A Model of Language Learning

Robert Worden

UCL Theoretical Neurobiology Group

rpworden@me.com

See the model running at:

http://www.bayeslanguage.org/demo/

Overview

- These slides accompany a working model of language learning, that can be seen running at http://www.bayeslanguage.org/demo/
- The model is described in three papers:
  1. A Computational model of language learning
  2. A Model of Language Learning: Foundations
  3. A Theorem of language Learning
- These slides are a summary of the papers.
- The model is very capable (fast, accurate, complete language learning), because it has a simple mathematical basis.

(1) A Model of Language Learning

Core Principles:

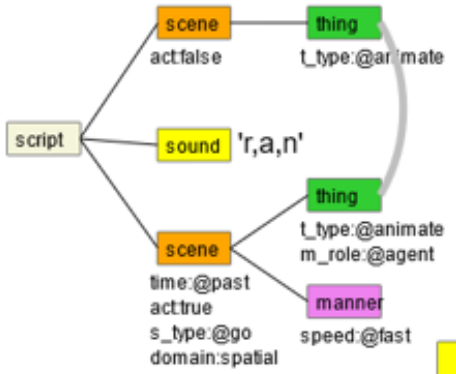
- Defined at Marr's Level 2 – algorithms and data structures (it is not a connectionist model).
- A Cognitive Linguistics Model (no autonomous syntax layer). Constructions = Feature Structures
- A Bayesian model (unification and generalisation are Bayesian maximum likelihood operations)
- Speakers and Listeners have a 'Common Ground' of mutual understanding of the current context

Core Learning Mechanisms

- Utterances are produced and understood by **unifying** feature structures, for words and other constructions (sounds and meanings)
- Learning examples are utterances, with their inferred meanings, represented as feature structures
- Word feature structures are learnt individually, by **generalising** learning examples containing the same word
- Generalisation and Unification are complementary operations; that is why learning works.
- Learning is approximately Bayes-optimal - done by permuting learning examples before generalisation

Running the Model

- http://www.bayeslanguage.org/demo/
- 40 English words are used to generate learning examples - short utterances, with meanings inferred by the learner
- Press the ‘Auto-Learn’ button, to run a cycle of 100 learning examples
- Word feature structures are learnt by generalising learning examples
- After each cycle, the model lists the words learnt
- Show the learnt feature structure for any word
- Replay the process used to learn it from examples



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)

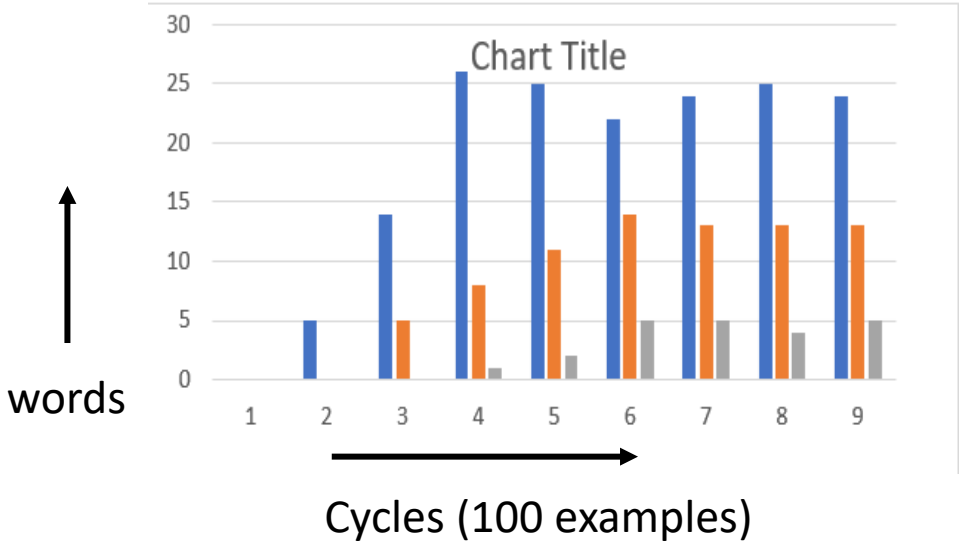
Slot: speed Value: @fast

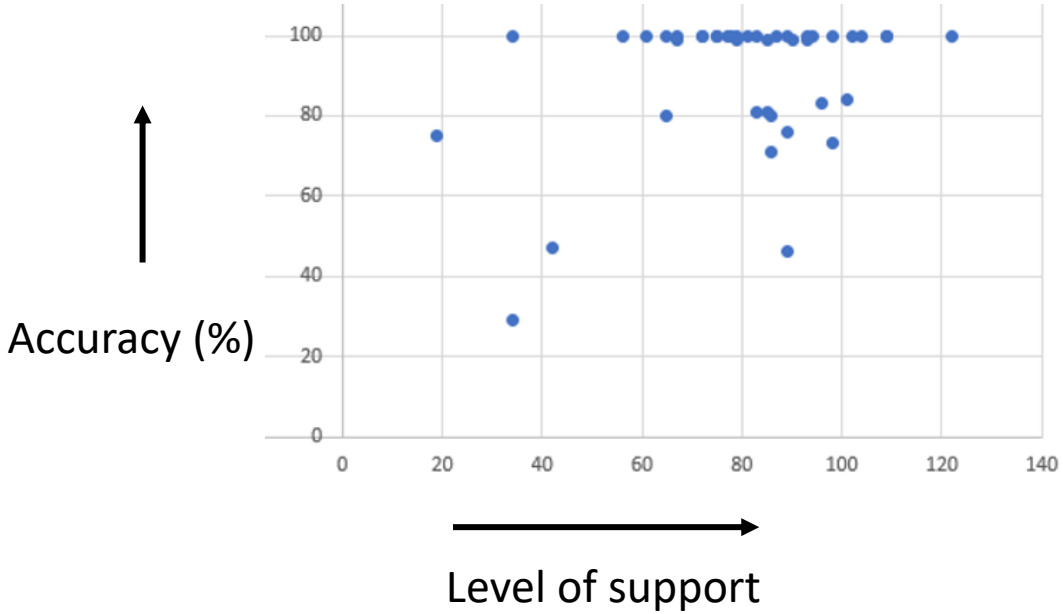
the speed at which some action (ty place

Values:
@slow, @fast,

Performance of the Model

- Learns 40 words from 800 learning examples (as children do)
- Word meanings are accurate (most words agree 100% with the words used to make the learning examples)
- The model learns productive syntax
- Fast execution (runs in a few seconds)
- Robust - no fine tuning of parameters
- Will work for any language (not yet tested)





Learning Word Segmentation

- Learning starts before infants understand meanings
- Statistical properties of sounds are used to learn word boundaries:
  - **Phase 1**: a sequence of sounds ‘biskit’ is a word if there is no partition of its sounds (e.g. ‘bis’ , ‘kit’) which would have made its occurrences by coincidence (this criterion finds the more common words)
  - **Phase 2**: Use common words to partition learning examples. The remaining sounds are words, if they cannot be partitioned further.
- This works well for the 40-word sample of English
- This may not be the best or only model for learning word segmentation; but it shows that there is enough information in the learning signal to learn word boundaries (even without using intonation)

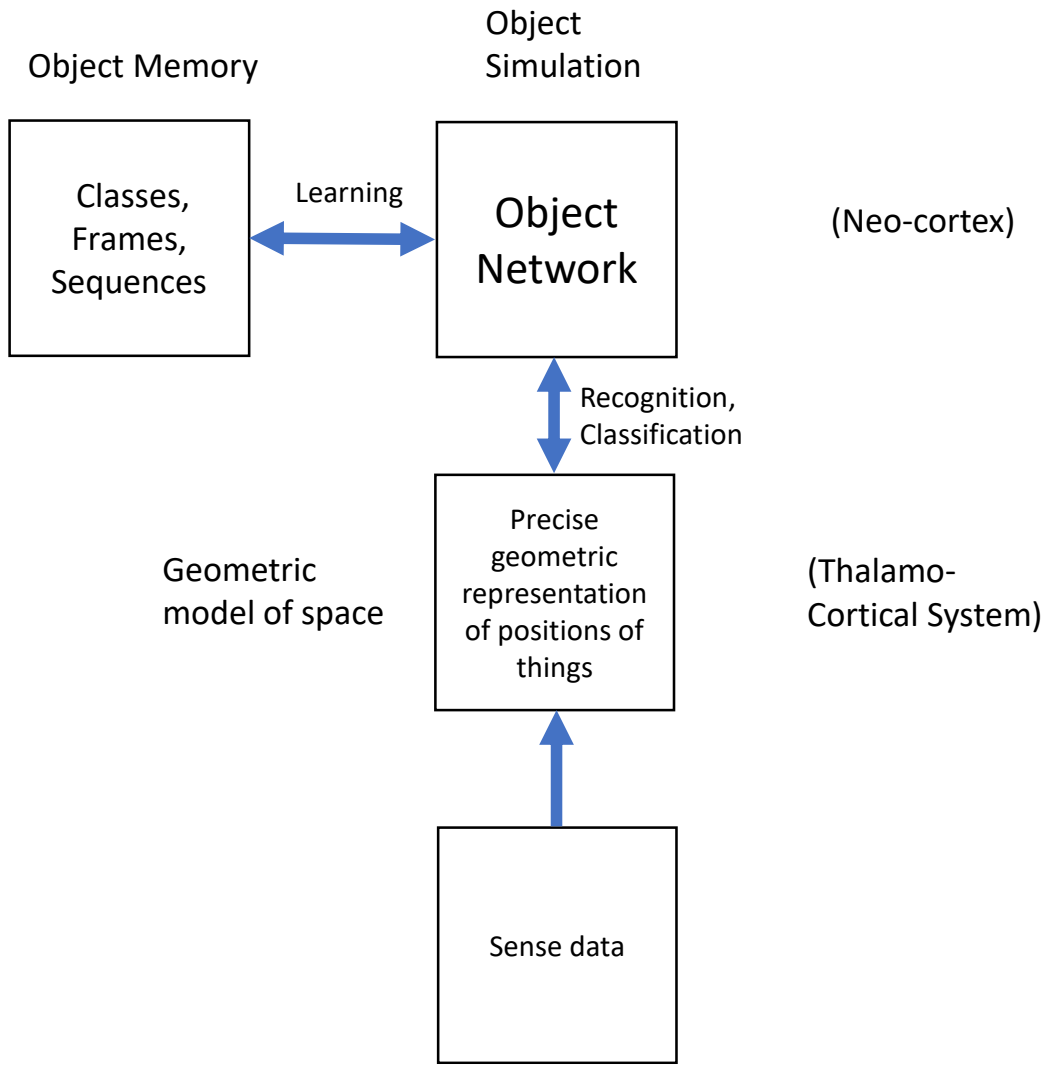
Other Computational Models of Language Learning

- There are very few computational models which can learn all aspects of a language from a standing start. Two such models are:
- Beekhuizen et al. [2015, 2017]:
  - Constructions are like feature structures
  - Understanding uses operations similar to unification
  - Learning uses operations similar to generalisation
  - Predicts ‘verb island’ learning (like this model)
- Abend et al. [2017, 2019]:
  - Learns a probabilistic Combinatorial Categorial Grammar (CCG)
  - 3 sets of probability parameters: SYNTAX, MEANING, and WORDS
  - Bayesian optimisation of all parameters
  - Predicts fast learning of regular grammar (?)

(2) Foundations of the Model

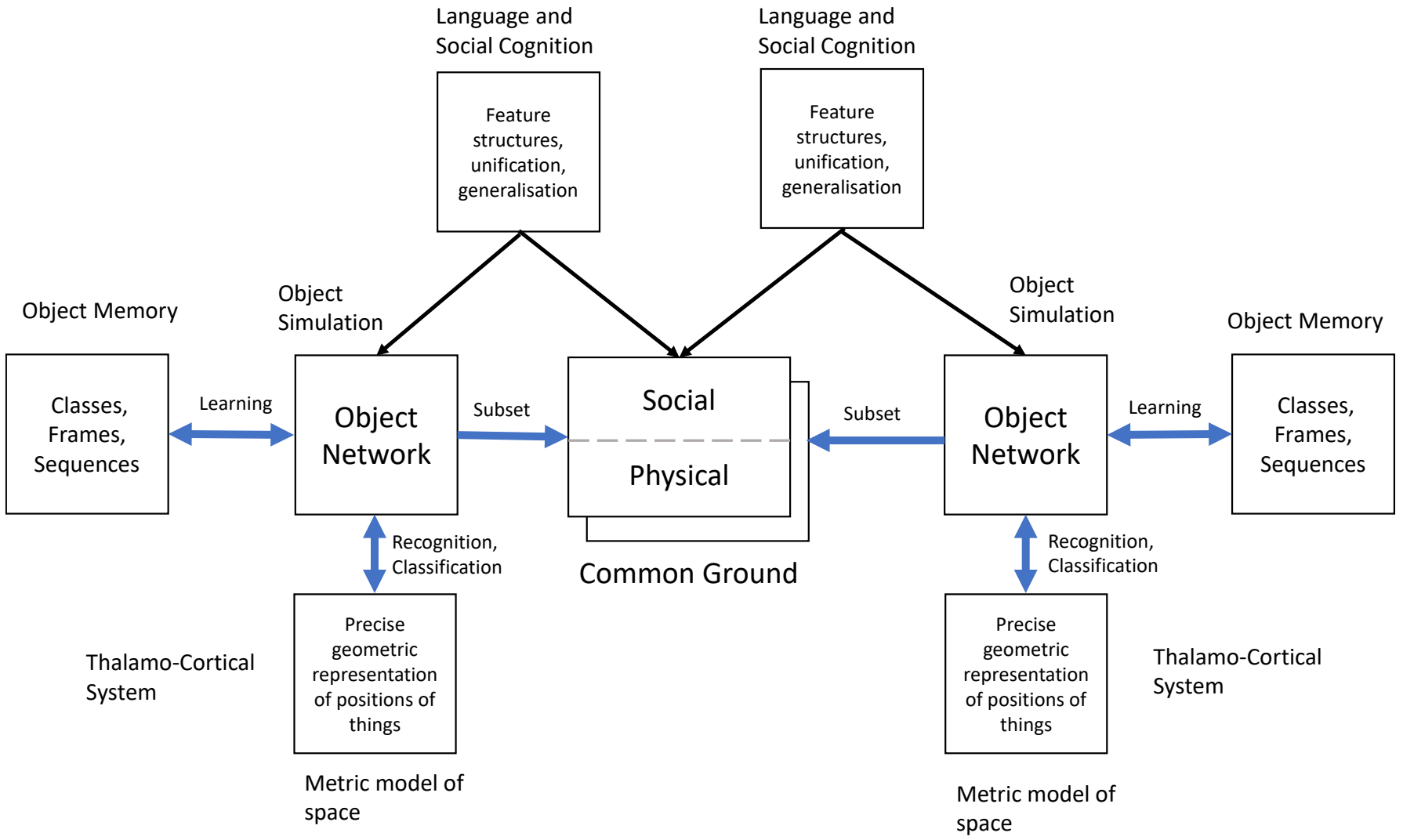
Architecture of the primate brain

- Geometric model of space: from sense data, precise geometry, may reside in thalamo-cortical system
- Object network: Used for understanding, simulation and planning actions (what happens next?)
- Object memory: learnt classes of objects with their behaviour
- Like the ‘Simulation Semantics’ of Embodied Cognitive Grammar



‘Common Ground’ in Language

Two people in conversation:



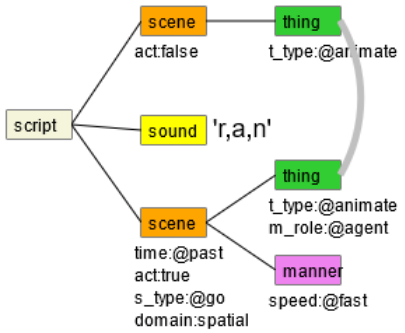
Cognitive Linguistics

- From the 1970s, many authors developed non-transformational approaches to parsing: **Unification–Based Grammars** (Gazdar, Kaplan & Bresnan, Pollard & Sag...). Many of these were built as computer models.
- At the same time, several authors developed **cognitive models of language semantics** (Croft, Fillmore, Lakoff, Langacker, Talmy,…)
- These converged to form Cognitive Linguistics:
  - Language overlaps with other cognitive faculties
  - All language consists of **Constructions** (pairings of meaning and sound)
  - There is no separate autonomous syntax layer
  - A key fact is the huge **diversity** of languages (Evans & Levinson 2013)

Constructions = Feature Structures

- Tree-like Feature Structures have a long history in language research
- A construction is a pairing of form and function (Goldberg): sound and meaning

(Subsumption links are curved lines, or arrows). They are the key to productive language.



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)

Tree notation (this model)

(equivalent)

8. A V- ("V minus") phrase, a phrase of the type **(cat V)(max -)**, consists of a lexical verb together with some or all of its non-subject complements or augments. I say "some or all" because some of them may be present at some distance from the V- constituent, just in case it is in topic or WH-phrase position. A non-maximal verb phrase built around the verb REMOVE, and incorporating all of its local, i.e., non-subject complements, is illustrated in Figure 8.

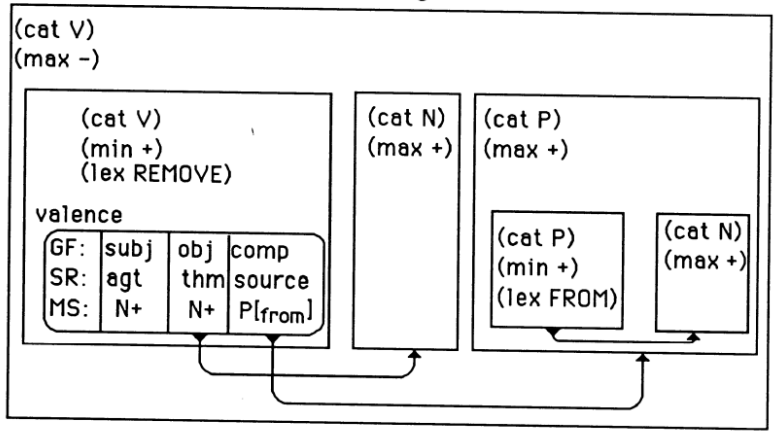


Figure 8

Nested box notation (Fillmore 1988, Langacker, Croft, Kay 2002)

Unification = Bayesian Maximum Likelihood Inference

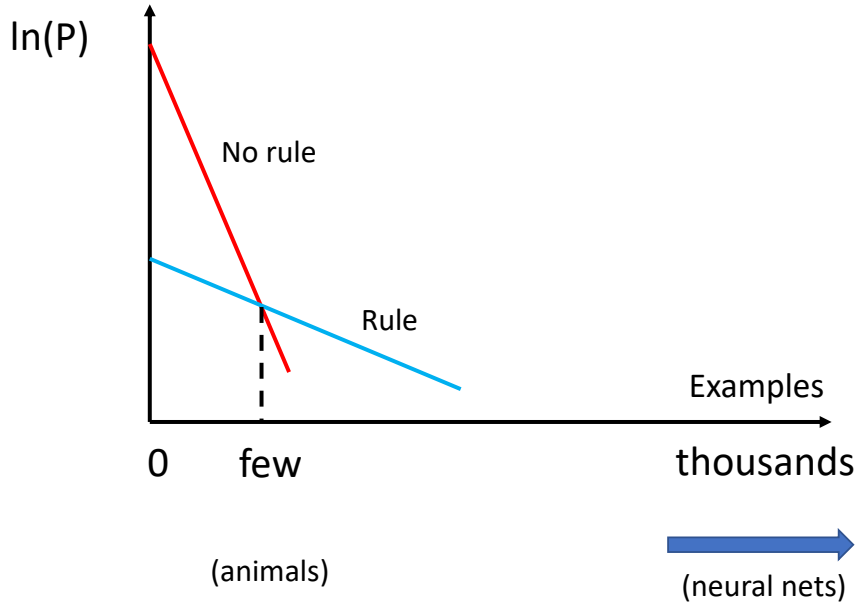
- Any feature structure represents **a set of situations** in the world
  - This is the **model-theoretic semantics** of feature structures (Kay 2002)
  - Tree structures give an unlimited space of meanings
  - A hyper-dimensional conceptual space (Goldberg 2019)
- If a feature structure has I bits of information in its slots, then the probability of the set of situations it represents is approximately 2^{-I}.
- The unification of two features (written as C = A U B) is the smallest feature structure which contains both A and B as sub-structures
- Unification combines A and B by matching nodes, keeping all nodes
- Unification tries to minimise the information content of C (best match between A and B)
- This maximises the likelihood of the set of situations described by C
- Unification is Bayesian Maximum Likelihood Inference (= Minimisation of Free Energy – Friston 2010)

Unification and Generalisation

- The generalisation of two features (written as $D = A \cap B$) is the largest feature structure which both A and B contain as sub-structures
- Generalisation combines A and B by matching nodes, throwing away nodes that do not match
- Generalisation tries to maximise the information content of D – the largest commonality between A and B
- Generalisation and unification are complementary operations
- This leads to many algebraic relations between them (like set theory)
- This underpins the coherence of the learning model

The Bayesian Theory of Learning

- Animals learn rapidly from noisy data – how do they do it?
- It depends on an optimal Bayesian Theory of Learning
- This theory defines the minimum number of training examples you need to learn any regularity (Anderson 1990; Worden 1995)
- The number of examples required to learn is usually very small
- Learning is robust against ‘noise’; it needs only a statistical correlation to learn.
- For learning feature structures, generalisation finds the most predictive rule possible



Zero Intercept = Logs of Bayesian priors.

Height of lines = Logs of Bayesian posterior probabilities.

Learning strength = difference between the two slopes; it depends on the strength of the correlation between events.

Evolution of the Capacity for Language

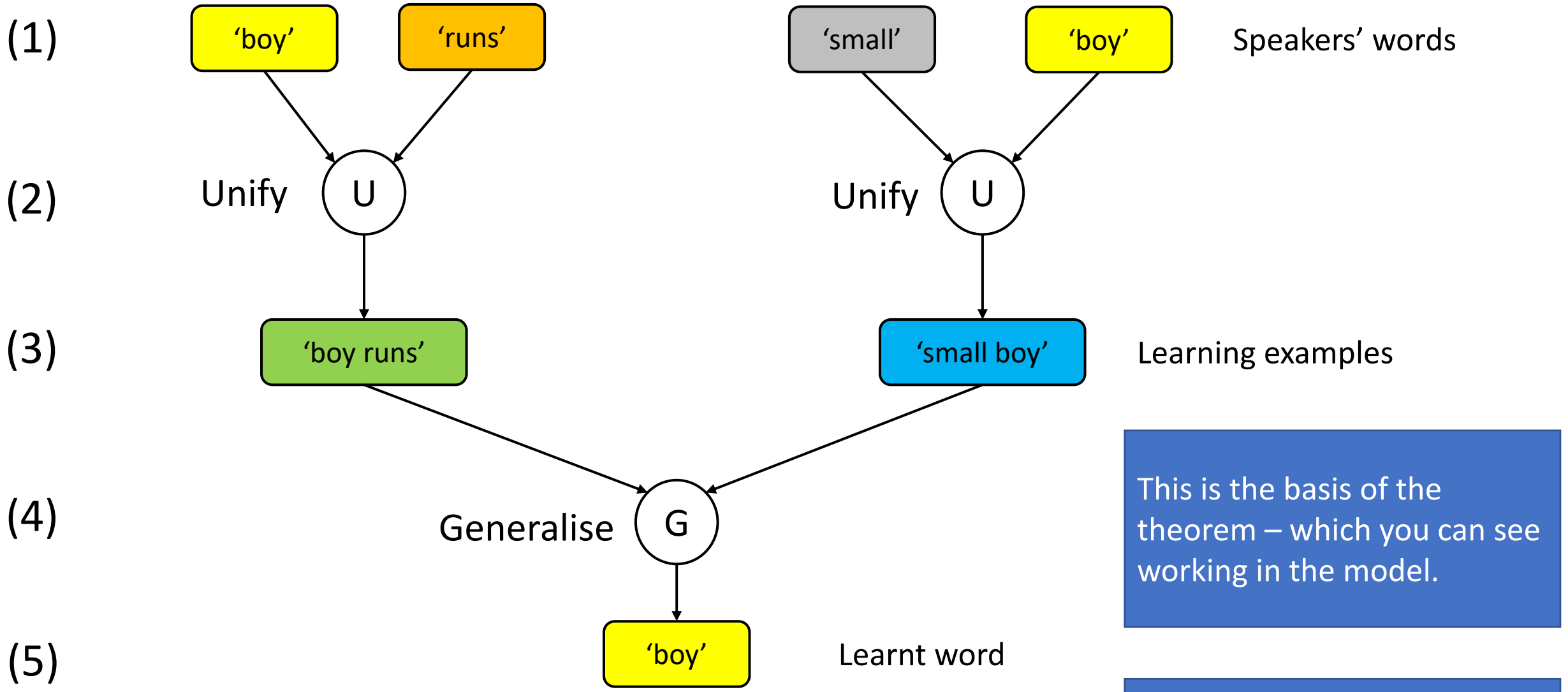
- Primates have tree-like feature structures:
  - To simulate what happens next
  - To plan complex physical actions
  - For social intelligence (Worden 1996)
- Primates need to have the operations on feature structures used for language:
  - Unification - for Bayesian maximum likelihood inference, to simulate events
  - Generalisation – for Bayesian fast learning
- So why do other primates not have language? Why only us?
- Language is massively **over-engineered** for any survival purpose; most primates do not need it.
  - Very fast
  - Huge vocabulary
  - Unbounded productivity
- Mankind went through a period of intense **sexual selection** for
  - General intelligence (Miller 2000)
  - Language ability (Worden 2017)
- We need to converse (better than other people) to get a mate
- Language is the **Peacock's Tail** of the human brain

(3) A Theorem of Language Learning

- Unification and generalisation form a simple mathematical structure, like set theory. In this structure, we can prove:

Through learning and use, any construction will replicate faithfully through the generations.

- This result is:
  - Universal - it applies to any set of constructions (meaning-sound pairs)
  - Applicable to any language
  - Underpins the durability, diversity, and power of languages (Evans & Levinson 2013)
- Analogy with DNA replication:
  - Replication of DNA is a faithful chemical process
  - Universal: it works for any sequence of base pairs
  - Underpins evolution and the diversity of all life
- The evolution of constructions (such as words) is a model of language change
- Regular grammar arises from word evolution



This is the basis of the theorem – which you can see working in the model.

Unification and generalisation complement one another.

Consequences of the Theorem

- Languages are **not constrained** by the theorem – they can contain any possible construction:
  - This underpins the remarkable **diversity** of the world's languages (Evans & Levinson 2013)
- It shows that the feature structure model **can apply to any language**:
  - Children can learn any construction at all without limit, initially in its unproductive form
  - Then they incrementally learn it with greater productivity
- Constructions are **accurately preserved** by replication:
  - Constructions can have consistent definitions across a speaking community – to serve their communicative purpose
  - Constructions can be refined over many generations, to better serve the needs of society
  - Constructions are subject to strong selection pressures, which lead to language change
  - Constructions, not languages, are the 'species' of language evolution

Evolution of Constructions, as Language Change

- The idea that language evolves is as old as the idea of biological evolution itself (Darwin)
- Today, it is very influential (e.g. Christiansen & Chater 2016; see Dediu et al. 2013 for an authoritative review.)
- Evolution is the dominant metaphor for historic language change
- In most discussions, ‘a language’ is taken as the evolving species
- The ‘language as species’ metaphor is often an unnoticed background assumption.
- In this model, the evolving species is not a language. Each construction is an evolving species. A language is like an ecology.
- This alters the terms of the discussion:
  - e.g. what would be the ‘design space’ for an ecology? (Dediu et al 2013)

Selection Pressures on Constructions

- Each construction (e.g. each word) is subject to strong selection pressures – often from other constructions in its language ecology.
- Each construction must get used, in order to be heard and to reproduce:
  - It must have a useful meaning, which is not better expressed by other constructions
  - It must be brief, while not creating intolerable ambiguities
  - If it can combine productively with other constructions, that will greatly increase its range of use (= fitness)
  - If different variants of the construction are related systematically, following other constructions (e.g. tenses of a regular verb) it is easier to learn.
- These selection pressures give a simple account of many prominent features of languages:
  - Productivity of many constructions
  - Partial semantic regularity (e.g. alignment of S/V verbs, spatial terms -Talmy 2000, Bowerman & Choi 2003)
  - Partial syntactic regularity
  - Language universals (Greenberg 1963, Hawkins 1994, Worden 2002)
  - Language universals are not universal (Evans & Levinson 2013)