Rules vs. example: past tense debate and beyond
LING 611 Spring 2022

Brian Dillon
Shota Momma

University of Massachusetts, Amherst
Department of Linguistics

2/14/2021

Big issue

Something has to be stored in the long term memory.

-> Minimally, idiosyncratic information of atomic units must be stored (“lexicon”).

We can always create novel expressions.

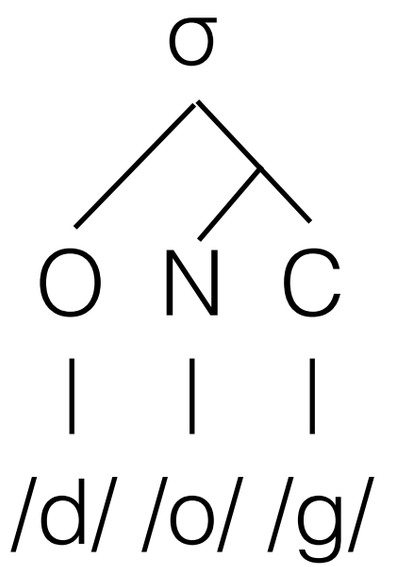
-> Long-term memory is not sufficient

-> Some sort of combinatorial mechanism has to be involved.

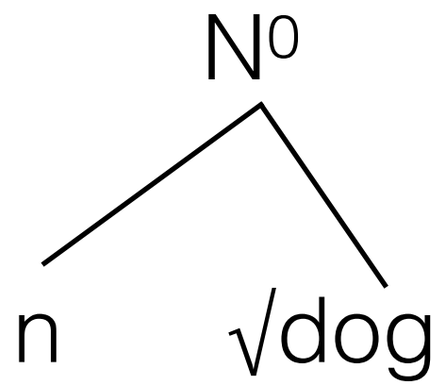
The relative contributions of memory and combinatorial mechanisms in language processing

Complex units in language

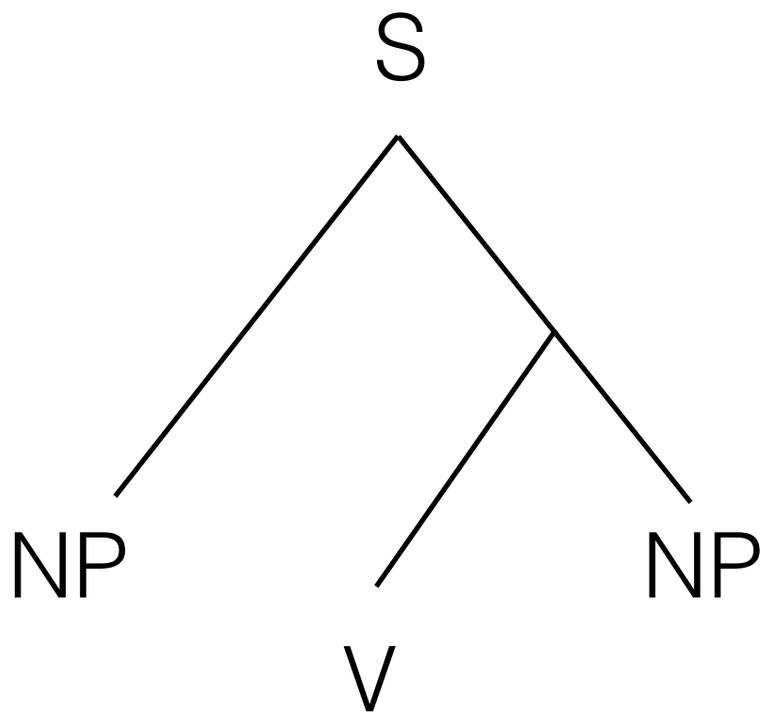
Phonology



Morphology



Syntax



The dog chases the cat.

What gets retrieved from LTM, and what gets constructed on the fly?

Past tense as a model species

Relatively “simple” system that exhibits regularity (‘walked’), irregularity (‘ate’) and semi-irregularity ('drank’).

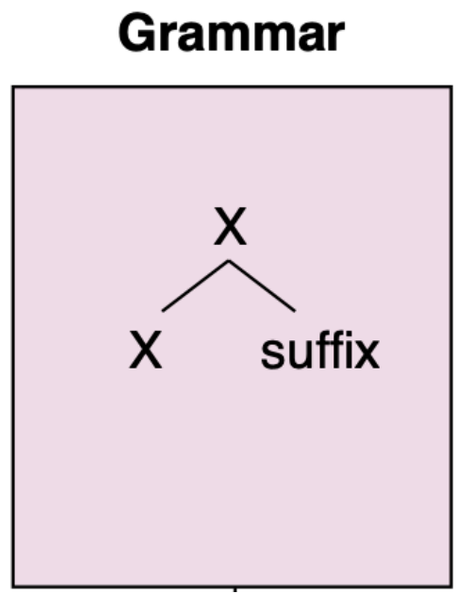
Ideal for studying the role of storage and online computations in general.

Drosophila

Symbolic system

Fundamentally **algebraic**: the ability to represent *variables*

Very good at capturing regular past tense

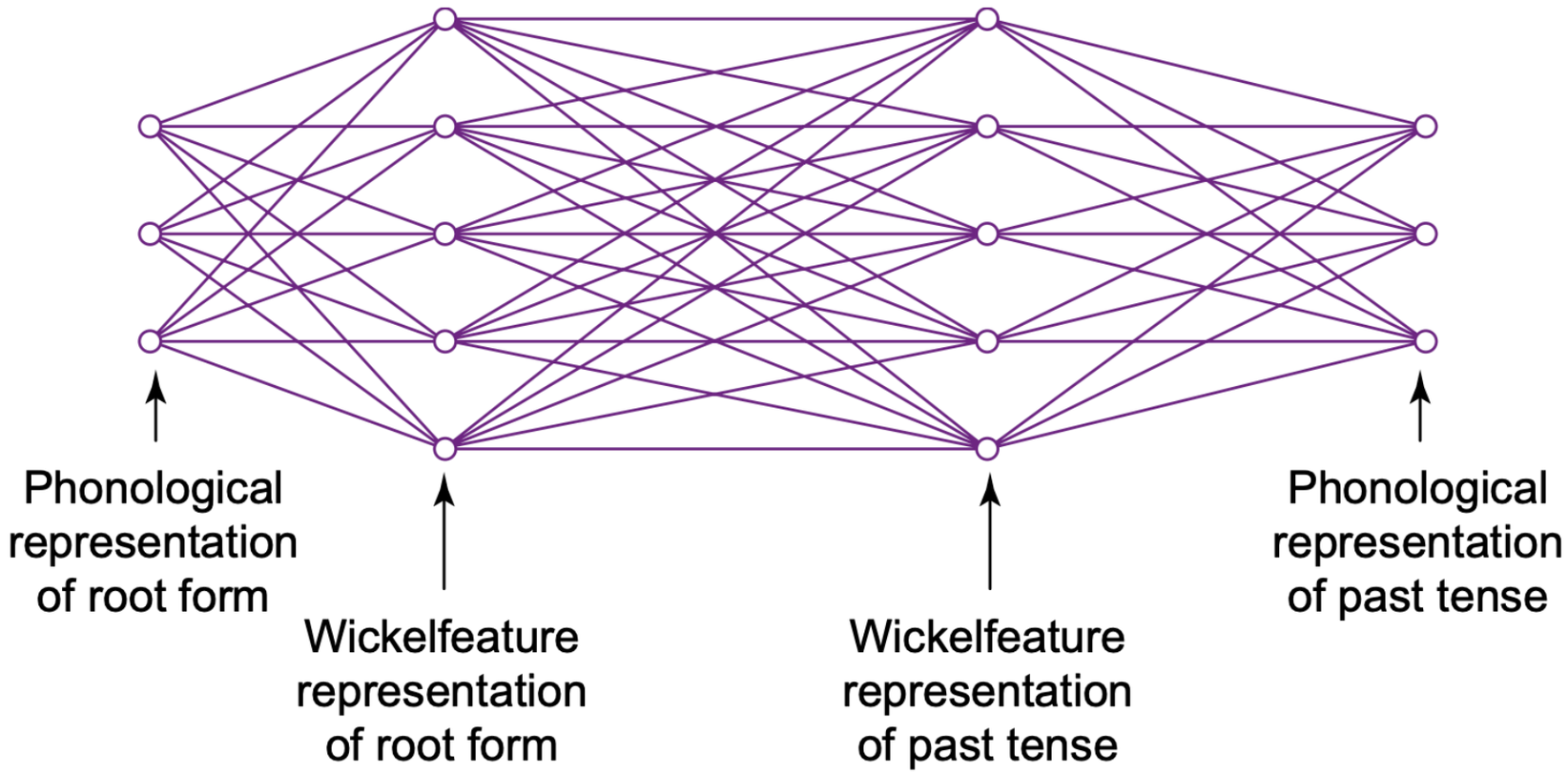


May capture some irregular patterns too.

- ring-rang, sink-sang, etc.
- With abstract forms (that never surface), empirical scope can be expanded (rin-run; sing-sang)

Associative network (Parallel Distributed Processing)

"We suggest that lawful behavior and judgments may be produced by a mechanism in which there is no explicit representation of the rule"



Wickelfeature: basically like a tri-phones but represented in a *distributed* fashion

Basics: AND

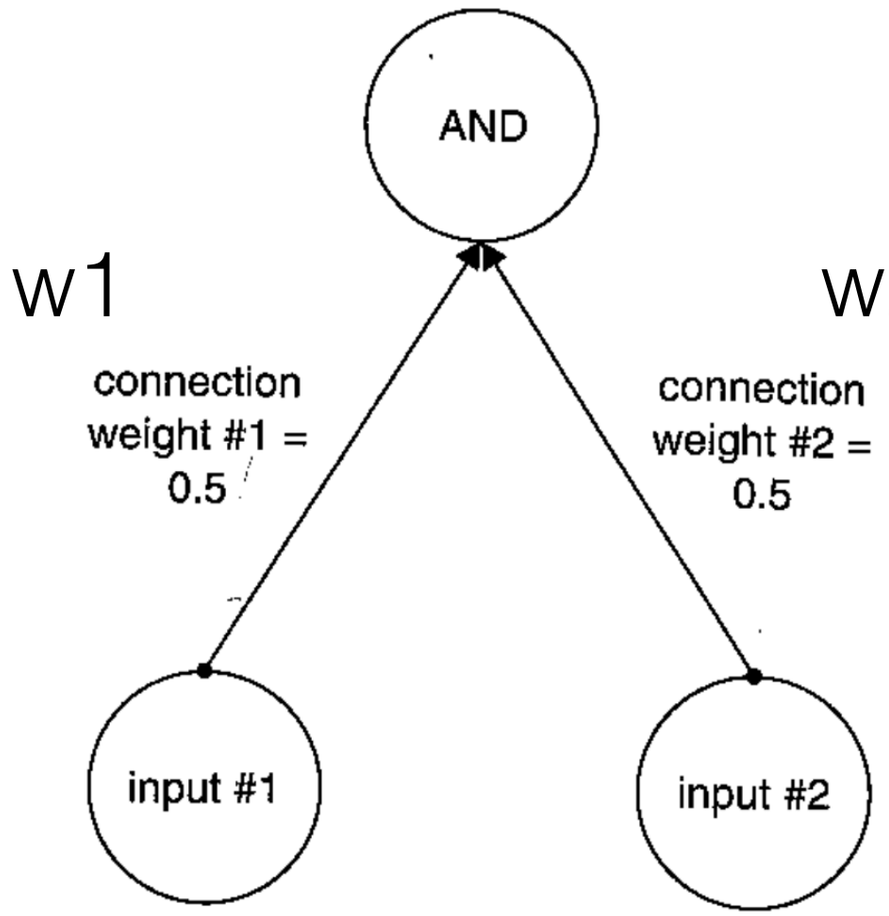
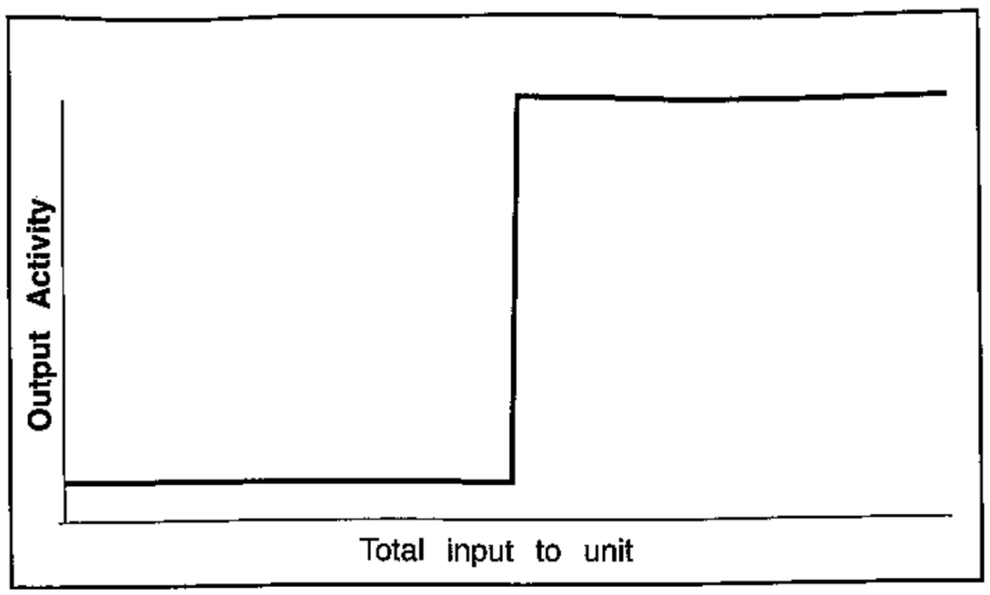


Figure 2.5
A two-layer perceptron that computes the function AND.

Activation of the output node:
f(x1*w1 + x2*w2 + bias)

f: Activation function



Basics: OR

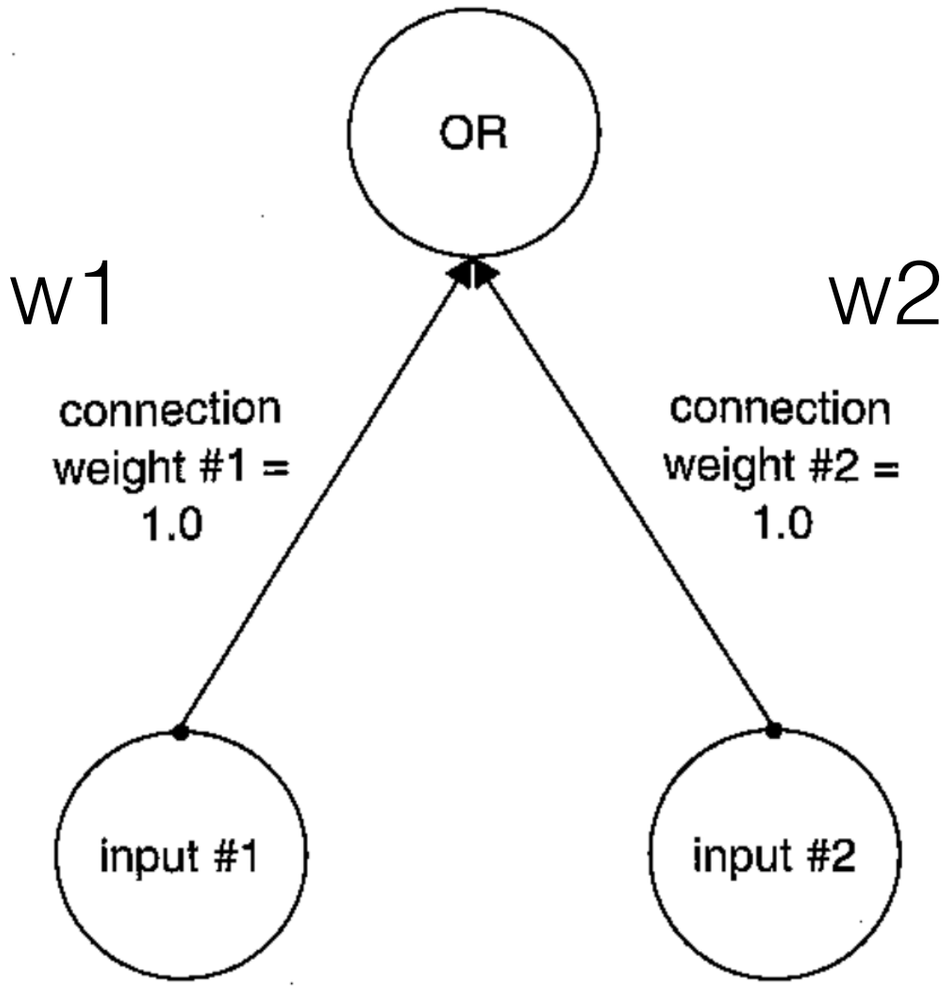
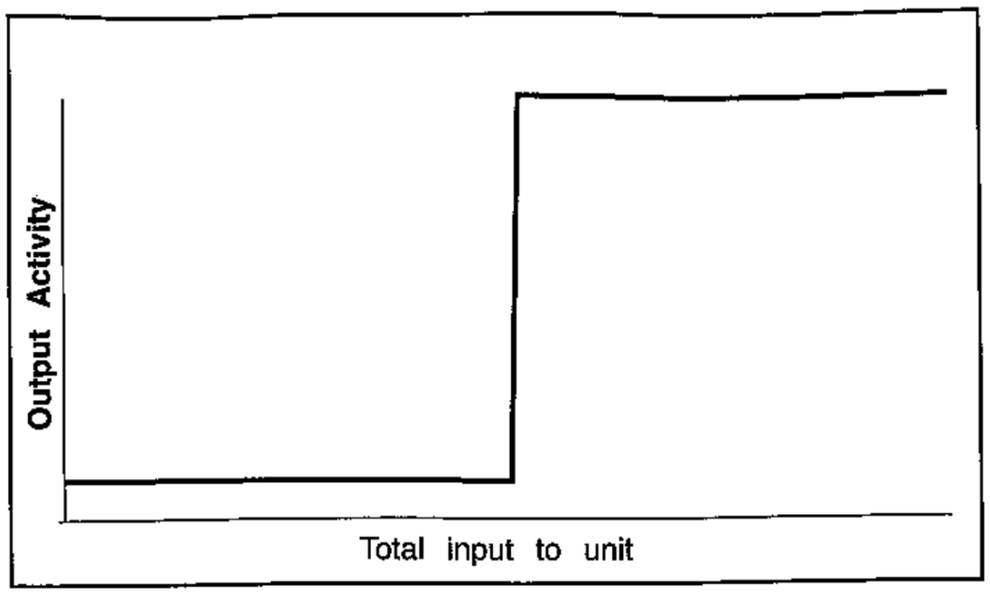


Figure 2.4
A two-layer perceptron that computes the function OR.

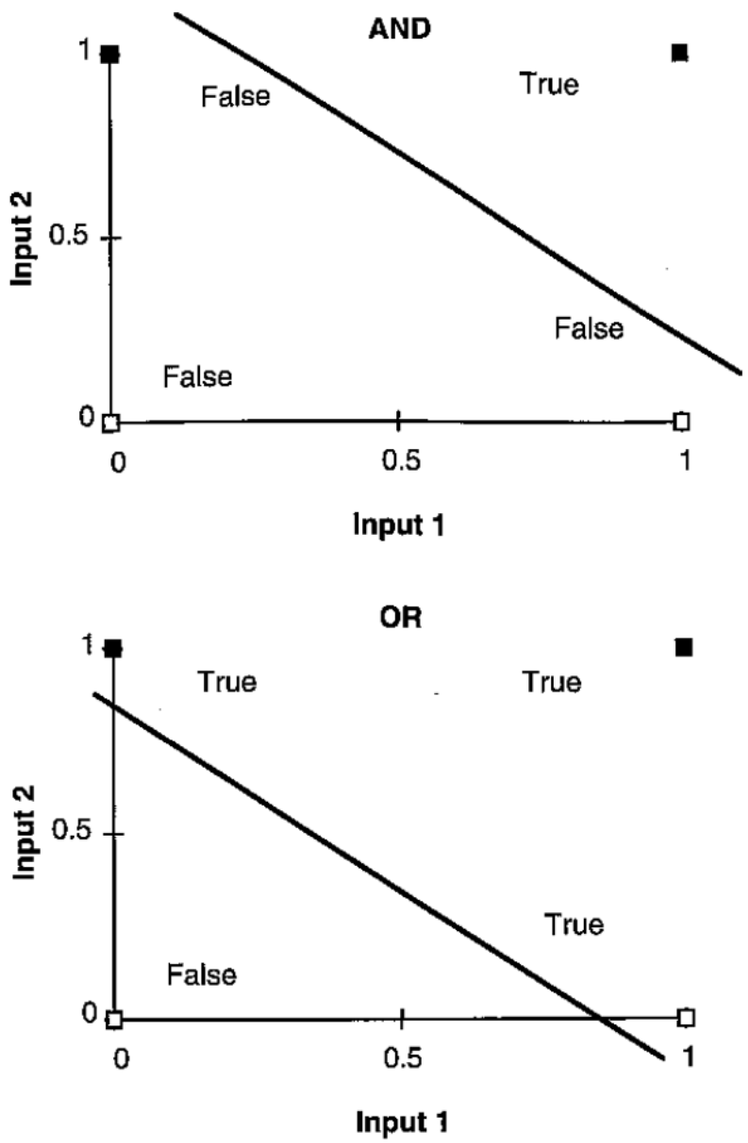
Activation of the output node:
f(x1*w1 + x2*w2 + bias)

f: Activation function



Basics: Linearly separability

Linearly separable



Linearly non-separable

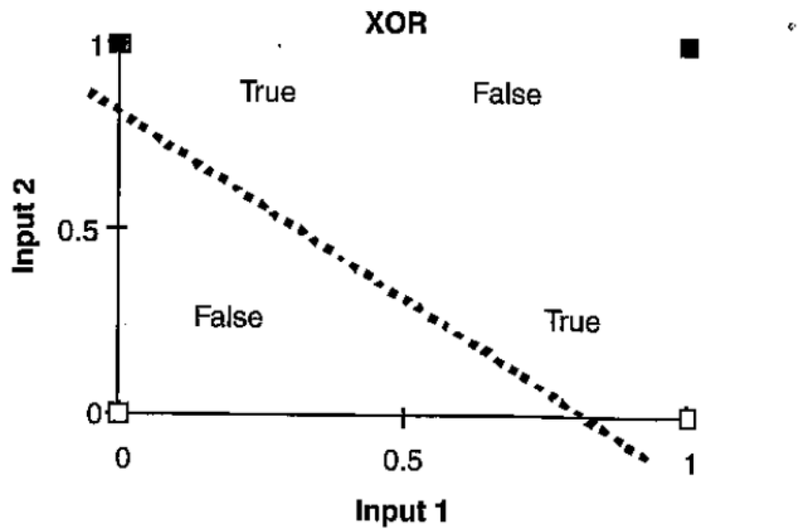
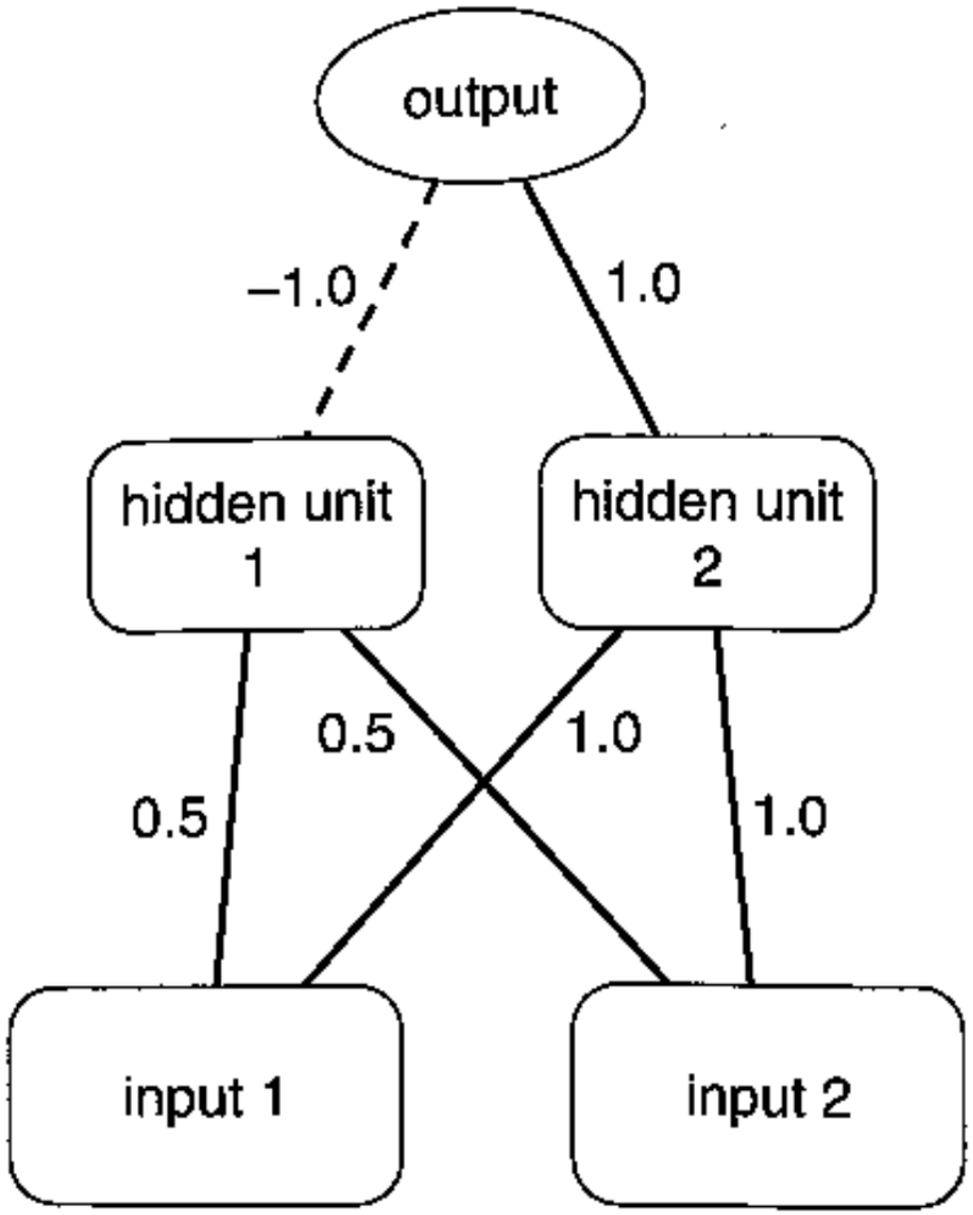


Figure 2.7
The exclusive or (XOR) function. No straight line can separate inputs that yield *true* from the inputs that yield *false*.

There is no way to do XOR classification (or linearly non-separable) with two-layer networks.

Basics: Adding hidden layer



Activation values of units in an exclusive or (XOR) network (see figure 2.8).

Input 1	Input 2	Input to hidden unit 1	Output from hidden unit 1	Input to hidden unit 2	Output from hidden unit 2	Input to output unit	Output
F = 0	F = 0	0	0	0	0	0	0
F = 0	T = 1	0.5	0	1	1	1	1
T = 1	F = 0	0.5	0	1	1	1	1
T = 1	T = 1	1	1	2	1	0	0

The network can solve XOR problem with a hidden layer.

Basics: Learning algorithms

The delta rule

$$\Delta w_{io} = \eta^* (target_o - observed_o)\, a_{i'}$$

Learning rate
(a parameter set by the modeler)

Error

Activation value of the input node
(bigger weight changes if the input node is more active)

Basics: Learning algorithms

Delta rule cannot simply be applied to networks with hidden layers, b/c the ‘target’ value for nodes in hidden layers cannot be known.

Backpropagation: The weights between a hidden node and an input node is adjusted proportionally to the ‘blame assignment score’ (instead of the difference between the observed and target activation value, which is unknown).

Backpropagation is a type of *gradient descent* algorithms.

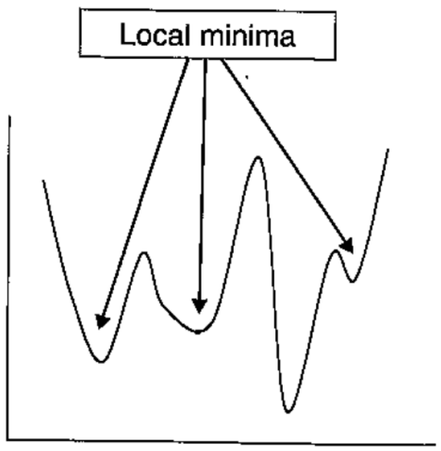
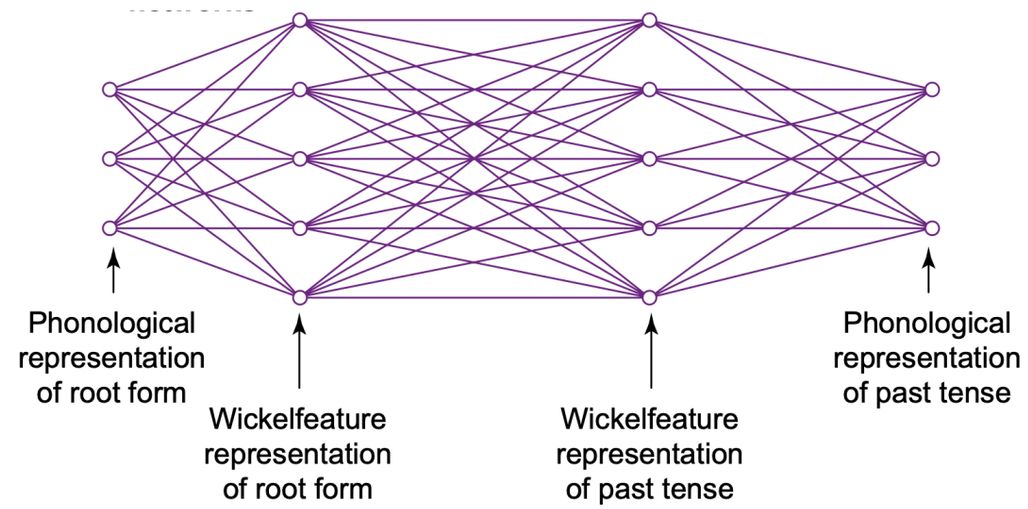


Figure 2.9
The hill-climbing metaphor. Arrows point to locations where error is low and small steps would lead only to greater error.

Associative network

Phonological rep. /kat/

Wickelphones $\#k_a$ $k a_t$ $a t\#$



Wickelfeatures [(000) (00) (000) (00) 1]
[(100) (10) (001) (01) 0] = $\#k_a$
[(001) (01) (010) (01) 0]

Interrupted vs. continuous vs. vowel

Interrupted: stop vs. nasal
Continuous: fricative vs. sonorant
Vowel: high vs. low

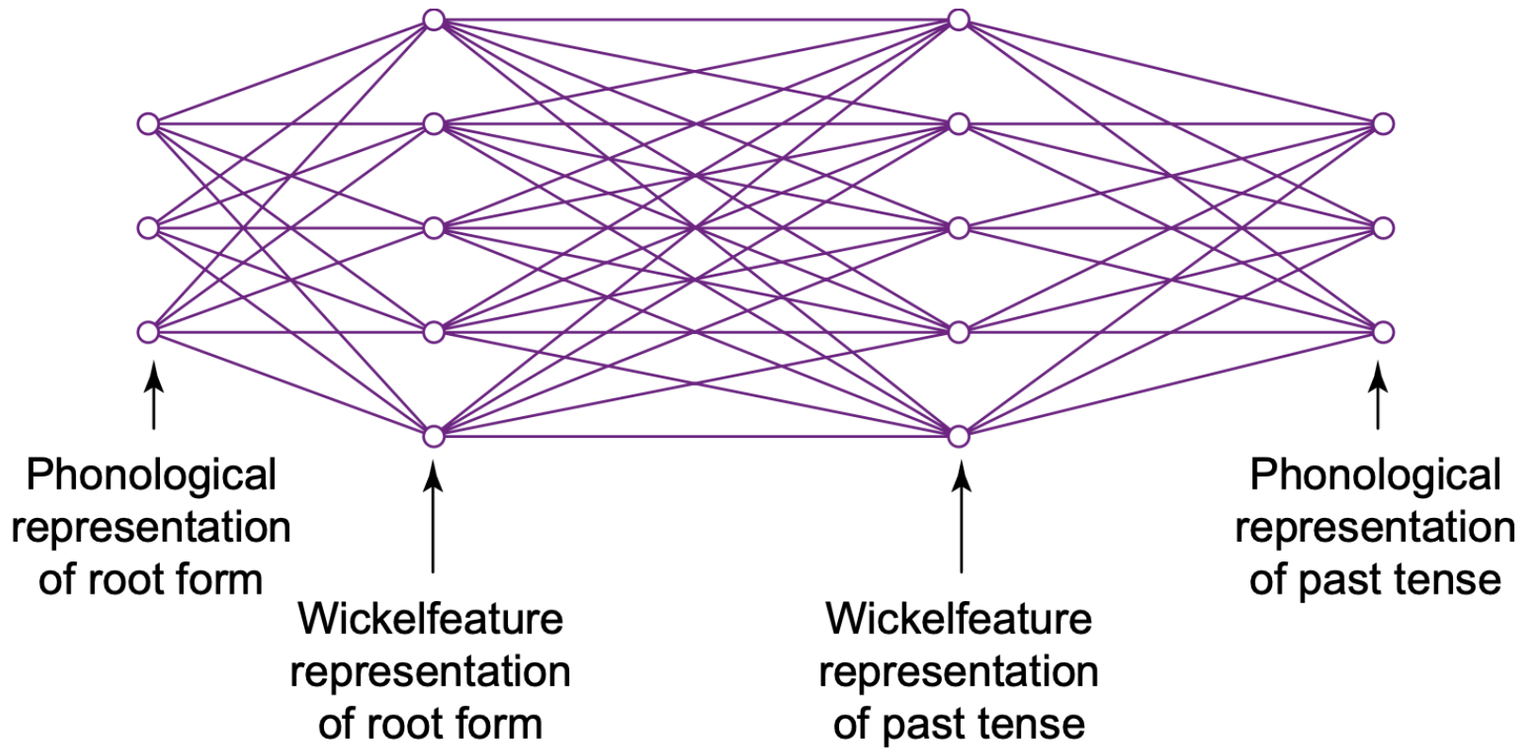
Voiced vs. voiceless
Long vs. short

Front, middle & back

Learning

If the computed activation of a given unit matches the correct value, no learning occurs. If a unit that should be active is not, the weights to that unit from each active input unit receive a small fixed increment, and the threshold is reduced

Associative network



TRENDS in Cognitive Sciences

Associative networks generalize b/c novel items share units and connections used in old items.

Associative network: good

Overregularization:

- Initially great performance on irregular verbs (‘went’)
- Performance drops due to overregularization (‘goed’)
- Performance recovers

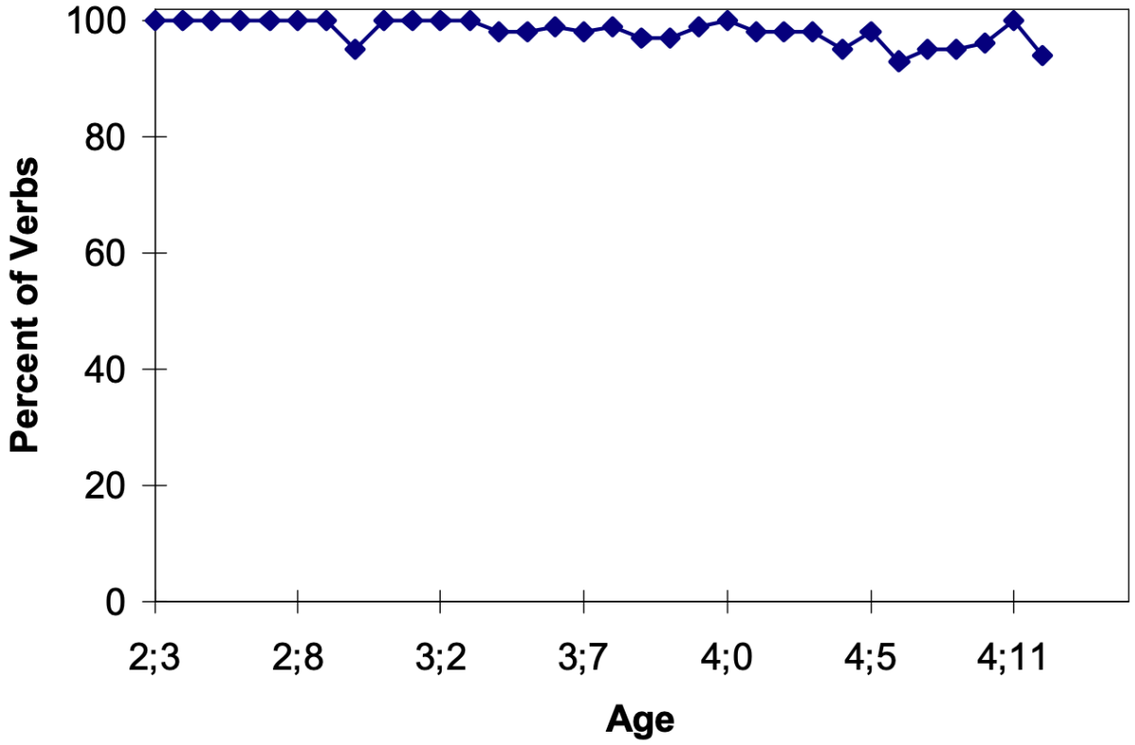
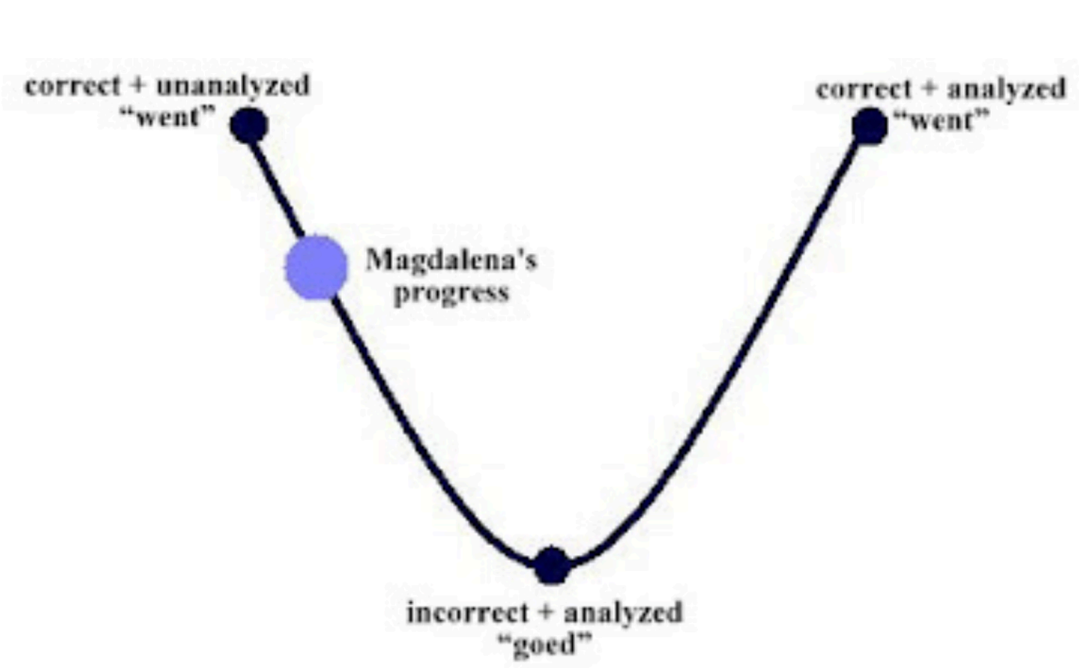


Figure 1: (1-overregularization rate) for Adam (reproduced from Marcus *et al*. (1992)).



Associative network: bad

Unusual errors

- Network errors are not something that humans would produce (mail-membled, trilb-treelilt)

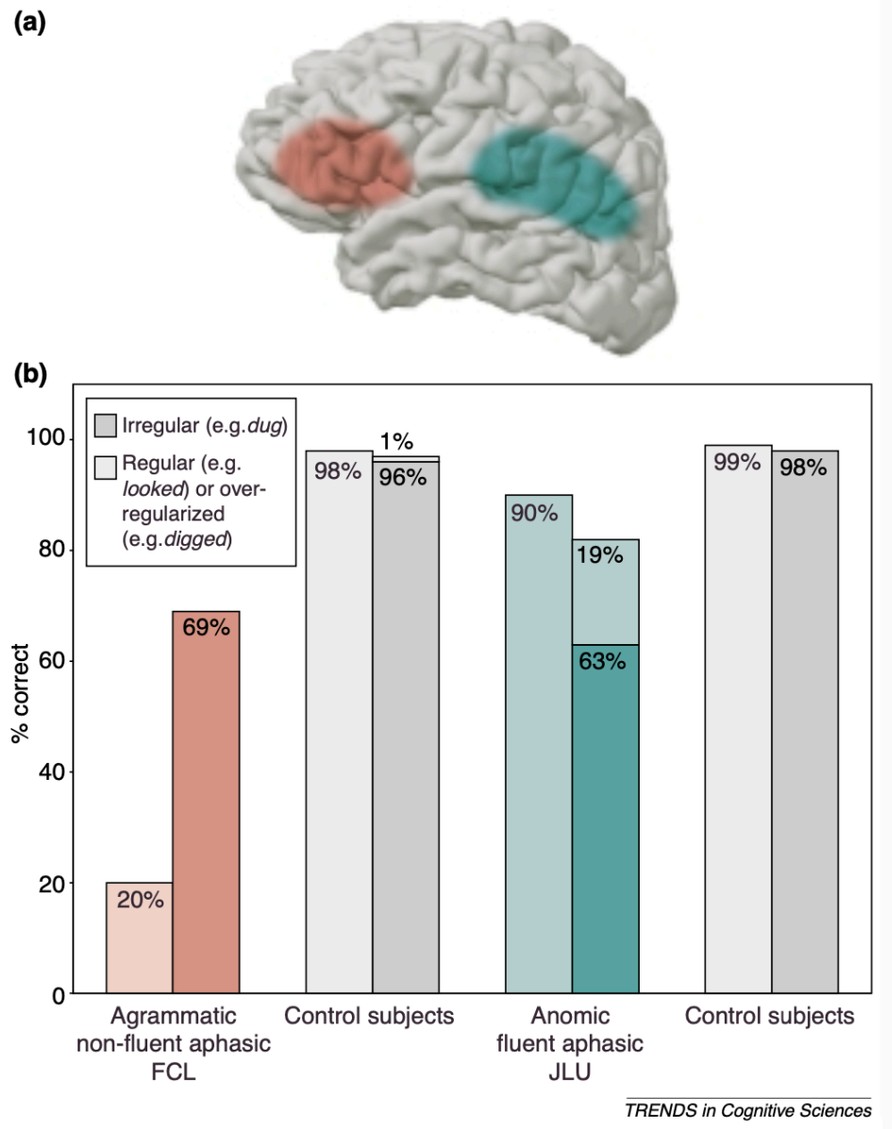
Systematic regularization

- grandstanded (not grandstood), ringed the city (not rang the city)

Associative network: bad

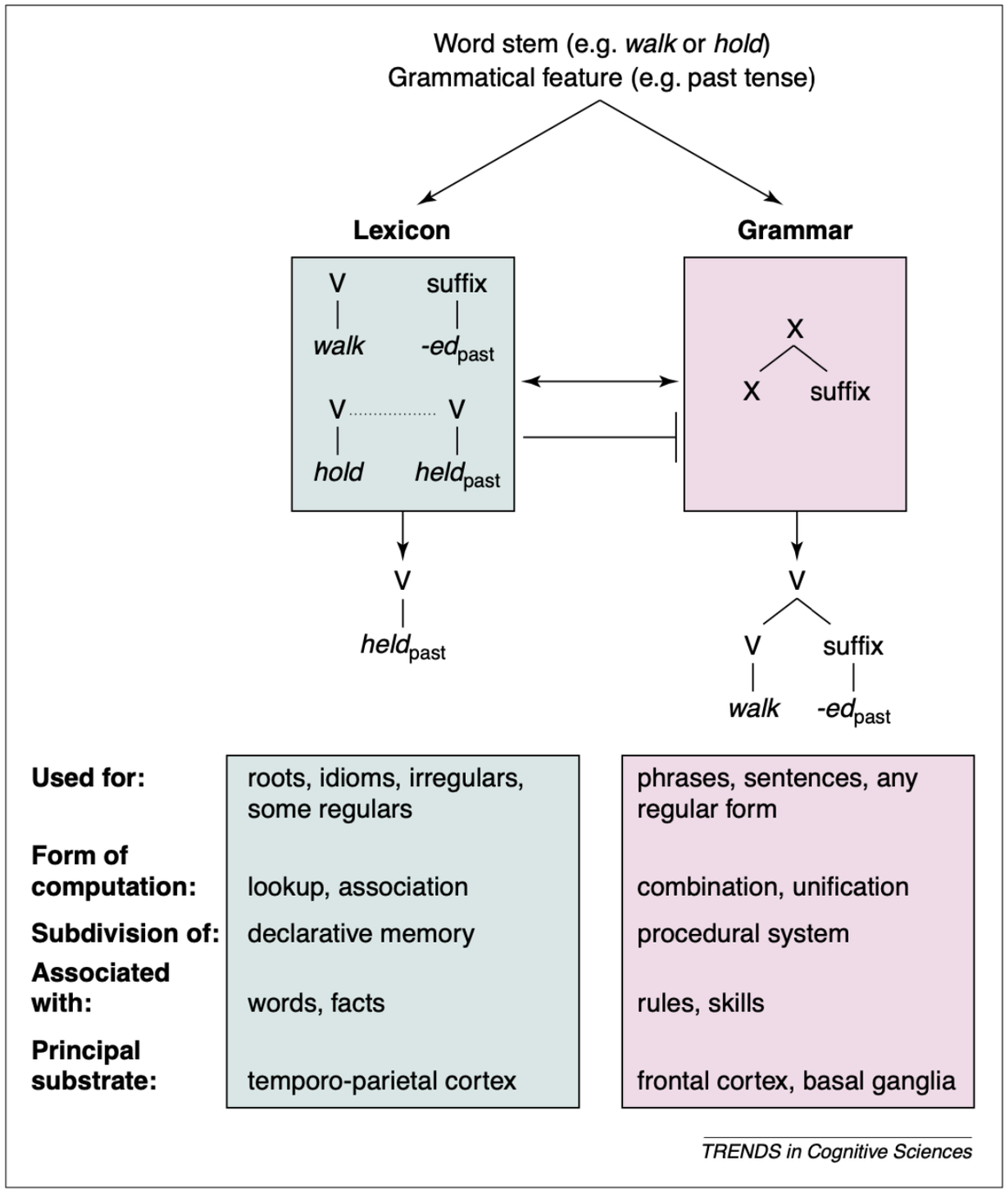
Double dissociation

- Network errors are not something that humans would produce (mail-membled, trilb-treelilt)



Best of both world

Word & Rule (WR) theory

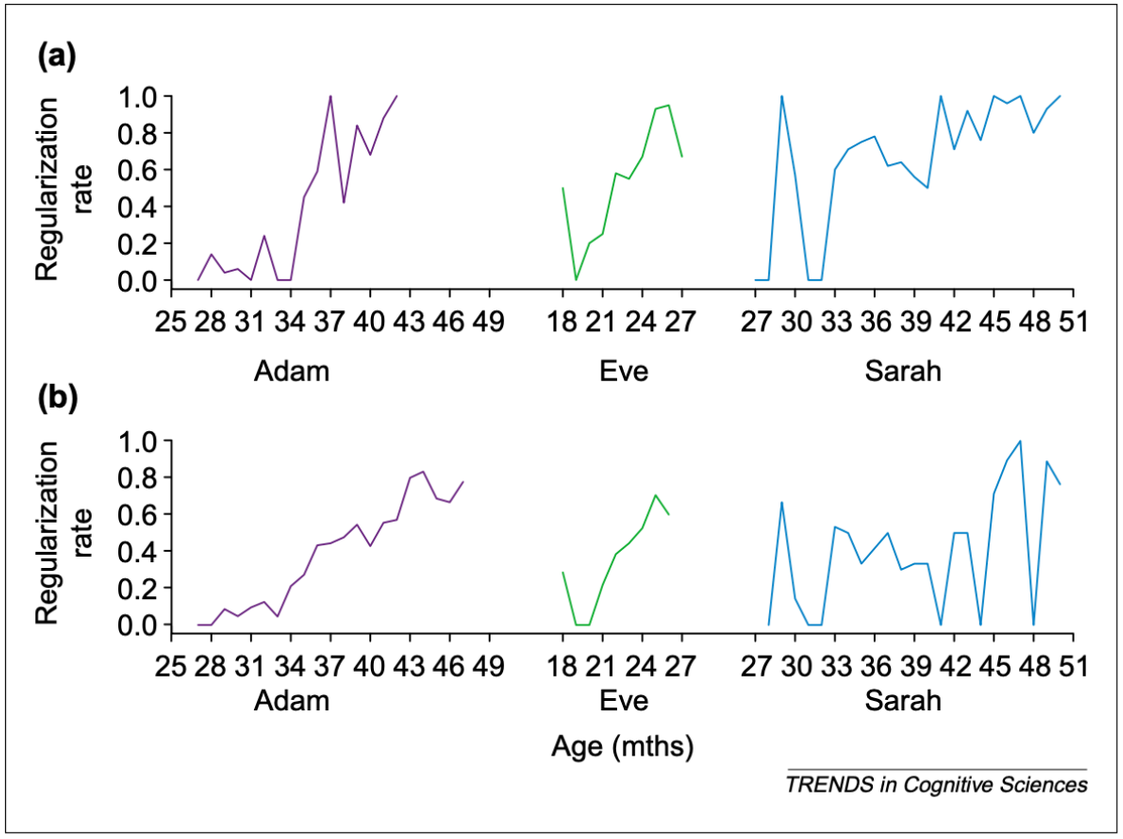


Convinced?

Gradual acquisition?

Table 1. Predicted and observed aspects of regular inflection

Aspect	Prediction from		Observed
	Symbolic Rules	Connectionist Models	
Acquisition	sudden	gradual	gradual
Sensitivity:			
to phonology	no	yes	yes
to semantics	no	yes	yes
in development	no	yes	yes
in German +*s* plural	no	yes	yes
Separability from exceptions:			
Genetically	yes	no	no
Neurologically	yes	no	no

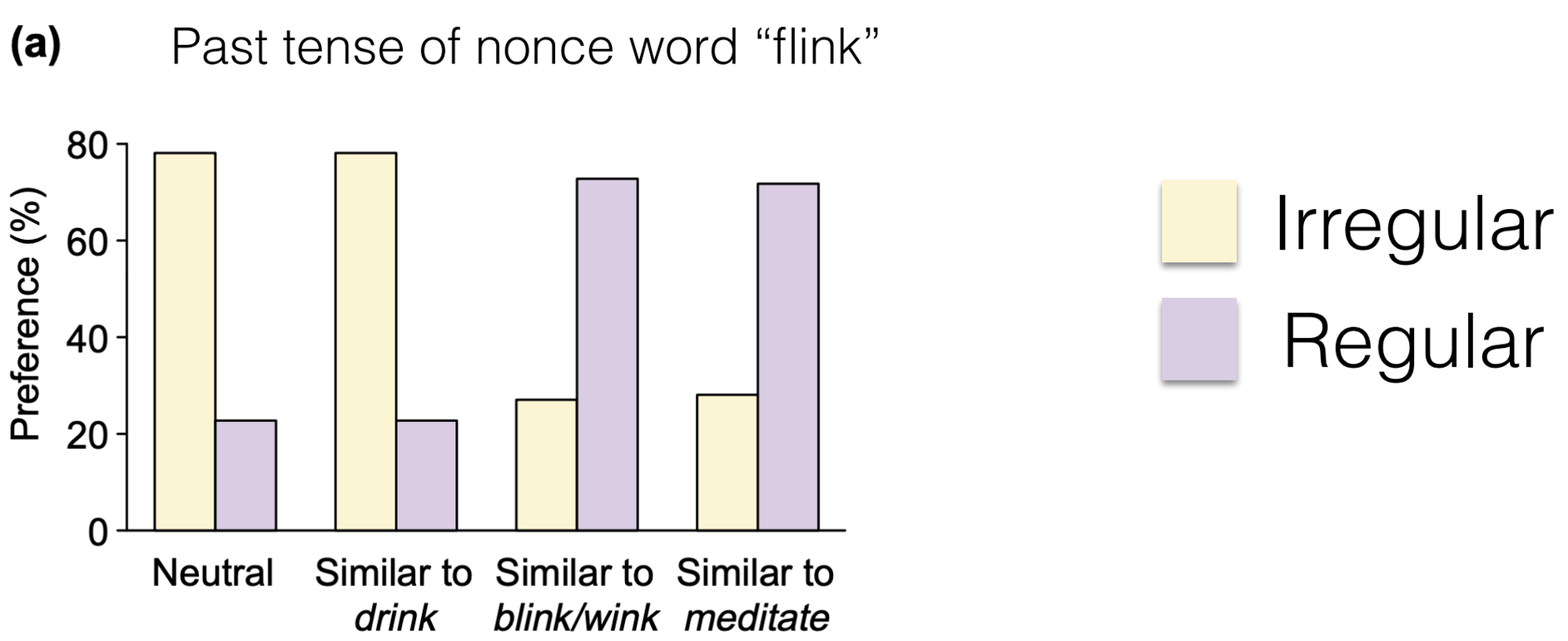


> There is always a considerable period… in which production-when-required is probabilistic. This is a fact that does not accord well with the notion that the acquisition of grammar is a matter of the acquisition of rules, since the rules… either apply or do not apply. One would expect rule acquisition to be sudden. (Ref. [17], p. 257)

Semantic effect on regularization

Systematic regularization
- grandstanded (not grandstood), ringed the city (not rang the city)

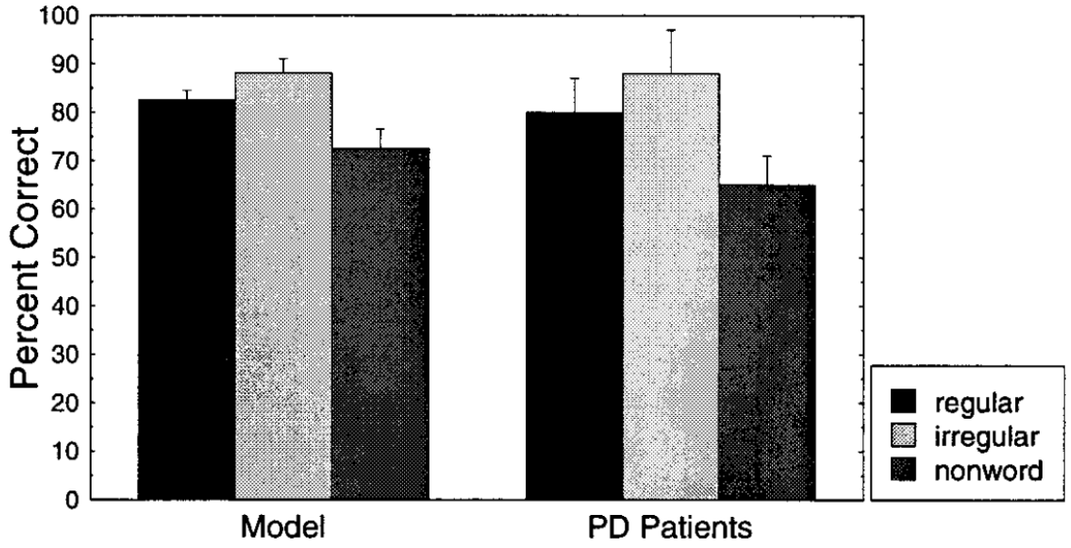
Perhaps this is due to semantics effects: words with different semantic content are classified differently?



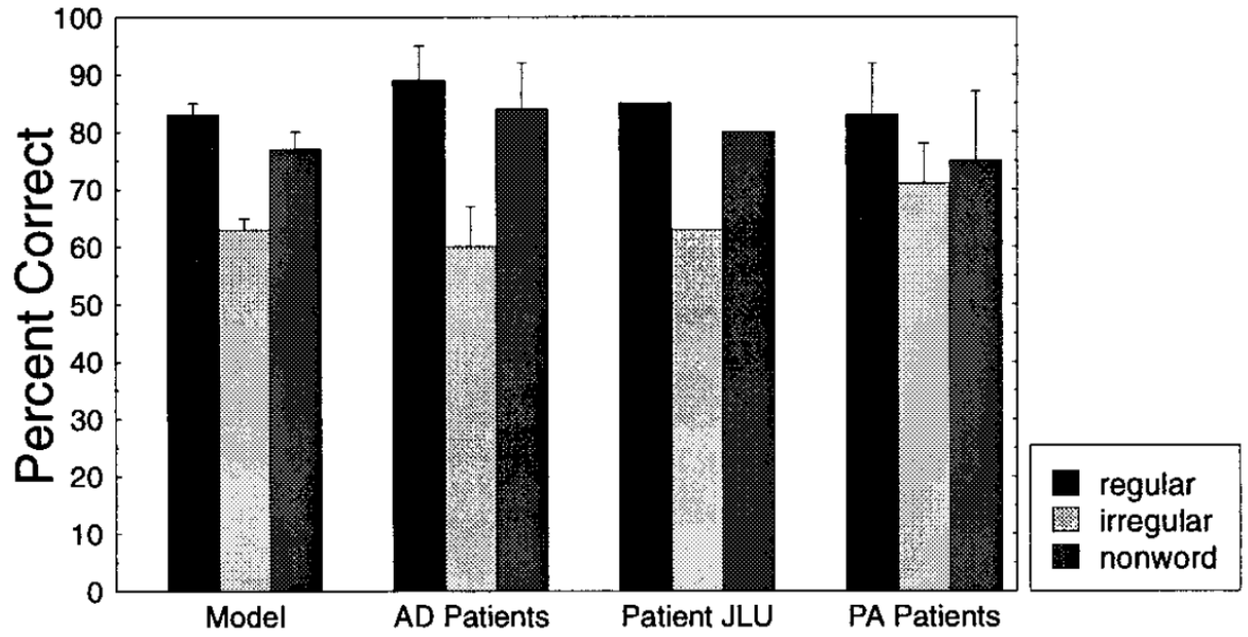
Ramscar (2002)

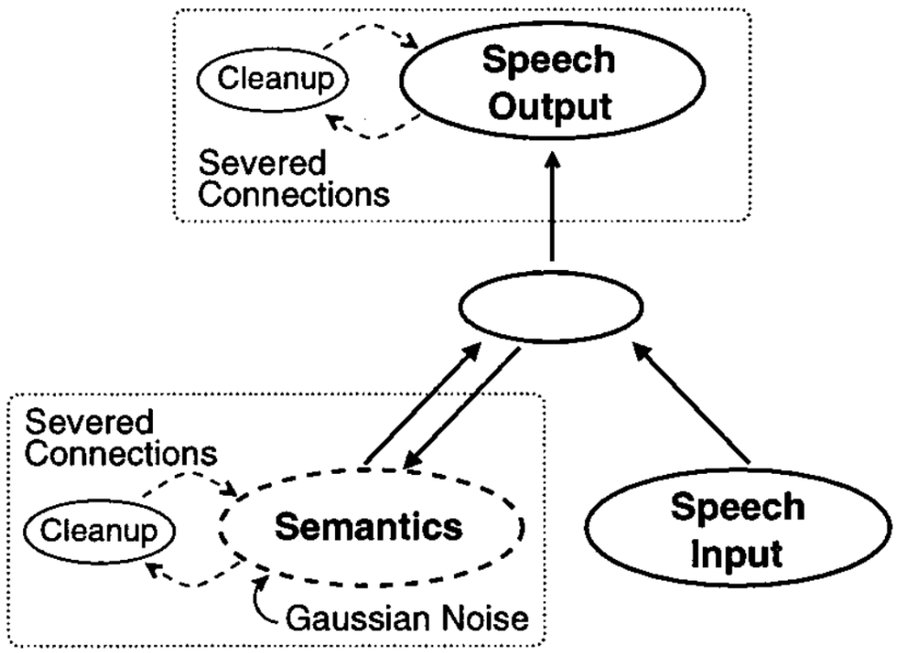
Double dissociation?

Phonological damage
Parkinson's disease



Semantic damage
Alzheimer's disease



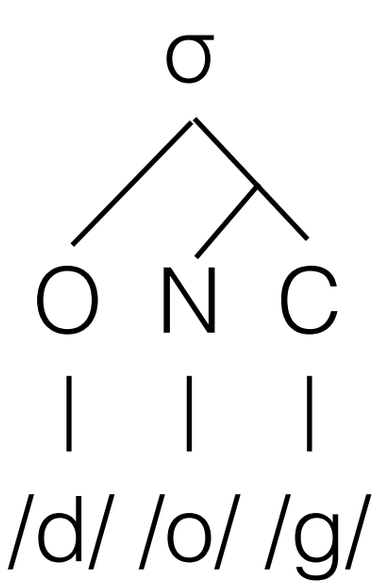


Joanisse & Seidenberg (1999)

Beyond past tense morphology

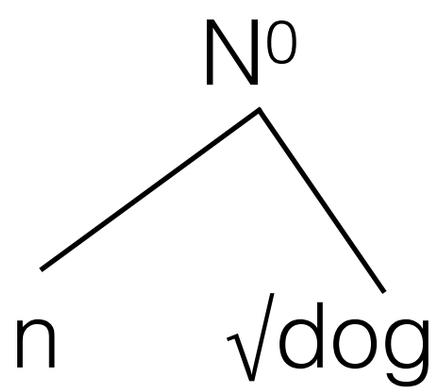
Do we decompose/compose complex units?

Phonology



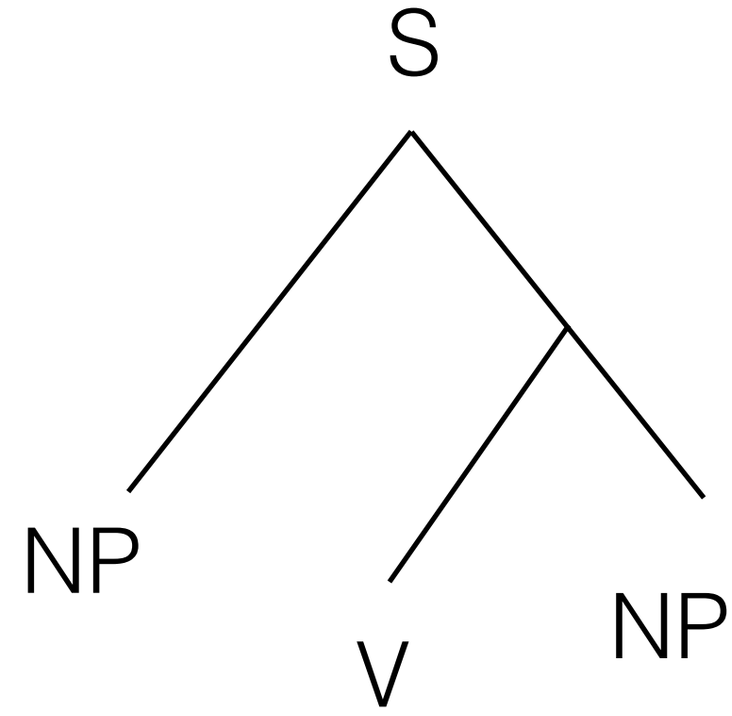
Syllables?

Morphology



Complex words?

Syntax



The dog chases the cat.

Phrases?
Idioms?