A Computational Model of Language Learning

Robert Worden

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

rpworden@me.com

DRAFT - June 2022

Abstract:

This paper describes a working computational model of language acquisition, which learns the word sound segmentation, syntax and semantics of productive language. The learning model is based on a Bayesian cognitive linguistic model of language. It can be seen running at http://www.bayeslanguage.org/demo/model , learning a vocabulary of about 40 words from a starting point of no linguistic knowledge, using 800 examples of generated learning data.

The model is robust, and is expected to scale well to larger vocabularies and complex syntax. It agrees with data on first language acquisition, including: (a) fast, robust learning of language; (b) an initial single-word learning phase (c) subsequent learning of productive constructions; and (d) 'verb island' learning of individual verbs before learning regular syntax.

The model is capable, for one reason: it has a simple mathematical basis, in a Bayesian theory of learning and inference. It can be proven that the model works.

Many computational models of language learning either do not learn word meanings, or assume that the learner starts with some linguistic knowledge; so they can only be compared with selected aspects of first language acquisition. Two models which are more nearly complete are the models of Beekhuizen et al. [2014, 2015] and Abend et al. [2019]. This model is compared with those models.

Keywords: Language acquisition; learning meanings; productivity; Bayesian cognition and learning; feature structures; unification and generalisation; cognitive linguistics

1. Introduction

First language acquisition has long been a key challenge for cognitive science, as emphasised by Chomsky [1965, 1980]. Even today there are very few working computational models of language acquisition which are near to complete, in meeting all the main requirements for language learning - learning word segmentation, semantics, and productive syntax, from a starting point of no linguistic knowledge.

The model of this paper is complete in that sense, and can be seen running online at http://www.bayeslanguage.org/demo/model. It learns the syntax, semantics and phonology of words from generated learning examples, starting with no linguistic knowledge, and learning rapidly as young children do. The words it learns support the productive generation and understanding of language.

This is the first of three linked papers describing the model:

1. 'A computational model of language learning': [Worden 2022a; this paper] 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] describes the cognitive and mathematical foundations of the model.
3. 'A theorem of language learning' [Worden 2022c]: derives a theorem in this model of language learning, which has important consequences for the scope of the model, for language diversity and for language change.

The model is fast, robust and reliable – not requiring large computing resources of fine tuning of parameters. It reproduces the main features of early language learning, including fast learning, a 'one word' stage of learning, the later learning of productive constructs, and the 'verb island' learning of the syntax of individual verbs before learning any regular syntax, observed by [Tomasello 2003, 2009].

The online demonstration shows the model learning about 40 English words from about 800 learning examples. Scaling the model to larger vocabularies and more complex syntax has not yet been tested, but there are theoretical reasons to expect that the model can learn any construction in any language [Worden 2022c], and will scale well with increasing vocabulary.

The model is based on the principles of cognitive linguistics [Langacker 1987; Fillmore 1982, 1995; Goldberg 1995; Croft 2001; Kay 2002; Sag, Boas & Kay 2012; Bybee 1985; Kaplan & Bresnan 1981; Lakoff 1987; Slobin 1986; Talmy 2000; Hilpert 2014] and Bayesian Cognition and Learning [Rao et al 2002; Friston, Kilner & Harrison 2006; Chater & Oaksford 2008; Friston 2010;]. In the model, words, sentences and constructions are all represented as tree-like feature structures.

I know of only two other computational models of language learning which are complete or near-complete in the same sense as this model. These are the models of Beekhuizen et al. [2014, 2015] and Abend et. al [2019]. This model is compared with those two models.

The model of language learning rests on two key operations on feature structures – the operations of **unification** and **generalisation**. These operations are mathematically defined through Bayesian optimal inference, and are complementary to each other [Worden 2022b].

This is a very capable model of language learning, for one reason: it has a simple mathematical basis, in the Bayesian theory of learning and inference. It can be shown mathematically [Worden 2022c] why the model works.

The learning model is available as a download from the demonstration site. This can be used to test the model, or to apply it to other languages.

2. Principles of the Model

The principles of the learning model are described in [Worden 2022b]. In summary:

1. **The model is defined at Marr's [1982] Level Two**: The structures and operations of the model are defined and implemented at Marr's Level 2 of data structures and algorithms. The neural implementation of those operations (At Marr's Level 3) is not defined. The model is not a connectionist neural net model.
2. **It is a Cognitive Linguistic Model**: The model follows the principles of Cognitive Linguistics, in which language is closely related to other cognitive faculties of the primate brain. The core data structures of the model are tree-like **feature structures**, which represent the constructions of cognitive linguistics. Feature structures are mappings between sounds and meanings. There is no intervening syntactic level, as there is in generative grammar [Chomsky 1980].
3. **It is a Bayesian Model of Cognition**: The two key operations on feature structures in the model are **unification** and **generalisation**. These are

mathematically defined Bayesian maximum likelihood operations, using symbolic matching of the nodes of feature structures. Unification is used for language production and understanding [Gazdar et al 1985; Kaplan & Bresnan 1981; Kay 2002]. Generalisation is the core operation for Bayesian optimal learning.

4. **Speakers and Listeners share a 'common ground' of understanding of the current situation and its context**: The common ground [Tomasello 2002]is not represented in the model, but it allows a learner to infer (not always reliably) what a speaker is referring to. If the common ground was to be represented explicitly, it would be an object-based simulation of the current physical and social situation and context, built using the principles of Object-Oriented Programming (OOP). Such models are used in Embodied Cognitive Grammar (ECG) [Bergen & Chang 20013]. All language use relates to the common ground.
5. **Learning examples are used to infer the feature structures for words**: It is assumed that adult speakers have a stock of word feature structures, which they use by unification to produce utterances. These are used by a learner as learning examples. On each learning example, a learner hears an utterance and infers its meaning, as a feature structure. By generalising learning examples which contain the sounds of some word, the learner learns the feature structure for each word (or other construction) – and can then use word feature structures to speak and understand, by unification.

3. Core Learning Mechanism

This section outlines the core learning mechanism used by the model. Further details are given in the Appendix, and in [Worden 2022b].

In the model, feature structures represent the sounds, meaning and syntax of any word or other construction; grammar is fully lexicalised. Utterances are produced or understood by unifying the feature structures for constructions – as has been done in computational linguistics for many years [Gazdar et al 1985; Kaplan & Bresnan 1981; Kay 2002].

The **unification** of two feature structures is defined as the smallest feature structure which contains them both as substructures.

The learning model is based on an operation that is complementary to unification – the operation of generalisation.

The **generalisation** of two feature structures is defined as the largest feature structure which they both contain as substructures [Worden 2022b].

When a child hears two utterances in which the same word is used, and correctly infers the speaker's meaning, each utterance and its meaning can be represented as a feature structure – a learning example. Each learning example contains (as a substructure) the sounds of any word it uses, and its meaning.

Then, the generalisation of these two learning examples will contain both the sounds of a shared word and its meaning[1]. It may contain a few other nodes and slot values, from random coincidences between the two examples – but it is a good first approximation to the feature structure for the word. When it is generalised with other learning examples containing the same word, any coincidental similarities which are not part of the word are rapidly removed.

So the core mechanism for learning a word (or other construction) is to generalise together a small set of learning examples in which the word is used. As will be seen by running the model, generalisation discovers the syntax of productive words, as well as their sounds and meanings.

Generalising the learning examples containing a word, in the order in which they are encountered, is a fairly reliable learning mechanism; but in a small proportion of cases, it can get off to a bad start, and does not recover. So serial learning is supplemented by another process, also using generalisation, which is closer to Bayesian optimal learning.

When a small number of learning examples have been encountered for any word[2], several different candidates for the word can be made – by permuting the learning examples in different random orders, and generalising them in those orders. Different orders give different starting words, and different final results. Choosing the best candidate (by criteria described in the appendix) gives robust and reliable learning of all word feature structures. This is called **permutation learning**. By permuting learning examples, and keeping the best learnt word result, it approximates Bayesian optimal learning from those examples.

Permutation learning can only be used if a child is able to retain several learning examples for a word, for long enough to permute them; so it is not fully compatible with the 'Now or Never' bottleneck discussed in [Christiansen & Chater 2016; Chater & Christiansen 2016]. More generally, it seems likely that the closer learning is to approach to Bayesian

[1] This holds as long as the word is not used in a nested context – as can be seen in the demonstration

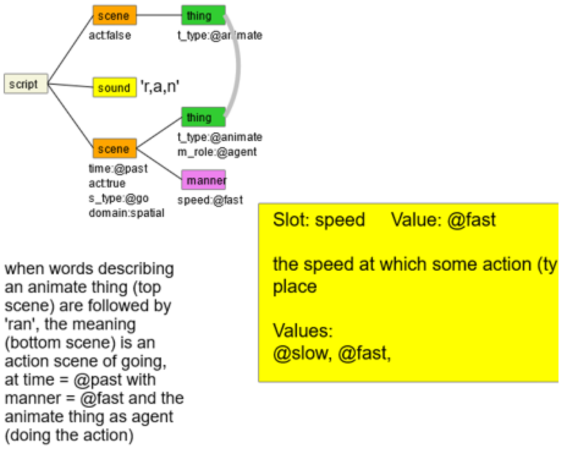
[2] In the model shown in the demonstration, this number is set at 20.

optimal learning, the more memory for learning examples is needed.

4. Running the Model

An on-line demonstration of the learning model can be seen at http://www.bayeslanguage.org/demo/model. I shall describe how the model runs by reference to that demonstration.

You can inspect the feature structure for any word by selecting it from the 'Words' menu in the demonstration. For instance:



Slot: speed Value: @fast

the speed at which some action (ty
place

Values:
@slow, @fast,

when words describing an animate thing (top scene) are followed by 'ran', the meaning (bottom scene) is an action scene of going, at time = @past with manner = @fast and the animate thing as agent (doing the action)

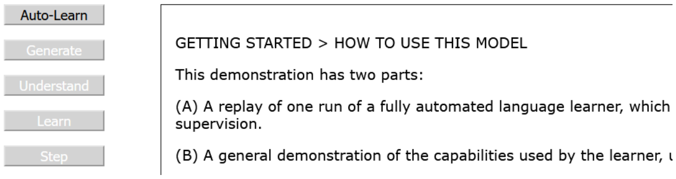
Figure 1: The feature structure for a typical word

The learning model learns these feature structures, starting with no knowledge of any words. In doing so, it learns the phonology, semantics and syntax of any word.

The word 'ran' is a productive word, with a productivity denoted as [1,0]. This means that it requires one phrase or word to come before the word sound - describing any animate thing that can run - and it requires no words or phrases after it. It can be used productively to describe any animate thing running.

The animate entity which precedes 'ran' is shown in the feature structure above – which is to be read in time order of its inputs, from top to bottom. The bottom 'scene' node is the meaning resulting from applying (unifying) the word. Adjectives have productivity [0,1], and simple transitive verbs have productivity [1,1]. Nouns have productivity [0,0].

A run of the learning model can be replayed by repeatedly pressing the 'Auto-Learn' button:



GETTING STARTED > HOW TO USE THIS MODEL

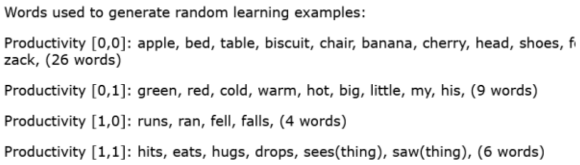
This demonstration has two parts:

(A) A replay of one run of a fully automated language learner, which
supervision.

(B) A general demonstration of the capabilities used by the learner, u

Figure 2: Starting the learning demonstration

When you first press the button, the model displays the words used to generate the learning examples which are the inputs to the learner:



Words used to generate random learning examples:

Productivity [0,0]: apple, bed, table, biscuit, chair, banana, cherry, head, shoes, f
zack, (26 words)

Productivity [0,1]: green, red, cold, warm, hot, big, little, my, his, (9 words)

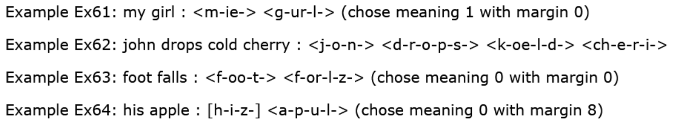
Productivity [1,0]: runs, ran, fell, falls, (4 words)

Productivity [1,1]: hits, eats, hugs, drops, sees(thing), saw(thing), (6 words)

The model has a pre-linguistic sound learning phase, described in the next section. In this phase the learner understands no meanings, and uses the sounds of learning examples to learn the phonetic boundaries of words and phrases, by a Bayesian-like statistical learning algorithm. This phase models the earliest months of an infant's life; so that when she is first able to infer the meanings of learning examples, she already knows the boundaries of spoken words. This phase has been implemented in the model, but is not shown running.

Each learning example is a sentence or phrase made by unifying feature structures for words selected at random from the generating set, using a context-free grammar (not known to the learner). Semantic restrictions of the words are applied by unification. These semantic restrictions remove some examples, leading to fairly sensible learning examples.

When you next press the 'Auto-Learn' button, the model creates a cycle of 100 random learning examples, which are shown in the lower screen.



Example Ex61: my girl : <m-ie-> <g-ur-l-> (chose meaning 1 with margin 0)

Example Ex62: john drops cold cherry : <j-o-n-> <d-r-o-p-s-> <k-oe-l-d-> <ch-e-r-i->

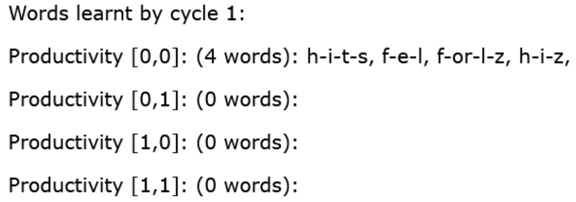
Example Ex63: foot falls : <f-oo-t-> <f-or-l-z-> (chose meaning 0 with margin 0)

Example Ex64: his apple : [h-i-z-] <a-p-u-l-> (chose meaning 0 with margin 8)

Because of the prior sound learning phase, the sounds of the learning examples can be segmented into words by the learner. The spelling of the words is phonetic, because the learner hears only phoneme-like units of sound.

Each learning example is described by a feature structure, which contains the sounds of the example, and a meaning which the learner infers from the context. Words are learnt by generalising the feature structures for learning examples which contain their sounds.

The model learns incrementally from the learning examples, stopping at the end of each cycle, and displaying the words that have been learnt. At the end of cycle 1 it shows:



Words learnt by cycle 1:

Productivity [0,0]: (4 words): h-i-t-s, f-e-l, f-or-l-z, h-i-z,

Productivity [0,1]: (0 words):

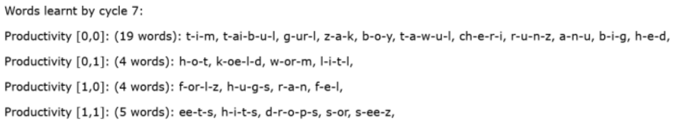
Productivity [1,0]: (0 words):

Productivity [1,1]: (0 words):

In this first cycle, when almost no words have been learnt, each word is learnt in an unproductive [0,0] form, because this learning requires no knowledge of any other words.

When some words are known, each learning example can be partly parsed by unifying it with feature structures for the known words. This replaces those word sounds by their meanings. By generalising these partly parsed examples, a productive word can be **promoted** to its more productive forms. Productive forms of words, which contain subsumption links, are learnt because generalisation of the parsed learning examples finds the subsumption links.

For instance, after 700 examples:



```
Words learnt by cycle 7:
Productivity [0,0]: (19 words): t-i-m, t-ai-b-u-l, g-ur-l, z-a-k, b-o-y, t-a-w-u-l, ch-e-r-i, r-u-n-z, a-n-u, b-i-g, h-e-d,
Productivity [0,1]: (4 words): h-o-t, k-oe-l-d, w-or-m, l-i-t-l,
Productivity [1,0]: (4 words): f-or-l-z, h-u-g-s, r-a-n, f-e-l,
Productivity [1,1]: (5 words): ee-t-s, h-i-t-s, d-r-o-p-s, s-or, s-ee-z,
```

After each learning cycle, you can use the model menus to show:

- What has been learned so far
- How it has been learned

Use the 'Learning' menu to display the feature structure for any word that has been learnt so far. Learnt words are grouped by their learned productivity:

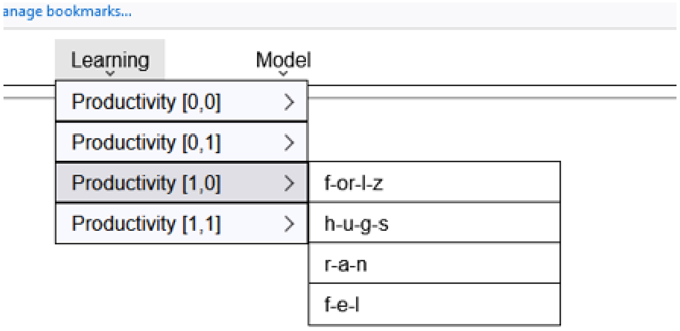


Figure 3: Menu to show the feature structures of words that have been learnt

Selecting 'r-a-n' from the Productivity [1,1] sub-menu will show the feature structure for the word, as it has been learnt. This is not the feature structure that was used to make the learning examples, but is the feature structure learnt from them. The two are usually identical, or very similar.

The representations of the meaning of words are simplified compared to the full meanings of words known by adults, which include many social and other associations. These associations could also be learned by the same learning mechanism. The choices of slots and slot values in the model are simplified, loosely following the work of Jackendoff [] and others. Had the choices of slots been different, the same mechanisms of unification and generalisation learning would still work.

You can see how generalisation learns any word by showing the word from the 'Learn' menu, pressing the 'Learn' button, then repeatedly pressing the 'Step' button. This will show feature structures for a few learning examples containing the word, and the results of generalising those feature structures together. You will see how generalisation projects out the sounds and meanings which are part of the word, and throws away other parts.

The learner starts with no knowledge of the structures, meanings or syntactic categories of words. The only initial knowledge of the learner is:

1. The segmentation of sounds into words, as learnt in the pre-linguistic phase
2. A set of node types, slots and slot values with which to represent the meanings of situations.
3. An ability to infer the meaning that a speaker expresses, from the context of 'common ground', with limited reliability.

5. Performance of the Model

There are two main measures of the performance of the learning model:

1. The speed with which it learns – how many learning examples it needs to learn any word or construction
2. The accuracy of its learning – how closely the learnt word feature structures resemble those used to make learning examples.

By each of these measures, the model performs well.

The number of words learnt of each productivity, as a function of the number of cycles of 100 learning examples, is shown below for a typical run of the program:

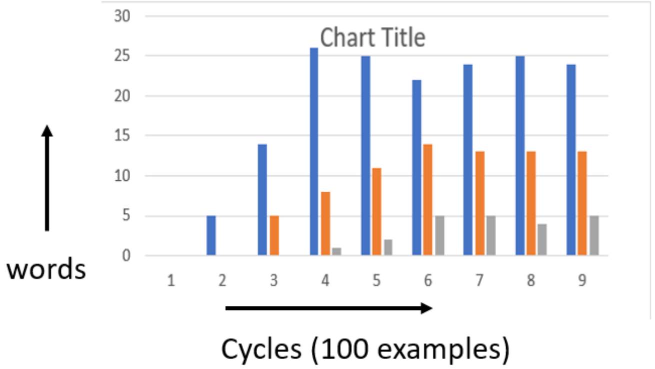


Figure 5 : number of words learnt of different productivities, as a function of the number of cycles of 100 learning examples.

In this run of 800 learning examples, the vocabulary used to make the examples was 45 words. Before the last cycle, 42 of these words had been learnt, with their correct productivities.

The blue bars represent words of productivity [0,0] (which are mainly nouns); the red bars are the sum of words with productivities [0,1] and [1,0] (adjectives and intransitive verbs); while the green bars are words of productivity [1,1] (transitive verbs). Each word is learnt initially in its unproductive [0,0] form, before being promoted to its full productivity.

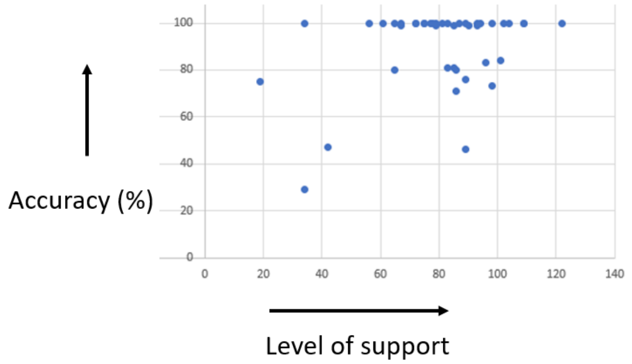
Learning 42 words from 800 learning examples is a fast rate or learning, comparable to the rates at which children learn their first words. This fast learning is consistent with the Bayesian theory of learning described in [Worden 2022b], but is not consistent with the much slower learning of neural nets.

The accuracy of learning word feature structures can be measured as follows: for any word, denote the 'correct' feature structure used to make learning examples by W_e; denote the learned word feature structure by W_l; and denote their generalisation by $W_g = (W_l \cap W_e)$. If learning is accurate, these three feature structures are equal; but any errors in learning will lead to gaps in W_g. Denote the information content of the feature structures by I_g, I_e and I_l respectively. Then define the percentage accuracy of a word to be

$$A = 100^* I_g / \max (I_e, I_l)$$

Any discrepancy between the two feature structures results in A being less than 100.

The next figure is a scatter plot of the accuracy as a function of the strength of belief in each word, for a typical run of the learner, after 800 examples:



There is a dot for each learned word; the horizontal axis is a measure of the strength of belief in the word; and the vertical axis is the measure A of its accuracy (measured against the 'oracle' of the words used to generate examples, which the learner does not know). If a learnt word either has extra information, or missing information, it scores less than 100%.

More than half the words are learnt with complete accuracy, and most of the others are learnt with accuracy around 80% (this may arise from a discrepancy of only one slot). Only three words have accuracy below 70%, and two of those words have only a low level of confidence.

This level of accuracy is more than sufficient for use of the word feature structures to be used to produce and understand language.

Other aspects of the performance and robustness of the model are:

1. **Parameter Insensitivity**: While there are variable parameters which affect the running of the model (some of which are described in the appendix), the model does not depend on fine tuning of the parameters. It runs well with a wide range of parameter values.
2. **Fast execution**: The speed of execution of the model cannot be seen in the online demonstration, which only replays a previous run of the model. Running the model to learn 40 words from 800 learning examples only takes a few seconds on a small PC. The model is efficient because it learns words individually, not needing to learn any wide cross-language regularities; and because generalisation, used to learn each word, requires only simple discrete pattern matching. This may imply that a neural implementation of the model, in the human brain, could also be fast and efficient.
3. **Learning any construction in any language:** To date, the model has only been tested on a small subset of English. However, there are reasons to suppose that generalisation learning will work for any construction in any language; these reasons are described in section 8 and in [Worden 2022c]. While the model will doubtless need refinement to learn languages very different from English, the prospects for successful learning of those languages by the model are good.

6. Pre-Linguistic Learning of Word Boundaries

If two learning examples both include the same word, generalisation of those learning examples makes a feature structure whose sounds include the sounds of the word, and does not include other sounds which the examples do not have in common. Generalisation works to project out the sounds of words – to learn about the segmentation of spoken sounds into words.

However, pairwise generalisation of learning examples can project out not only the sounds of words, such as 'r-a-n', but also the sounds of non-words, such as 'd-r-a-n'. If the model relied only on generalisation to project out the sounds of words, it would try to learn non-words, such as 'dran'.

The model assumes that in the early months of an infant's life, before she can infer the meanings of any utterances, she can nevertheless use the statistical distributions of the sounds she hears to segment those sounds into words.

The model does this by a simple form of statistical learning, described in the Appendix. When trained with the sounds of learning examples, this successfully learns more than 90% of the words used to make the examples, and learns very few non-words.

The model's statistical learning may or may not be a realistic model of the learning used by an infant. In any case, it shows that there is enough information in the distribution of sounds in the learning examples, to successfully learn the boundaries of words in the infant's native language, before starting to learn their meanings.

The point is not that this model of learning sound segmentation into words is the only model, or the best model – only that there must be models that learn sound segmentation, and do it well. Several capable models exist, such as the Chunk-Based Learning (CBL) model of [McAuley, Christiansen & Chater 2014]

7. Other Models of First Language Acquisition

In recent years there have been many computational models of language acquisition [Pinker 1984, 1989; Elman 1990; Niyogi 2002; Rohde 2002; Reali & Christiansen 2005; Chang Dell & Bock 2006; Chang 2008; Perfors Tenenbaum & Wonnacott 2010; Alishahi & Stevenson 2010; Nematzadeh Fazly & Stevenson 2012; Barak Fazly & Stevenson 2014; Ambridge & Blything 2015; Barak Floyd & Goldberg 2019; McAuley& Christiansen 2019]. Most of these are partial models, in that either: (a) they do not address some important aspect of language learning (such as the learning of word semantics, or of syntax), or (b) they assume that the learner already has some important linguistic knowledge at the start of the learning process.

Being partial models of learning, these models cannot be compared with the full range of data about child language learning, starting from an initial non-linguistic state; they can only be compared with selected aspects of child learning data.

There are very few models of language learning which can be said to be complete, in the sense that:

1. They model the language learning process from a start of no linguistic knowledge
2. They learn all the major aspects of a language – including segmentation of the sound stream into words, syntax, and semantics.

The model of this paper is complete in this sense. I know of only two other working computational models which approach this level of completeness. These are the model of Beekhuizen et al. [2014, 2015] and the model of Abend et. al [2019]. I shall compare the model of this paper with those two models, before comparing it with a few selected other models.

(A) The model of Beekhuizen et al:

The model of [Beekhuizen et al. 2014, 2015] differs from the model of this paper in many respects, but at a high level there are similarities:

- Language representations (including constructions and learning examples) are tree-like feature structures, which contain both sound and meaning, with no abstract syntax layer in between – as is usual in cognitive linguistic models.
- Both models use partial understanding of learning examples using known words, before using the examples to learn from.
- For parsing examples, Beekhuizen et al's COMBINATION operation is compatible with unification, as used in the model of this paper. The other three parsing operations in Beekhuizen et al are not like unification, but are broadly compatible with the 'partial parsing' of each learning example, using known words, used in the model of this paper.
- Both models use a best possible partial parse of each learning example, as the route to learning the productivity of constructions.
- For learning, Beekhuizen et al's ASSOCIATE operation (which they describe as 'simple cross-situational learning over the memory buffer', looking for overlapping subgraphs) appears to be compatible with generalisation, as used in the model of this paper to learn constructions of zero productivity.
- For learning, in the model of Beekhuizen et al., SYNTAGMATISATION creates a maximally concrete new and larger construction, and PARADIGMATISATION makes it more abstract (more productive). Combining these operations is similar to the way in which generalisation discovers subsumption links, as used in the model of this paper.
- The model uses 2000 artificially generated learning examples to reach 95% comprehension of examples – which is similar to the performance of this model. The two models appear to learn at comparable speeds (although it is not stated how many words the Beekhuizen et al. model learns)

A key difference between the two models is that the model of this paper has a concise mathematical basis in feature structures, unification and generalisation; and it has a rationale for this foundation, from Bayesian cognition and learning, as described in [Worden 2022b]. Beekhuizen at al.

do not present any mathematical basis for their model, but describe a set of algorithms for parsing and learning. These algorithms appear to be broadly compatible with the operations of unification and generalisation; so the two models may well have similar structures and algorithms, with different names.

The learning framework of Beekhuizen et al. depends on incrementing use counts of constructions, whereas the Bayesian model can learn with fairly small use counts, through a Bayesian learning criterion. I do not know whether this will lead to different rates of learning as the models scale to larger vocabularies.

I have not yet been able to draw out contrasting qualitative predictions from the two models; but they both appear to learn constructions as direct mappings between word sounds and meanings, in a 'verb island'- like manner [Tomasello 2003].

(B) The model of Abend et al:

The differences between the model of Abend at al. [2019] and the model of this paper appear to be more deep-seated, and (unlike the model of Beekhuizen et al), their model cannot be said to be a similar the model of this paper, using different terminology. The models make distinctly different predictions for the course of learning.

The most important differences are that the model of Abend et al. learns words in the syntactic categories of Combinatorial Categorial Grammar (CCG); and that the mapping between the sounds of a word and its meaning goes through an intermediate syntactic category, which is used for parsing and determines the functional form (λ-expression) of the word meaning (the constants in the λ-expression are word-specific.)

In this regard, the model of Abend et al. resembles the models of generative grammar – in which an intermediate syntax layer is used – rather than the models of cognitive linguistics, which have no such layer. Unlike most models in generative grammar, the model of Abend et al. is not a parameter-setting model; it does not rely on sudden discovery of discrete parameter values.

The model of Abend at al. uses a probabilistic grammar, which it learns by optimising the probability parameters of the grammar, incrementally as learning examples accumulate. I do not know how many parameters are varied at each learning step; but since the grammar is fully lexicalised, there must be at least some variable parameters for each word, as well as language-wide parameters.

In the model, the semantics of each word are modelled as a λ-expression (depending on its syntactic category), which has a tree-like structure, much like a feature structure. Composing λ-expressions is a form of unification. The leaves of each λ-expression are atomic values like '*doggie*', rather than sets of slots and values like 'animate = true', 'furry = true', 'legs = 4' as in the model of this paper. This is a design choice by Abend et. al., and could be changed – if, for instance, they wanted their model to embody semantic constraints of words, such as: 'only animate things can run' (such constraints are learnt in the model of this paper).

Although the two models are both Bayesian models, they are Bayesian in different senses:

- The model of Abend et. al finds the Bayesian best fit of a whole (fully lexicalised) probabilistic grammar, by varying the real probability parameters of the grammar as learning examples accumulate.
- Their model has three sets of parameters, making three products of probabilities: (a) P_{SYNTAX} is a product of the probabilities of the rules of the syntax (function applications); (b) $P_{MEANING}$ is a product of the probabilities with which each of the leaf syntactic category child nodes is assigned a given meaning; (c) P_{WORDS} is the product of the probabilities with which each meaning corresponds to a word in the sentence. The three sets of parameters are varied simultaneously to find the Bayesian best fit.
- The model of this paper is based on the operations of unification and generalisation, which are individually and locally Bayesian operations. Unification is a discrete maximum likelihood pattern match, and finds the Bayesian most likely fit to a sequence of word sounds in a meaning context; while generalisation (of a small set of learning examples, all containing the sounds of some word) is also a discrete pattern match, finding the meaning of the word which accounts for the largest part of the meanings of the examples.

So the learning model of Abend et al. is more global 'whole language' learning, and the model of this paper is more local and 'word by word'.

The model Abend et al. requires a set of real probability parameters to be optimised incrementally for each learning example, searching a multi-dimensional space of probability variables – unlike the model of this paper, where each learning example requires only discrete symbolic pattern-matching operations. Discrete operations are usually computationally less expensive than searching a high-dimension parameter space.

Some other differences between the two models:

1. The model of Abend et al. has been tested on real child-directed speech from the CHILDES database[MacWhinney 2000], rather than artificially generated learning examples. For this purpose, the CHILDES data has been semantically enriched.

2. The model assumes the correct segmentation of the sounds of learning examples into words – so it does not yet address the sound segmentation problem faced by a learning child. This limitation could be removed several ways – for instance, using the statistical learning word sounds as in this paper.
3. The model learns regular syntax (such as SVO order for transitive verbs) at the same time as it learns the syntax of individual verbs. So the model of Abend et al. appears to be less compatible than the model of this paper with the 'verb island' form of learning observed by Tomasello [2003] and others.

I have not been able to compare the overall learning rates between the two models.

The difference (3) in rates and onsets of learning for regular syntax appears to be the main difference in predictions for the course of early learning, between the model of Abend et al. on the one hand, and the models of Beekhuizen et al. and the model of this paper, on the other hand. It would require more detailed comparisons between the models to turn this qualitative difference into a quantitative difference to be tested against data.

(C) Other models of Language Learning

This section does not attempt to survey models of language learning – only to mention a few that relate specifically to the model of this paper.

As was described in section 6, the statistical model of pre-language learning of word segmentation is not the only possible model or the best model. One of these models of learning word segmentation is the Chunk Based Learning model (CBL) of [Christiansen & Chater 2016].

The CBL model learns to segment language into 'chunks' which are words or short phrases, using the statistics of distributions of words in large corpora, or of the words heard by individual children in those corpora. The statistics CBL uses are Backward Transition Probabilities (BTP), using these to infer chunk boundaries, and so to produce a 'shallow parse' of what a learning child hears. This shows that statistical learning techniques other than the simple learning used in this model can learn word segmentation, and can go further than that, learning the beginnings of syntax.

Some other models of language learning use structure matching operations which are similar to generalisation. For instance, the ECG model of [Bailey et al 1987] uses a 'model merge' operation which is like generalisation applied to shallow feature structures, to project out semantic features.

Several other models such as [Alishahi & Stevenson 2010] represent meanings as flat lists of feature values. In themodel of this paper, words are represented by tree-like feature structures, which may have semantic slots on any nodes; and all these slots can be learnt. It would seem that to account for both the semantics and syntax of most words, this more powerful form of learning is necessary.

8. A Theorem of Language Learning

This section shows that the learning mechanism can be used to learn any construction, in any language.

Constructions in any language can be represented as feature structures, and can be used to produce or understand speech by unification. If, as in this model, constructions are learnt by generalisation of feature structures, then a fundamental theorem of language learning can be proved:

Theorem: Suppose that speakers have a set of feature structures for words and other constructions, and produce sentences by unification of these feature structures. Suppose that learners hear those sentences, infer their meanings from the context, and learn constructions by generalising the resulting feature structures.

Through this process, feature structures for words and other constructions are replicated accurately from speakers to learners.

This result follows from the mathematical properties of unification and generalisation, because they are complementary operations. The result is proved in [Worden 2022c], and it can be seen working in examples in the demonstration. The working of the theorem is illustrated in the figure below:

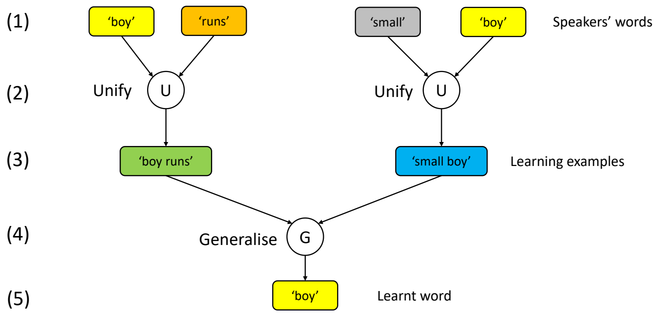


Figure 6: How generalisation learning leads to the theorem of language learning. The stages are (1) feature structures for words are in speakers' minds, including 'boy'; (2)-(3) speaking by unifying the words; (4) the learner generalises the learning examples; (5) the result is the word 'boy'

This shows how several learning examples, all using the word 'boy', all contain its sounds and its meaning as substructures. When those examples are generalised together, the feature structure for the word is recovered.

The result has important consequences for the diversity and durability of the world's languages.

Here I note one consequence: that the learning mechanism can be expected to work well for any construction, in any language. Through the theorem, any construction in any language will be passed faithfully from speakers to learners by generalisation learning. Initially, each construction is learnt in its unproductive [0,0] form. Later, as the construction appears in other learning examples with other known words, it is promoted to its most productive form, by the same generalisation mechanism (as in the demonstration).

The theorem gives strong reasons to expect that this model of language learning will be applicable to any language. Other consequences of the theorem are discussed in [Worden 2022c].

Furthermore, the Bayesian Learning Theory implies that the learning of any construction will be optimally fast or nearly so, requiring only a few learning examples which include the construction - the smallest number of examples which are needed to learn it reliably.

9. Learning Regular Grammar

Since [Pinker 1984, 1989] much attention has been devoted [Perfors et al 2010; Ambridge et al 2012; Barak et al. 2014; Ambridge & Blything 2015; Goldberg 2019] to how children learn the approximate grammatical regularities of their language, analysing childrens' propensity to make transient errors of over-regularisation, and their ability to handle novel words in examples like 'She wugs the ball'.

The model of this paper does not learn regular grammar. It learns in an entirely 'verb island' manner – learning that 'hits' has SVO order, and that 'hugs' has SVO order, and making no attempt to generalise to SVO order for novel verbs. It cannot yet be compared with learning data about grammatical regularities.

The model can be extended to learn grammatical regularities, by the same mechanism of generalisation which it uses to learn individual words and constructions. As soon as a few transitive verbs like 'hits', 'hugs' and 'eats' have been learned, it is possible to generalise their feature structures together, to learn the regular SVO pattern. If the model was extended in this way, it would predict:

- Grammatical regularities can only be learned after a significant number of words have been learned – so the learning of regularities must follow well after the learning of word islands – in agreement with child learning data [Tomasello 2003].
- Learning of any regular pattern (such as '-ed' for the past tense) depends on the type frequency of verbs that obey that pattern; whereas the learning of irregular forms such as 'went' depends on the token frequency of that word. The tradeoff between the two kinds of learning (as seen, for instance, in errors of over-regularisation) depends in a complex way on token frequencies and type frequencies. Predictions will depend on detailed data and models.

10. Discussion

This paper has described a working computational model of language learning, which can be seen running online.

In learning all aspects of a language rapidly from a standing start, this may be one of the most capable models of language learning in existence. The claimed merits of the model, and of the underlying theory of language are:

1. **Complete language learning, starting from no linguistic knowledge**: The model learns all the aspects of words and other constructions, including word segmentation, productive syntax, and rich semantics. It learns these from a standing start of no linguistic knowledge.
2. **Working computational model**: The model is fully implemented, learning with no manual supervision. It is available for testing by running it on different languages, different learning data sets, and so on.
3. **Proof that the model works**: Because the model has a simple mathematical basis in Bayesian inference and learning, it can be proven that the model works [Worden 2022c]. This is an advance on models that need experimentation and tuning to make them work.
4. **Fast and robust learning**: The model learns quickly, robustly and accurately – learning all aspects of a language needed to understand and use it productively.
5. **Agreement with data on first language acquisition**: the key facts are: (a) fast and robust learning; (a) initial learning of single words and constructions; (c) later learning of productive words; (d) 'verb island' learning of individual verbs, before learning regular syntax.
6. **Bayesian optimal model of language**: The core operations of unification and generalisation are Bayesian maximum likelihood operations. This is a Bayesian model of language learning and processing, aligning it with the many empirical confirmations of Bayesian cognition [Rao et al 2003; Chater & Oaksford 2008]
7. **Simple mathematical and computational basis**: In its Bayesian foundations, the model has a simple mathematical and computational basis - of feature structures, unification, and generalisation. These operations fit into a simple algebraic structure, underpinning the consistency and wide applicability of the model [Worden 2022b, 2022c].

8. **General applicability, to learn any language**: The theorem of language learning, described in section 8 and derived in [Worden 2022c], shows that in this model of learning, any construction in any language can be learnt from a few examples of its use.
9. **Requires only modest extension of primate cognitive abilities**: Language evolved in less than about a million years, so it cannot require major new cognitive faculties in the human brain – because they could not have evolved in only a few thousand generations [Worden 1996]. The faculties used in this model are all needed to support complex primate behaviour and learning [Worden 1994, 2022b]. Language uses existing primate cognitive faculties, with only modest extensions.

There are very few models of language learning in existence which have comparable strengths. This theory should therefore merit serious consideration and future testing.

There are many possible future tests of the learning model, such as:

1. Testing the model with real child language learning data, instead of artificially generated learning examples
2. Using the model to learn dialogue pragmatics and social uses of language.
3. Using the model to learn other languages, or larger subsets of a language

Appendix A: Implementation Details

This appendix describes some parameters and design choices of the computational model, which affect its performance.

The model can be run by downloading the model program from the demonstration site. The download is a compressed zip file with the following contents:

- An executable java .jar file to run the program
- A folder of data files which the program uses and updates
- Instructions to run the program
- The java source code

To run the learner, you only need to unzip this file, and then double-click the .jar file. It is an interactive program, with a graphical user interface similar to the web demonstration. As in the online demonstration, there is an 'Auto-Learn' button to run the learner. It uses randomly generated learning examples, and gives different results on each run.

Parameters of the learner are defined in a small file 'generator.xml' which can be edited to test different parameter values. The feature structures for words to be learned are defined in another data file. The program has a graphical editor to create or modify feature structures for words, and to save them in a data file. This can be used to apply the learner to a different subset of English, or to a different language.

The program includes an object-oriented Java implementation of feature structures and the three key operations of subsumption, unification and generalisation, as defined in [Worden 2022b]. These operations are used by the learning framework, to automatically learn words from a starting point of no linguistic knowledge.

Some details of the learning framework:

1. **Generation of Learning Examples**: To generate the learning examples, a set of words has been defined. The sounds, meanings and syntax of word feature structures have built with the editor, using semantic slots based on the work of Jackendoff [] and others. The words are classified into a few syntactic categories such as noun, adjective, or transitive verb. These categories are used in the production rules of a simple grammar to randomly generate candidate sentences and phrases. Some candidates are rejected, where the words cannot be unified because of their semantic constraints (e.g. you can only 'eat' a thing of type 'food'). Valid candidate sentences are used as learning examples, with meanings defined by unifying the words, and with duplicate examples removed.
2. **Pre-Linguistic learning of word sound segmentation**: Using the sounds of 1000 learning examples made as in (1), word boundaries are learned in two stages. In the first stage, a word with sounds such as [b-i-s-k-i-t-] is inferred to exist, if, for any partition of the sounds such as [b-i-s-] and [k-i-t], the occurrences of [b-i-s-k-i-t-] cannot be accounted for statistically as random coincidences of [b-i-s] followed by [k-i-t-]. This criterion reliably learns the more frequent words, and does not learn any non-words. In the second stage, the common words are used to partition the sounds of all learning examples; and the remaining sound sequences, when they cannot be further partitioned, are taken to be words. This two-step process learns the sound sequences of nearly all the words used to make learning examples, and usually learns no more than 1-3 non-words (which are typically concatenations of two words)
3. **Distractor Meanings**: For a proportion of the learning examples, it is assumed that the learner does not correctly infer the speaker's meaning, so the true meaning is replaced by a randomly chosen 'distractor' meaning of a different valid utterance. Each learning example is presented to the learner with one correct meaning and (N-1) distractors.

Currently N = 2, so that 50% of meanings are distractors.

4. **Extra Meaning Observed by the Learner**: When the learner correctly infers the speaker's meaning, he or she may observe extra meaning in the situation, which was not part of the meaning the speaker expressed. This extra meaning is modelled as random extra slot values in the feature structure for the inferred meaning. An average of 3 extra slots are added, to randomly chosen nodes in the inferred meaning, with a Poisson distribution of the number of extra slots, and with randomly chosen slots and slot values. Slot values which would conflict with the correct meanings of words in the example are not added.
5. **Zero-Knowledge Start**: The learner starts with no knowledge of the words or their syntactic categories. The learner observes only the learning examples – which are sounds and inferred meaning combined in a feature structure. Sounds occur in small phoneme-like units, but because of the pre-linguistic learning, these can be segmented into valid words. The learner only attempts to learn such valid words.
6. **Rejecting Distractor Meanings**: The learner tries to distinguish the correct meaning of each learning example from possible distractor meanings. When the learner knows some words in an example, the learner attempts to parse the example by unifying it with the known words, to partly understand it. It then chooses the example meaning with best aggregate match to the supported word meanings (the largest information content in the generalisation). When some words are known, this is a powerful way to distinguish the correct meaning of an example from distractors – typically rejecting more than 90% of distractors. When no words are known, the learner can only make a random choice, with probability 50% of choosing a distractor.
7. **The criterion for starting a new candidate word**: When two learning examples (which both contain the same word) are generalised together to make a new candidate word, the resulting feature structure has a meaning scene which taken to be the meaning of the word, with a small number of extra slots arising from random coincidences between the examples. If one of the learning examples is a distractor rather than the speaker's intended meaning, the information content of the generalisation will be very small, and of no use. A new candidate word is only started when the meaning information content is more than 4.0 bits, in the case where both learning examples have some known words (which helps to remove distractors); or 8.0 bits if one or other example has no known words, and so has a 50% chance of being a distractor meaning.
8. **The success criterion for refining a candidate word**: When a new learning example contains a candidate word, then the example is generalised with the candidate word, in an attempt to refine it. This refinement may, for instance, result in the removal of a slot value which has occurred by coincidence in the previous learning examples for the word, but is not part of the meaning of the word. However, the generalisation may remove too much meaning – particularly if the example meaning is a distractor. In that case, the learning example is rejected. The criterion for rejection of an example is that generalisation with it removes more than half the information content of the word meaning, or if it removes some subsumption link from a productive word.
9. **Permutation Learning**: The sequential learning of words, using learning examples in the order in which they are encountered, works fairly well, learning correct feature structures for more than 50% of words. However, a few words get off to a bad start (for instance, from a distractor learning example meaning) and then never recover. These words can be recognised by a high level of failures when refining them by generalisation with further learning examples. For those words, the following procedure is used: as soon as there are 20 learning examples for the word, 40 different candidate feature structures for the word are made. Each candidate is made by randomly permuting the 20 learning examples, making and refining the word as above. These 40 candidates are split into groups having the same learned feature structure (i.e. with the same meaning). Only the largest groups, of 10 or more candidates with the same meaning, are retained. The chosen meaning for the word is a word taken from one of those groups with the highest success rate in refining it. By choosing the commonest and most successful word learned from sets of randomly permuted learning examples, this procedure approximates to Bayesian optimal learning from those examples, and gives a high level of accurate word meanings.
10. **The criterion for accepting a word as 'supported' (known)**: This criterion is still the subject of experimentation. It might be defined from first principles of the Bayesian learning theory described in [Worden 2022b], but this has not yet been done. Instead, a word is taken as 'supported' when the proportion its 20 most recent learning examples (or all its examples if they are fewer than 20), which have been used successfully to refine it, exceeds a threshold.

11. **Measuring Information Content**: some choices made by the automatic learner, such as the criterion for success of refining a word by a learning example, depend on the information content of feature structures. This information content depends on the frequencies of occurrence of slots and their values. The information content of each slot value is measured from the frequencies of occurrence of that value in a sample of 2000 learning examples – so that the information content of feature structures is related to the learning examples observed by the learner, It depends on the frequency of occurrence of slots, as well as of their different values.
12. **Promotion of words to higher productivity**: As soon as enough words are known in two learning examples which share some productive word, to unify those words and know the input meaning scenes of the word being learned, that word can be learnt in its productive form, discovering the subsumption links by generalisation. Examples of this can be seen in the demonstration. From generalising only two learning examples, there may be some random extra meaning slots in the word. But the same word will have previously been learnt in its less productive form, using several learning examples. To avoid throwing away this information, there is a method to 'promote' a word to a more productive form, by taking its meaning from the less productive form, and taking the subsumption links from the latest generalisation of learning examples. This speeds up the learning process for productive words, and does not alter the principle of learning by generalisation.
13. **Preference for 'thing' words**: In the model, words which denote things (such as nouns and pronouns) have productivity [0,0]. So in understanding any learning example, those words can be unified before any other words are unified. However, at the early stages of learning, all words have learned productivity [0,0], and the learner does not know about syntactic categories such as verbs which can be promoted to higher productivity. There is a risk, for instance, that the learner will treat a verb or an adjective (whose eventual productivity is to be [1,0] or [0,1]) as a noun, because it has productivity [0,0], and may then use known adjectives to promote a noun to productivity [1,0], which would be incorrect. To reduce this risk, the learner is biased to unify words in examples which denote things, before unifying other words. Words denoting things are recognised by having higher information content in those slots (particularly the slot 't_type') which are used to classify things. This 'thing preference' can be interpreted as a young child having an earlier ability to understand words for things, before understanding words for actions and properties.
14. **Absence of Articles**: Articles such as 'a' and 'the' frequently occur just before nouns, and so the learner can easily interpret them as part of the sound of the noun they precede. To avoid this, I have assumed that very young children ignore articles in the sounds they hear. This assumption has been relaxed in some runs of the model, with the result that…
15. **Memory for Learning Examples**: Unlike the 'Now or Never' bottleneck model of learning [], this model requires some memory for learning examples. It needs to have some memory because Bayesian optimal learning cannot be done without it. For any word being learnt, the learner is assumed to retain at least the last 20 learning examples which contain the word, to support the near-Bayesian learning described above.
16. **Multiple Senses of Words**: The model does not yet learn multiple senses of the same word, but it is believed that the Bayesian-like learning by random permutation of learning examples can be easily extended to do this.

References

Abend O, Kwiatkowski T, Smith N J, Goldwater S and Steedman M (2019) Bootstrapping Language Acquisition, University of Edinburgh preprint

Alishahi A, Stevenson S. (2010) A computational model of learning semantic roles from child-directed speech. Lang Cogn Process, 25:50–93.

Ambridge B, Pine J M, and Rowland C F (2012) Semantics versus statistics in the retreat from locative overgeneralization errors. Cognition, 123(2):2

Ambridge B, Blything R. P. (2015.) A connectionist model of the retreat from verb argument structure overgeneralization. Journal of child language, pages 1–32.

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

Baldwin, D. A., & Tomasello, M. (1998). Word learning: A window on early pragmatic understanding. In E. V. Clark (Ed.), The proceedings of the twenty-ninth annual child language research forum (pp. 3–23). Center for the Study of Language and Information

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. Proceedings of the National Academy of Sciences, 106, 17284–17289

Barak L, Fazly A, and Stevenson S. (2014) Learning verb classes in an incremental model. In Proceedings of the 5th Workshop on Cognitive Modeling and Computational

Linguistics (CMCL 2014). Association for Computational Linguistics

Barak L. Goldberg A. & Stevenson S. (2016) Comparing Computational Cognitive Models of Generalization in a Language Acquisition Task, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 96–106

Barak, L. Floyd, S, and Goldberg, A. (2019) Modeling the Acquisition of Words with Multiple Meanings, Proceedings of the Society for Computation in Linguistics: Vol. 2 , Article 23

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.

Beekhuizen B (2015) Constructions Emerging; A Usage-Based Model of the Acquisition of Grammar, PhD thesis, University of Leiden

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. L. (1985). Morphology: A study into the relation between meaning and form. Amsterdam: John Benjamins

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. Psychological Review, 113, 234–272.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. Cognitive Systems Research, 9, 198–213.

Chang, N. C. L. (2008). Constructing grammar: A computational model of the emergence of early constructions. Unpublished doctoral dissertation. University of California, Berkeley.

Chater N and Oaksford M (2008) The Probabilistic Mind: Prospects for Bayesian Cognitive Science (eds)B, Oxford

Chater, N. & Christiansen, M.H. (2010). Language acquisition meets language evolution. Cognitive Science, 34, 1131-1157

Chater, N. & Christiansen, M.H. (2016). Squeezing through the Now-or-Never bottleneck: Reconnecting language processing, acquisition, change and structure. Behavioral & Brain Sciences, 39, e62.

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

Chomsky N.(1980). Rules and Representations. Oxford: Basil Blackwell.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39, e62.

Christiansen, M.H. & Chater, N. (2016). Creating language: Integrating evolution, acquisition, and processing. Cambridge, MA: MIT Press.

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

Elman, J. L. (1990) Finding structure in time. Cognitive Science 14:179–211.

Fillmore, C. (1982). Frame semantics. The Linguistic Society of Korea (ed.), Linguistics in the Morning Calm (pp. 111–137). Seoul: Hanshin Publishing Company

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

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.

Gowanlock D, Tervo R, Tenenbaum J and Gershman S (2016) Toward the neural implementation of structure learning, Current Opinion in Neurobiology 37:99–105

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.

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

Lakoff, G. (1987) Women, fire, and dangerous things: What categories reveal about the mind. University of Chicago Press.

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.

Lieven E. et al (2003) Early syntactic creativity: a usage-based approach, J. Child Lang. 30, 333–370.

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

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

McCauley, S.M. & Christiansen, M.H. (2014). Prospects for usage-based computational models of grammatical development: Argument structure and semantic roles. Wiley Interdisciplinary Reviews: Cognitive Science, 5, 489-499.

McCauley, S.M. & Christiansen, M.H. (2019) Language Learning as Language Use: A Cross-linguistic Model of Child Language Development, Psychological Review. DOI: 10.1037/rev0000126

Nematzadeh A., Fazly A., and Stevenson S. (2012). A computational model of memory, attention, and word learning. In Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics, pages 80–89. Association for Computational Linguistics.

Niyogi S. (2002) Bayesian learning at the syntax-semantics interface. In: Proceedings of the 24th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates; 697–702

Perfors A. Tenenbaum J. and Wonnacott E. (2010) Variability, negative evidence, and the acquisition of verb argument constructions. Journal of Child Language, 37(03):607–642

Pinker, S. (1984) Language learnability and language development. Harvard University Press

Pinker, S. (1989) Learnability and cognition: The acquisition of argument structure. MIT Press.

Rao R Olshausen B and Lewicki M (2002) Probabilistic Models of the Brain, MIT

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. Cognitive Science, 29, 1007–1028.

Rohde D. L. (2002) A connectionist model of sentence comprehension and production. Unpublished Doctoral Dissertation, Carnegie Mellon University

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.

Slobin, D. I. (1986). The crosslinguistic study of language acquisition: London: Psychology Press.

Talmy, L. (2000) Towards a cognitive semantics. MIT Press

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

Tomasello, M. (2009) The usage-based theory of language acquisition. In: The Cambridge handbook of child language, ed. E .L. Bavin, pp. 69–87. Cambridge University Press.

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

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

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

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