A Theorem of Language Learning

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 computational model of language acquisition and use 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 model shows how the same feature structures are used through unification to understand the meaning of heard speech, and to produce speech from meanings. The model embodies two key features of language – rapid learning (to learn any word from observing a few examples of its use); and unbounded productivity, combining words and constructs to express new meanings.

The learning model works, because it can be proven to work. This paper contains the proof, as a theorem. The theorem says that if language production and understanding is done by unification, and language learning is done by generalisation, then word feature structures replicate faithfully from speakers to learners.

The theorem implies that the model of language learning can apply to learning all constructions in any language. The faithful replication of word feature structures also underpins the durability, the remarkable diversity, and productivity of languages – just as the faithful replication of DNA underpins the diversity of life. This view leads to a model of language change as the evolution of constructions as species – rather than as the evolution of languages as species.

Keywords: Cognitive linguistics; feature structures; unification; generalisation; theorem of language learning; language diversity; evolution of constructions; emergence of grammar.

1. Introduction

There are very few working computational models of language acquisition which address the whole problem of language learning, from first words up to productive adult-like capabilities. A model which does so can be seen at http://www.bayeslanguage.org/demo, demonstrating every stage of language acquisition, and showing how learnt words are used, to comprehend and to produce language.

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

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

The model of language is based on feature structures, unification and generalisation. The operations of unification and generalisation are mathematically defined, and together form an algebraic structure, which is described in the second paper [Worden 2022b].

This leads to a fundamental theorem of language learning, which is proved in this paper - that through the learning processes in the model, feature structures for words and other constructions replicate accurately from speakers to learners.

The theorem is a proof that the model of language learning works.

The accurate and unconstrained replication of constructions is the underlying reason why languages are so diverse and can still be passed faithfully through the generations and across speaking communities – enabling languages to serve their communicative function, and to adapt over time to meet the needs of their users.

The learning theorem underpins all language, just as the faithful replication of DNA underpins all life.

When a DNA double helix splits and reproduces, each daughter DNA sequence is a faithful copy of the parent – so that information in the DNA sequence persists to later generations. Without this faithful replication of DNA, complex life could not emerge.

In the same way, as word feature structures are replicated through the speaking and learning process, their faithful replication (as guaranteed by the theorem) allows each word to evolve to serve the speaking community. Because the language is largely preserved across generations, it can adapt and improve, as the words in it evolve.

Just as the DNA sequence of base pairs is not constrained by the DNA replication process, in the same way the meanings and sounds in feature structures are not constrained by their replication process. This underpins the most remarkable empirical fact about language [Evans and Levinson 2013] – the huge **diversity** of human languages. Replication of feature structures does not constrain languages, and so allows this great diversity. The approximate universals of languages arise not from innate constraints, but from mechanisms of language change described in this paper.

The theorem implies that, while the computational model has so far been only applied to a small subset of English, it could be extended to any language, implementing the constructions of the language as feature structures. Because of the theorem, constructions are not restricted, and any construction can replicate faithfully through learning.

The evolution of the feature structures of constructions is the underlying reason why languages change over historic time. Each feature structure for a construction is under selection pressures to become more productive, less ambiguous, and easier to learn. These pressures lead to the high productivity and partial grammatical regularity of languages – including language universals.

In this model, regular syntax is not a fundamental property of language, and is not innate in the brain. Regular grammar is an after-effect of the evolution of words over historic time.

2. Mathematical Properties of Feature Structures

This section re-states from [Worden 2022b] some properties of feature structures which underpin the theorem.

A feature structure is a directed acyclic graph (DAG) of connected nodes. In the online model, these DAGs are shown as tree structures with cross-links (subsumption links). Each node may contain number of defined slot values. Each feature structure has an information content I,

which is approximately the sum of the information content of its slot values. A feature structure represents a set of situations in the world; the higher its information content, the smaller is the probability of that set of situations occurring. The probability is approximately 2^{-I}.

The unification C of two feature structures A and B is denoted by C = A U B, and their generalisation D is denoted by D = A ∩ B. These operations are defined in terms of a more basic operation of **subsumption**.

A feature structure A subsumes another feature structure B, (written as A > B) if and only if all the information contained in A is also contained in. B. The set of situations described by B is a subset of the set described by A; that is one way to remember the meaning of the '>' symbol.

Structurally, B is a larger structure than A; if A > B, all the nodes, slots, and edges in A are also 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.

Unification and generalisation are defined as follows:

(1) When C = A U B, C is the feature structure with least possible information content which obeys both A > C and B > C. If A and B are not consistent with one another, (A U B) does not exist; otherwise, A > (A U B) and B > (A U B).

(2) When D = A ∩ B, D is the feature structure with largest possible information content which obeys both D > A and D > B. (A ∩ B) always exists, and (A ∩ B) > A, and (A ∩ B) > B.

These definitions imply:

If A > B, then (A U B) = B (3)

If A > B, then (A ∩ B)= A (4)

A U B = B U A (5)

A ∩ B = B ∩ A (6)

For language to be productive, word feature structures must have subsumption links – which in effect, convey information from one branch of a feature structure tree to another, making it behave more like a DAG. These are described in [Worden 2022b], and ilustrated in the appendix of this paper. They do not alter the mathematical properties of feature structures described here.

Subsumption is transitive:

If A > B and B > C, then A > C (7)

It follows that

if A > B and A > C, then A > (B ∩ C). (8)

and that

if B > A and C > A, them (B U C) > A (9)

3. The Theorem of Language Learning

This section states the theorem, then gives its derivation.

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 constructs by generalising the resulting feature structures.

Through this process, feature structures for words and other constructs replicate accurately from speakers to learners.

Before deriving the result, I first summarise the model of language use, and summarise the model of language learning.

The model of language use between adult speakers is as follows:

- Adults in a speaking community all have a set of feature structures X, Y,… for the words and other constructions they know.
- Every word feature structure W has a sound part W_s, consisting of the sounds of the word, and a meaning part W_m.
- The meaning which a speaker intends to express in an utterance is denoted by a feature structure M.
- Usually a speaker cannot express all of this meaning, but can only express a part of it M' > M, depending on the words he knows.
- To do so, the speaker selects a word W whose meaning expresses a large part of M, and which subsumes M (W_m > M), and forms the unification W U M.
- When W is a productive word, this creates further intermediate meaning structures, which can be expressed by unification with further words.
- The utterance Z is produced by successive unifications, such as Z = X U (Y U (W U M))), where in this case the words are X, Y and W.
- The speaker then says the sound part Z_s, and the listener hears it.
- The listener understands the sounds by unifying the same words in the reverse order, making a feature structure such as C = W U (Y U (X U Z_s)))
- It can be shown that the meaning C_m understood by the listener is the same as the meaning which the speaker was able to express: C_m = M'.

The final result, derived in the appendix, shows how language acts as a faithful medium of communication, in a speaking community who all share the same word feature structures X, Y,.. The result can be derived from the

mathematical properties of feature structures. The derivation is a bit lengthy, and is given in Appendix A. It is worth working through it to get a clear idea of how unification is used for both language production and understanding. Alternatively, you can inspect many examples of language production and understanding in the online model.

Given this model of communication by language, the model of language learning is as follows (for convenience, a language learner may be denoted by 'she'):

- Speakers produce a set of utterances which serve as learning examples for a learner.
- Each learning example is denoted by a feature structure L_i (i = 1, 2 ,…). Each L_i has a sound part L_{is} and a meaning part L_{im}.
- A learning example made from a meaning M using words A, B, C.. has a feature structure given by L_i = A U (B U (C U M), in an appropriate order of unification.
- When a learner hears a learning example L_{is}, she cannot find the meaning part L_{im} by unification of all the words (as she does not know all the words), but on some occasions can infer L_{im} from the context.
- If the learner knows some of the words X, Y… in a learning example, she tries to unify all the words she knows with the example, making a partly unified example denoted by $L_k' = X \cup (Y \cup L_k)$
- When a learner has heard a small set of learning examples L_j which all contain some unknown word W, she forms a generalisation $G = L_j' \cap (L_k' \cap (L_m' \cap L_n'))$ of these examples (in any order).
- It can be shown that the unknown word W subsumes G (i.e., W > G); and that with increasing numbers of examples, G becomes an increasingly good approximation to W.

This is the learning process used by the model of these papers, and many examples of this word learning can be seen in the online model. The last point (the faithful replication of words through learning) is the result to be proved in this paper. The proof follows.

We first need to make one further assumption: that no learning examples for a word use that word in a nested context. In a nested context, such as the sentence: 'John said the pie was hot', the 'thing' node representing the pie occurs nested at greater depth in the meaning tree, than it would in a simple sentence such as 'the pie was hot'. So if these two meaning structures (nested and un-nested) are generalised together, the nodes will not match and the meaning 'pie' does not appear in the result.

Words like 'said' which introduce nested contexts have subsumption links connecting nodes at different depths in the tree. The word 'said' creates a nested context, but need not itself occur within one. It is a reasonable assumption that a learner can learn any word from its simplest uses, in non-nested contexts.

I first consider non-productive words and constructs – fixed forms which cannot (on their own) combine productively with other forms (they need the other forms to be productive). In the model, feature structures for these constructs have no subsumption links.

Any non-productive word can be learnt without knowing any other words. In this case, the learner cannot partly unify any learning examples, so all $L_i' = L_i$.

Consider a set of learning examples L_j all using some unproductive word W. Each learning example for W was made by a sequence of unifications which includes a unification with W, and unifications with other words X, Y and so on, such as

$$L_5 = X \cup (W \cup (Y \cup (A \cup M)))$$

Using the result (1) that A > (A U B), and the transitivity of subsumption (7), we can show that

$$W > L_i$$

for any learning example L_i that uses W (not in a nested context[1]). Now if the learner generalises several of these learning examples[2] to make G:

$$G = L_i \cap (L_j \cap (L_k \cap L_m))$$

Repeated application of (8) implies that

$$W > G$$

As the number of learning examples increases, G rapidly converges towards W. This is because any extra information (in G but not in W) can only come into G from other words, which occur in all the learning examples Li. As the number of examples increases, the probability of any such coincidence decreases exponentially. So after a fairly small number of learning examples, to a good approximation

$$G = W$$

This implies that G has the same sounds as the word, and the same meaning:

[1] Technically, the relation A > A U B always holds for whole feature structures, because unification is a structure-growing operation; but it only holds for final meaning parts on their own, if the unification involves no subsumption links which create nested contexts in the final meaning part.

[2] The learning model needs to have a mechanism for rejecting any learning examples in which the meaning has not been correctly inferred from the context.

$G_s = W_s$

$G_m = W_m$

This proves the result for any non-productive word or construction W. The combination of unification (for use of the word by speakers) and generalisation (for learning) leads to faithful replication of the feature structure for the word.

To extend this result to productive constructions, we need to consider subsumption links, which are the source of language productivity in the model. A productive feature structure (like that shown in the appendix, figure 3) has the following parts:

- It has a sequence of **sound segments**.
- Interspersed with these (possibly before them, or after them, or intermingled with them) are one or more **input meaning parts**, which are meaning subtrees beneath a 'scene' node.
- It has one **result part**, which is a meaning subtree describing its meaning (and in the model, always appears at the bottom of the diagram)
- Each input meaning part is the start of one **subsumption link**, which ends on a node in the result meaning part.

In order to learn the feature structure for any productive word, it is necessary first to know some other words – those other words whose meanings are the input meaning parts of the word to be learned, in the learning examples.

When sufficient other words are known, the learner can use those other words to partially unify learning examples L_i, giving examples L_i' in which all the input meaning parts of a word to be learned have been unified (and so are meaning scenes rather than sound nodes). In all of these learning examples, there is some node n_2 in the example meaning L_{im}', and some node n_1 in the input meaning scene (got by unifying known words) P_{im}, which obey a subsumption relation:

$n_1 > n_2$

This relation between nodes can hold because L_{im}' is the full meaning of the learning example (inferred by the learner from the context), whereas P_{im} is a part of the full meaning (got by unifying known words); so there can be a subsumption relation between nodes in them. Because of this relation, generalising the learning examples discovers a subsumption link between the nodes, giving the full feature structure for a productive word.

Examples of generalisation discovering subsumption links can be seen in the online demonstration.

This means that the core result for non-productive constructs (that through learning, a word feature structure replicates faithfully in both its sounds and its meanings) carries over to productive words and constructions. The algebraic relations between learning examples and their generalisations carry through for learning productive words, just as for unproductive words.

This proves the main result – that because of the complementary nature of unification (used in language generation and understanding) and generalisation (used in learning), language generation and learning together form a precise replication process for feature structures. Feature structures for all constructions pass accurately across speaking communities and through generations, by the complementary processes of language use and learning.

Every stage in the process is illustrated by examples in the online model.

The theorem is summarised for the word 'boy' in the diagram below:

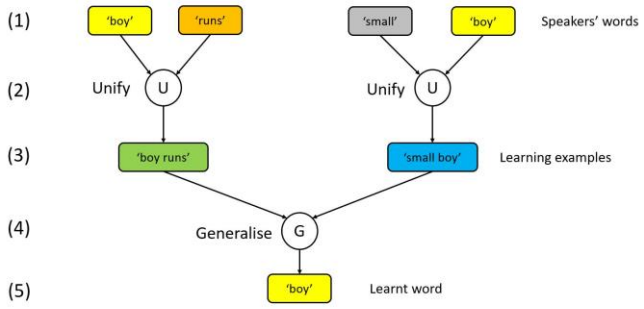

Figure 1: Illustration of how feature structure unification (used in speech) and generalisation (used in learning) accurately transmit the sound and meaning of any word.

The steps in this diagram are:

1. Speakers know the feature structures for words.
2. To form sentences, they unify these feature structures.
3. A child hears the sentences including a word, and infers the speaker's meaning from the context, to make learning examples for the word.
4. The child generalises the learning examples together, which include any one word. ('boy' above)
5. This recovers the feature structure for the word used by the speakers

4. Universality of the Theorem

The first thing to note about this theorem is its universality. The result depends only on the mathematical properties of feature structures and their operations of unification and generalisation. It does not depend on the detailed form of the semantic representations of meanings (we could use different slots and node types), and it does not depend on phonetic details (we could use different units of sound, or even sign language or printed text); with any of these changes, the result would still hold.

The result applies to any construction in any language, however complex is its semantic representation or its realisation in sound - as long as the meaning of the construction can be captured in a tree-like feature structure, and is expressed in some sequence of sounds. Both meaning trees and sound sequences have unbounded information capacity.

This universality is the core reason why the world's languages can take such diverse forms [Evens & Levinson 2013]. Languages do not need to fit into any syntactic, semantic or phonological straitjacket, in order to persist and spread through learning.

It is also a reason to believe that the learning model can apply to learning all languages, from childhood learning up to productive adult use of those languages. Any construction in any language can first be learned in a fixed, unproductive form – however complex is the concept that it expresses, or the stream of sounds it uses. Then the construction can later be extended to more productive uses, as examples of these uses are heard and understood. No regular syntax is needed for this process; it will work, however irregular the language.

5. The Analogy with Replication of DNA

The theorem states that when word feature structures replicate by unification and generalisation, they do so accurately – information is neither gained nor lost.

This theorem is to language, as DNA replication is to life.

When a DNA double helix splits in order to replicate, each strand picks up complementary bases from the surrounding medium – forming an exact replication of the double helix, preserving the information in the sequence of base pairs with very few errors and no constraints on the order of bases. Any sequence of base pairs can replicate in this way.

This precise replication of DNA has enabled every life form to reproduce faithfully, retaining its best features and slowly accumulating other useful features. The faithful replication of DNA, for any sequence of base pairs, has enabled all life on earth to survive and evolve.

In the same way, the faithful replication of word feature structures enables any language to be transmitted faithfully through the generations and across communities. Were it not for this faithful replication, the useful features of words oculd be lost to errors of replication. Words could not survive over many generations, or slowly improve to better meet the needs of their users, as society changes. Languages could not persist and improve.

It is important to note that, just as the replication of DNA is 'neutral' in that it does not constrain the order of DNA base pairs (and so does not restrict the genetic information carried by the DNA), so the replication of feature structures does not constrain either their sound part or their meaning part. This allows the feature structure replication to support the remarkable **diversity** of human languages [Evans & Levinson 2013] – just as DNA replication supports the remarkable diversity of life.

This replication theorem is as fundamental to language, as DNA replication is to life.

While DNA replication is very precise – with error rates less than 10^{-9} per base pair – the DNA of any species typically has a lot of diversity. Sometimes, boundaries between species are not precisely defined. Similarly, while the word replication theorem allows any word feature structure to replicate precisely, actual words are only as precise as their speaking community needs them to be. Typical words have a variety of overlapping senses, which may change with time. Different speakers may have learnt a different mix of senses of any word. The fidelity of word replication allows this to happen, ensuring that meaning details are not lost in replication, and that word meanings are not constrained by replication.

This is analogous to the diversity of genotypes within a biological species. If a word has several different senses, with different feature structures, over time a learner can learn them all. Words which are not learned do not survive as word species.

6. Word Evolution and Language Change

Analogies between language change and biological evolution have been made since Darwin. For a recent review by several leading authors, see [Dediu et al 2013]. These analogies are now important in many approaches to language such as [Christiansen & Chater 2010; 2016]. The theorem proved here suggests that the evolving unit is not, as has been commonly supposed, a language; but that words and other constructions are the 'species' which evolve. A language is more like an ecology than a species[3]. A construction such as a word may thrive in several language ecologies.

This viewpoint alters the perspective on language change as evolution, raising new questions. See [Croft 2007] for other considerations about replicators and evolution in language.

[3] The 'language as ecology' analogy is motivated as follows: an ecology is a set of many species in a region, in which every species depends on other species for its survival. A language is a set of many word species (more precisely, many construction species) in a speaking population, in which every word species depends on other word species for its use and survival.

Through the processes of unification and generalisation, the feature structure for any word can replicate across many generations, with very few errors; and it can be varied deliberately as new expressive needs arise. Therefore each word in a language is like a species, which evolves as society changes, to make it more useful to speakers in each generation. Every species of word is subject to natural selection, so that only the fittest species survive. For instance, when two languages merge through migration or conquest, words in one language can be displaced by more useful words in the other language. A word can become extinct, in only a few generations.

There are several selection pressures on each species of word. Each word needs to be used sufficiently often to be heard and acquired by the next generation, in order to survive. Some of these selection pressures on a word species are:

1. A word must express a useful meaning, so that people need to use it.
2. The meaning must be sufficiently unique that it cannot be expressed more easily by other words.
3. Words should be as short as is needed to convey their information content, and should not create ambiguities that a listener cannot easily resolve.
4. If the word can be used productively in combination with other words, that can greatly increase the range of circumstances in which it is used - increasing its fitness
5. If different variants of a word (such as different tenses and numbers of a verb) are related in a regular fashion, that makes them easier to learn together, and increases their fitness.

These selection pressures on words are very strong, and the evolution of word species occurs within a few human generations - thousands of times faster than the evolution of the human brain.

Given these diverse selection pressures on words, there is no simple way to predict which words will thrive and which will disappear. I suggest that one selection pressure is particularly important – the pressure towards productivity. This is because greater productivity multiplies the number of occasions in which a word or construct will be used – greatly increasing the opportunities for learning it.

7. Regularity and Language Universals

One evolutionary pressure has important holistic effects on any language – the pressure towards syntactic regularity.

In the evolution of words, there is a network effect, leading words to cluster in groups with regular grammar. Having a regular grammar enables a word to be used productively, in combination with other words which have compatible grammar, increasing its productivity and fitness. That is why every language in the world is partly regular, particularly in its less frequently used words. The most common words in a language must be used frequently enough to enable them to replicate, and have less need for regularity. For the most commonly used words, the need for brevity trumps regularity.

There are also selection pressures that lead towards semantic regularity, as opposed to syntactic regularity This can be illustrated by the different systems of verbs of motion in many languages.

Our internal semantic representation of motion appears to encode both for the path of motion, and for the manner of motion - as might be expected, if feature structure systems evolved before language partly to plan physical movements [Worden 2022b]. However, if a verb of motion encoded both the path and the manner of motion, its meaning might be too narrow - giving it too few occasions of use, so it could not be learned and survive. Hence there is a selection pressure on verbs of motion to encode less meaning, and have broader meanings.

There is a known tendency for verbs in a language either to encode manner of motion, or to encode path, depending on the language, but not to encode both [Slobin 1996]. Consider a mixed single-language ecology, in which some words encode path, and some words encode manner. There will then be occasions – where, as often happens, the speaker knows both the path and the manner – where two different verbs are equally applicable. Effectively, the two words compete for the same niche in meaning space – approaching it along different axes - which lessens the fitness of both words.

The disadvantage is clearly greater for words with the less common of the two meaning axes – because those words have a larger number of words using the other meaning axis to compete with them. This is an unstable situation, and will soon resolve itself, in any language, to have predominantly words with only one of the meaning axes – path or manner. This is what we see in many languages, such as English (manner) or Spanish (path). There are similar contrasts in prepositions of location, for instance between English and Korean [Bowerman & Choi 2003].

Generalising from these cases, I conjecture that in many cases, what may appears to be a syntactic regularity may turn out, on closer inspection, to be a semantic regularity. Words line up in regular patterns of meaning.

For instance, [Worden 2002] has shown how several of the Greenberg-Hawkins Universals [Greenberg 1963; Harkins 1994] can be motivated by a semantic requirement to avoid hard-to-resolve ambiguities. The English phrase 'the lid of the box on the table' is ambiguous; but because English obeys Greenberg's Universal No. 2, the two possible parses – (the lid of (the box on the table)) and ((the lid of the box) on the table) – both denote a kind of lid; so understanding can proceed in the presence of the ambiguity, and it can be

resolved later. If English did not obey the universal, this would not be the case; the two parses would denote different things. So a syntactic regularity about word order has a semantic motivation – it results from semantic selection pressures on words, to avoid hard semantic ambiguities.

Since the evolution of words happens thousands of times faster than the evolution of the human brain, the partially regular grammars of the world's languages tell us little about the structure of the human brain [Christiansen & Chater 2016]; grammatical regularity tells us only about which words survive in a language. This goes against notions of the innateness of grammar [Chomsky 1965].

In the study of language, grammar has always been Exhibit A: 'you start from here.' Regular syntax has such an intellectually satisfying, almost mathematical structure, that people have always thought: 'Surely grammar must be telling us something fundamental about language'.

In this model of language, it is not. Grammar is like the 'steam' you see coming from the funnel of a steam engine. When you first see steam, you may think: 'those big puffy white clouds must be important; surely they are telling us how the locomotive moves'. It turns out that what you called 'steam' is not steam at all – it is water vapour, which is an after-effect of hot steam meeting cold air. The real steam is a superheated transparent gas in the cylinders, which really pushes the locomotive along.

In the same way, the real locomotive force behind language is semantics – the need to say meaningful things to people. This is what has driven the evolution of words (and other constructions, like canned phrases and idioms) to say useful things – and then to be productive and say countless new things in combination with other words. Only after achieving productivity did words arrange themselves into partially regular grammars – so as to make each word easier to learn, and easier to combine with other words without ambiguity.

In this picture, regular grammar is an after-effect of the survival strategies of individual word species – the means adopted by less commonly used words in order to survive the competition of language use. Common words like 'be', 'have' and so on did not need regular grammar in order to survive – and have stayed short and irregular.

There is no need to give every construct of every language a place in a grammar. When you encounter a novel sentence like: 'Fred sneezed the napkin off the table' [Goldberg 1995], you need not worry that 'sneeze' is normally an intransitive verb, without an object like 'napkin'. This sentence is just what somebody happened to say, because their language allowed them to do, so and people understood it, by making a mental image of the act. If such a construct later becomes a part of a regular grammar, that has no primary importance. It does not tell us anything fundamental about language. It merely tells us about the historic process of language change.

From another viewpoint the forces that lead to regular grammar are chaotic, in the mathematical sense of chaos theory [Gleick 1987]. In language change, there is a high degree of amplification of small changes, leading to unpredictable outcomes. Some grammatical construct may arise in one construction, in some small speaking community; then it may spread to a larger population, and attract other word species to form an island of regularity – or alternatively, some different island of regularity may engulf it. There is so much amplification in this process, that it is best described by chaos theory.

The insight that the historical processes of language change should be seen as the evolution of word feature structures in language ecologies, rather than (as is commonly supposed) the evolution of language organisms, is related to the view that a language has no centralised grammar, existing apart from its words and other constructions. Grammar is fully lexicalised, in that each construction carries its own piece of grammar around with it, in its feature structure(s). If many words in a language ecology agree on the shape of their grammar, that is just a consequence of the evolution of those words over historic time.

8. Conclusions

In contrast to Chomskyan Generative grammar [Chomsky 1980, 1981, 1995], which has placed strong emphasis on formal mathematical structures of language, work in Cognitive Linguistics has placed less emphasis on formal mathematics, and more emphasis on the integration between language and other cognitive faculties – which typically are not seen as having a concise mathematical description.

The model of language learning described in these papers [Worden 2022a, b, c] is a cognitive linguistic model. So in its mathematical foundations it differs from much work in cognitive linguistics.

The mathematical foundations of this model of language and language learning derive from the theories of Bayesian optimal cognition, and Bayesian optimal learning [Worden 2022b]. As has been shown by the many applications of the Free Energy Principle [Friston et al 2006; Friston 2010], principles of Bayesian optimality can be used to derive a precise mathematical framework for cognition, and that framework can lead to many insights – as well as working computational models.

So it is not a surprise that this Bayesian model of language has concise mathematical foundations – in the relations between the operations of subsumption, unification, and generalisation, which are formally similar to the relations of set theory. Those relations lead to useful simplifications and

constraints on the computational model, and lead to the theorem proved in this paper.

That theorem implies that through the complementary operations of unification and generalisation, language learning leads to faithful replication of word feature structures in a speaking community. The faithful replication of word feature structures is analogous to the faithful replication of DNA; I suggest that its consequences may be as important for language, as DNA replication is for life.

This paper has begun to explore those consequences:

1. The theorem implies that any feature structure, for any construction, will replicate faithfully though learning. So the replication of feature structures is not a constraint on the feature structures of any language – consistent with the longevity and diversity of the world's languages.
2. The theorem implies that the model of learning works well. While the model has so far only been tested on a subset of English, we expect it to be able to learn all constructions in any language
3. The theorem leads to a model of historic language change, through the evolution of words as 'species' in an 'ecology' defined by a language. This model is consistent with many properties of languages – such as their partial syntactic regularity, and known language universals.

Appendix A: Faithful Communication

As background to the main result of this paper (which is the faithful replication of word feature structures through the generalisation learning mechanism) it helps to derive a more basic result, about faithful communication.

This result is the reason why language works as a communication mechanism. Precisely, it says that:

- if speakers and listeners know the same set of word feature structures, and
- if speakers create utterance sounds from meanings by unifying word feature structures, and
- if listeners re-create meanings from the sounds by unifying the same word feature structures
- then the listeners re-create the same meanings that the speakers expressed.

The informal proof of this result proceeds by recursion, from the simplest one-word utterances up to more complex utterances. In the online model you can see examples illustrating each stage of the proof.

In the model, the feature structure for a simple unproductive word like 'boy' is as shown in figure 2:

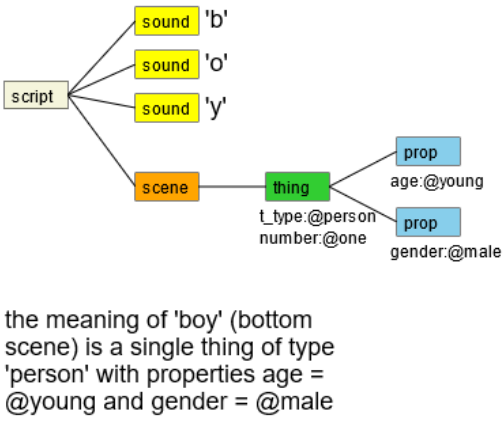

Figure 2: Feature structure for the word 'boy'

This feature structure (which we will call B for 'boy') has two parts – a sound part, consisting of the three yellow sound nodes making the sound 'boy', and a meaning part, consisting of the 'scene' node and the tree below it.

The sound part is denoted by B_s, and the meaning part is denoted by B_m. These are both sub-structures of B, so they obey:

$$B_s > B$$

$$B_m > B$$

A speaker wishing to express the thought 'boy' starts with only a meaning (denoted by a feature structure M) in his mind. He uses this meaning to retrieve a word feature structure whose meaning best matches it[4] – in this case, the feature structure B for the word 'boy' which partially describes his meaning:

$$B_m > M$$

As it obeys the subsumption relation, B_m does not add any information to M – so B is allowable as a word, to express M and nothing more. The speaker unifies the word feature structure B with the meaning feature structure M, forming

$$P = (M \cup B)$$

The sound part of the result comes only from B, so it obeys

$$P_s = B_s$$

That is, its sound part is the sound of the word 'boy' from the word – which the speaker then says.

[4] The meaning in the speaker's mind may have more information, such as the colour of the boy's hair. The speaker chooses a word which expresses as much of this meaning as possible – as long as the word meaning subsumes the speaker's meaning – i.e. the word does not contradict the meaning or add to it.

A listener hears the sound part B_s. She uses these sounds to retrieve the best-matching word feature structure that she knows – in this case, the feature structure B for the word 'boy', because it has the same sequence of sounds. She then unifies B with a feature structure containing the sounds she has heard, to form the feature structure

$$Q = (B \cup B_s).$$

Equation (3) then implies that

$$Q = B$$

So that

$$Q_m = B_m$$

In other words, the meaning constructed by the listener is the meaning of the word 'boy', which was expressed by the speaker. So if speaker and listener both know the same feature structure B for 'boy', the meaning is conveyed faithfully from speaker to listener.

The same result applies to more complex multi-word constructions such as "How d'you do?", using a single unification both to express the construction and to understand it, when it is used in an unproductive form.

This slightly laboured derivation shows how a single word meaning can be conveyed faithfully from a speaker to a listener, by encoding it in sound, then decoding it. You can follow though these steps in the online demonstration, and they are the foundation understanding how more complex utterances work.

When productive words are used, more unifications are involved. I illustrate this by a two-word sentence 'boy runs'. The feature structure for the productive word 'runs' is shown in figure 3:

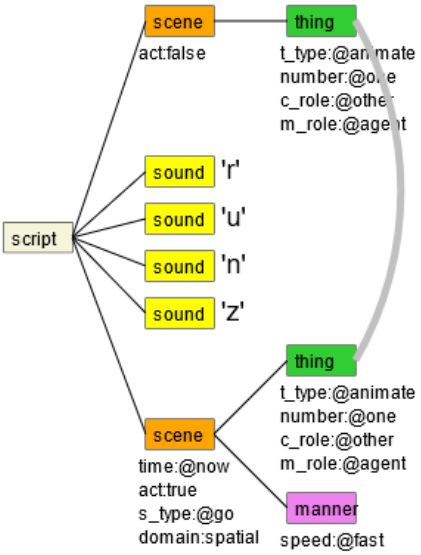

Figure 3: Feature structure for a productive word 'runs'. The curved line is a subsumption link.

The feature structure for 'runs' is denoted by R, and it has a sound part R_s and a meaning part R_m. In this case, the meaning part has two 'scene' subtrees, joined by a subsumption link.

Now suppose a speaker intends to express the meaning 'boy runs', which he has in his mind as the feature structure:

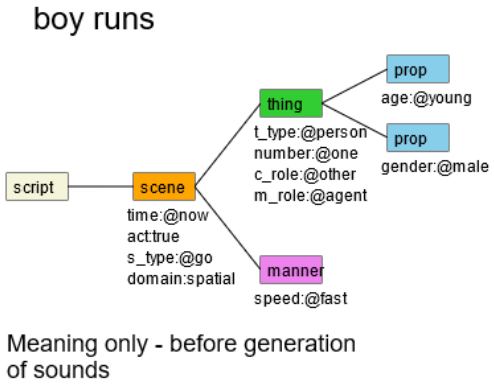

Figure 4: Feature structure for the meaning 'boy runs'

This feature structure (which is denoted as N) describes a scene of movement in which the moving thing is young and male – that is, a boy. Since N has only a meaning, and no sounds, it obeys:

$$N_s = \text{empty}$$

$$N_m = N$$

The meaning part N_m matches to the second meaning scene of the word 'runs' – but differs from it, in that it gives more information about the running thing. Formally:

$$R_m > N_m$$

Since the feature structure for 'runs' expresses a large part of the meaning N, but (because it subsumes N) does not add any extra information which is not in N, the word 'runs' is eligible to express the meaning. Suppose that it expresses more of the meaning N than any other word known to the speaker. This causes the speaker to retrieve the feature structure R for 'runs' and unify it with the meaning, to form

$$N' = (R \cup N_m)$$

The resulting feature structure N' is shown in figure 5.

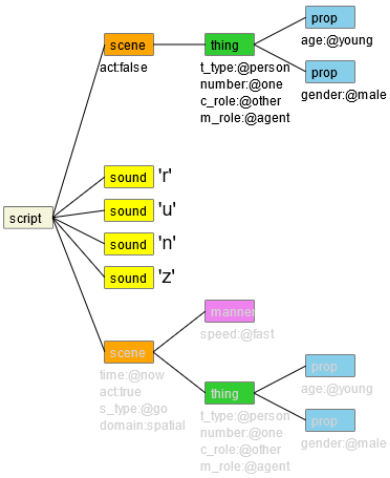

Figure 5: Feature structure N' =R U N

Because N' was made by unifying the feature structure R for the word 'runs' with the meaning N, it has a similar form to R; in particular, its sound part N'$_s$ is the same sound 'runs' as the word sound. There are three differences between R and N':

1. The bottom meaning scene has been 'greyed out' to denote that it has already been matched, and is no longer available to be unified with other words
2. Information has been added to the upper meaning scene, because of the subsumption link (curved line) in the word 'runs'. The extra information in the speaker's meaning N about the boy (the lower 'thing' subtree in figure 4, with gender male and age young) has been conveyed upwards to the upper 'thing' node.
3. N' has no subsumption link

Another unification is then required to express the boy-like upper meaning node in N'. This proceeds in the same way as the first example, where the speaker expressed the idea 'boy' on its own.

First the speaker selects the most appropriate noun feature structure – to match as much as possible of the upper meaning tree, without adding any information to it. This is the feature structure B for the word 'boy', shown in figure 1. Then the speaker does the unification:

N" = B U N' = B U (R U N).

The result N" is shown in the figure 6:

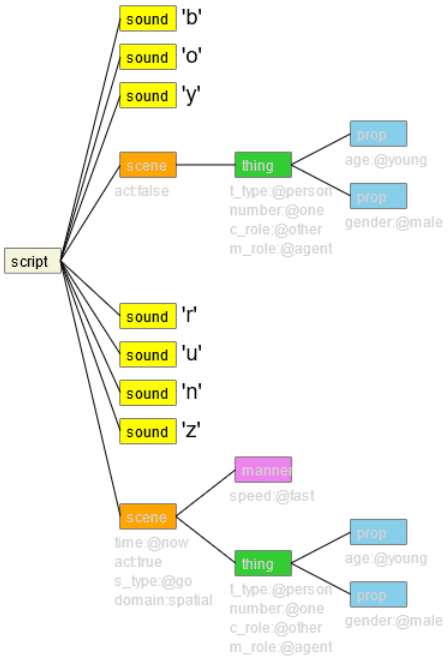

Figure 6: Feature structure N" = B U (R U N)

In this figure, the two meaning parts have both been greyed out, because they have both been matched with word feature structures. All that remains are the sounds in N"$_s$ for 'boy runs' – which the speaker says, and the listener hears.

To understand the sentence N"$_s$, the listener unifies the two word feature structures in reverse order, to form Q' :

$$Q = B \cup N"_s$$

$$Q' = R \cup Q = R \cup (B \cup N"_s)$$

The first unification to make Q acts exactly as it did in the first example, of expressing 'boy' on its own; so the meaning part Q_m is the meaning of the word 'boy':

$$Q_m = B_m$$

In the second unification Q' = R U Q, the subsumption link in R acts in the opposite direction – moving the meaning of 'boy' downward from Q_m into the running meaning scene. This means that the final meaning Q'_m constructed by the listener is the same as the speaker's original meaning:

$$Q'_m = N_m$$

So when a single non-productive word (boy) is combined with a single productive word (runs), the subsumption link in the productive word acts in sentence generation to extract the meaning of 'boy' from the running scene; and in understanding, the same link acts in the reverse direction, to move the 'boy' meaning back into the running meaning scene. By this process, the listener faithfully reconstructs the speaker's meaning. It is a novel meaning, for which no single word exists; language is productive.

Longer utterances require more unifications – one unification for each word or construction, in either

generation or understanding. Provided the unifications for understanding are done in the reverse order of those for generation, the same result can be shown, by recursion in the increasing number of unifications. In all cases, listeners can faithfully reconstruct speakers' meanings.

You can see how this works in the online demonstration, for more than 50 utterances of up to 6 words, involving many different English parts of speech.

This model of language production and understanding is a simplified model, in that:

1. The word unifications probably do not proceed in a simple serial manner, but may proceed in parallel, and more in a 'first heard, first unified' manner, consistent with the 'Now or Never' bottleneck of [Chater & Christiansen 2016].
2. Contextual information in the 'common ground' object model of shared context [Worden 2022b] may also be unified in parallel with heard words, to help remove ambiguities as early as possible.

Nevertheless, the simplified model allows us to understand the faithful nature of language communication.

References

Bowerman, M., & Choi, S. (2003). Space under construction: Language-specific spatial categorization in first language acquisition. In D. Gentner, & S. Goldin-Meadow (Eds.), Language in Mind: Advances in the Study of Language and Thought (pp.387-428). Cambridge, MA: MIT Press.

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.

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

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

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

Croft, W. (1996). Linguistic selection: An utterance-based evolutionary theory of language. Nordic Journal of Linguistics, 19, 99-139. Cambridge University Press

Croft, W. (2000). Explaining language change: An evolutionary approach. Harlow, Essex: Longman.

Dediu, D., Cysouw, M., Levinson, S. C., Baronchelli, A., Christiansen, M. H., Croft, W., Evans, N., Garrod, S., Gray, R., Kandler, A. & Lieven, E. (2013) Cultural evolution of language. In: Cultural evolution: Society, technology, language and religion, ed. P. J. Richerson & M. H. Christiansen, pp. 303–32. MIT Press.

Evans N. and Levinson S. (2013) The myth of language universals: Language diversity and its importance for cognitive science, BBS (2009) 32, 429–492

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

Gleick, James (1987). Chaos: Making a New Science. London: Cardinal.

Greenberg, J. H. (1963) Some universals of grammar with particular reference to the order of meaningful elements. In: Universals of language, ed. J. H. Greenberg, pp. 72–113. MIT Press.

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

Hawkins J. A. (1994) A performance theory of order and constituency, Cambridge University press

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

Slobin, D. I. (1996). Two ways to travel: Verbs of motion in English and Spanish. In M. Shibatani & S. A. Thompson (eds), Grammatical constructions: Their form and meaning (pp. 195–217). Oxford: Oxford University Press.

Worden R. P. (2002) Linguistic Structure and the Evolution of Words, in 'Linguistic Evolution Through Language Acquisition', Briscoe E (ed.) Cambridge

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

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

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