specificity, and exaggeration beyond the point of greatest fitness, to the point of being a handicap to individuals – for instance, in the huge energy requirements of the brain, increased size of the birth canal, and extended period of parental care.

Sexual selection cannot occur without display of the sexually selected traits; in order to drive selection, those traits must be visible to potential mates. So if there is sexual selection for greater general intelligence, there must be ways to display intelligence to potential partners.

If we could only display intelligence by making a better spear, or by cooking a tastier rabbit, or by carving a better statue, that would be a slow and inefficient form of display. Language is something we can rapidly display at any time of the day, addressing any topic. As such, it is the most efficient available display of intelligence. So sexual selection for intelligence includes sexual selection for language.

The best way to display your intelligence to a potential partner is to tell them something they do not know, relating it to what they already know. For you to know what they know, they must be able to tell it to you in a cooperative dialogue - and you must be able to infer what they know from what they say.

So to act as the best display for intelligence, language must involve:

- Mind-reading, to know what the other person knows and does not know
- Cooperative dialogues, to exchange knowledge and build up a sustained context for the mutual display of intelligence
- Complex syntax and large vocabulary, to express complex ideas quickly
- Language abilities which are symmetric between the sexes, to facilitate dialogues with potential mates.

In this picture, therefore, language is a kind of courtship dance for the mutual display of intelligence. Fluency in conversation is a sexually selected trait.

Sexual selection accounts for facts that other accounts of the evolution of language do not account for:

- That language evolved so rapidly, in less than two million years
- That it allows us to communicate so fast, conveying complex messages within seconds - much faster than we need to for survival purposes
- That it has such elaborate syntax and lexicon - more complex than is needed to convey most meanings
- That it constitutes an evolutionary discontinuity between mankind and the great apes
- For most of the period during which we have had language, it has not made us fitter – mankind has only just survived to the present day.

These factors are consistent with fast complex language as a competitive sexual asset, attractive to both sexes, and used for the display of intelligence - as it appears to be today.

9. Discussion

This is the second of three related papers, describing a working computational model of language acquisition. This paper has described the cognitive and mathematical foundations of the model.

It first described the broad foundations, in the cognitive capacities which we share with other great apes, which are required for language. Key amongst these is an object model of current reality, used by primates to simulate and plan complex physical actions. This object model is the basis of the simulation semantics of embodied cognitive grammar [Feldman 2006].

The object model must be able to represent, compute with, and learn composite multi-node data structures. It must do so in the face of uncertainty, in a Bayesian optimal manner. This leads to a requirement for the objects to be composite feature structures, which are unified for optimal Bayesian inference, and are generalised for optimal Bayesian learning.

This lays the computational foundations of the model of language and language learning – in feature structures, unification and generalisation. Because of their Bayesian foundation, these operations fit together in a simple mathematical structure – an algebra – which underpins the model. This leads to a theorem of language learning, which implies that the model can learn any construction, in any language. That theorem is the subject of the third paper in the series [Worden 2022c].

Finally, this paper discussed the question of why language has evolved only in mankind, but not in other great apes, which have similar cognitive capabilities – and why language is so remarkably powerful and over-engineered. A possible answer is that language evolved through sexual selection, as a means to display greater intelligence to potential mates. This answer accounts for many features of language, which other theories of language evolution do not explain.

References

Anderson, J.R (1990) The Adaptive character of thought, Lawrence Erlbaum Associates

Bailey D, Feldman J, Narayanan s, and Lakoff G (1997) Modelling Embodied Lexical Development, ICSI preprint, Berkeley, Calif.

Beekhuizen B. Fazly A. Stevenson S. Bod R. and Verhagen A. (2014) A Usage-Based Model of Early Grammatical Development, Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics, pages 46–54, Baltimore, Maryland USA, June 26 2014.

Bergen B and Chang N (2013) Embodied Construction Grammar, in the Oxford Handbook of Construction Grammar, , Hoffman T and Trousdale G (eds), Oxford

Bybee J. (2010) Language, usage and cognition. Cambridge University Press

Chomsky N. (1965) Aspects of the theory of syntax. MIT Press.

Chomsky N. (1981) Lectures on government and binding. Foris.

Chomsky N. (1995) The minimalist program. MIT Press.

Croft, W. (2001). Radical construction grammar: Syntactic theory in typological perspective. Oxford University Press..

Falkenhainer B. , Forbus K. and Gentner D. (1989) the Structure Mapping Engine: Algorithm and Examples, Artificial Intelligence 41, 1- 63

Feldman J (2006) From molecule to metaphor: a neural theory of language, MIT

Fillmore, C. (1985). Frames and the semantics of understanding. Quaderni di Semantica, 6, 222–254.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. Cognitive Science, 41. 1152-1201.

Friston K., Kilner, J. & Harrison, L. (2006) A free energy principle for the brain. J. Physiol. Paris 100, 70–87

Friston K. (2010) The free-energy principle: a unified brain theory? Nature Reviews Neuroscience

Goldberg, A. E. (1995). Constructions: A Construction Grammar approach to argument structure. Chicago/London: University of Chicago Press.

Hilpert M. (2014) Construction Grammar and its application to English, Edinburgh

Jackendoff, R. (1983), Semantics and Cognition, MIT Press, Cambridge Mass.

Jackendoff, R. (1991) Semantic Structures, MIT Press, Cambridge Mass.

Jacobsen I, Christerson M Jonsson P and Overgaard G (1992). Object Oriented Software Engineering. Addison-Wesley ACM Press.ISBN 978-0-201-54435-0.

Kaplan, R. M. and J. Bresnan (1981) Lexical Functional Grammar: a Formal System for Grammatical Representation

Kay P. (2002) An Informal Sketch of a Formal Architecture for Construction Grammar, Grammars 5: 1–19

Lande, R. (1981) Models of speciation by sexual selection on polygenic traits, Proc. Nat. Acad. Sci USA 78, 3721-5

Langacker, R. (1987). Foundations of Cognitive Grammar, Volume I. Stanford: Stanford University Press.

Langacker, R. (1991). Foundations of Cognitive Grammar, Volume II. Stanford: Stanford University Press.

Langacker, R. (2008) Cognitive grammar: A basic introduction. Oxford University Press.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521, 436–444. doi: 10.1038/nature14539

Marr, D. (1982) Vision, W.H.Freeman

Maynard-Smith, J (1982) Evolution and the Theory of Games, Cambridge University Press

Mirza M, Adams R, Mathys C and Friston K (2016) Scene Construction, Visual Foraging, and Active Inference, Front. Comput. Neurosci., 14 June 2016

Rumelhart, D. E. and McLelland, J. L. (1986) Parallel Distributed Processing, MIT Press

Sag I. A., Boas H. C, & Kay P. (2012) Introducing Sign-Based Construction Grammar, in Sign-Based Construction Grammar, Eds: Boas H. C. & Sag I. A.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, US: Harvard University Press.

Tomasello, M. (2008) The origins of human communication. MIT Press.

Worden R. P. (1995) An optimal yardstick for cognition, Psycoloquy, Vol 7. Expanded version on ResearchGate

Worden, R.P. (1996) Primate Social Intelligence, Cognitive Science 20, 579 - 616.

Worden, R. P. (1997) A Theory of language learning, https://arxiv.org/abs/2106.14612 .

Worden R. P. (2022a) A model of Language Learning, unpublished paper

Worden R. P. (2022b) A model of Language Acquisition: Foundations (this paper), unpublished paper

Worden R. P. (2022c) A theorem of Language Learning, unpublished paper