# Does the success of deep neural network language processing mean – finally! – the end of theoretical linguistics?

Paul Smolensky

Department of Cognitive Science

Johns Hopkins University

&

Jennifer Culbertson

Department of Linguistics and English Language

University of Edinburgh

SIGNLL Conference on Computational Natural Language Learning
31 July 2015 Beijing

# Our question

$\mathcal{Q}$: Shouldn't your DNNs for language processing have structured representations (e.g., tensor product representations *)?

** see below*

$\mathcal{A}$: Do you want

a. a pretty theory or

b. a system that works?

☞ Obviously:

✦ a

A Cog Sci sketch for such understanding:

Linguistics: how to understand linguistic cognition in terms of *structure processing*

Neural Networks: how to understand such structure processing as cognitively-motivated *NN processing*

Understanding: manifest in *theorem-proving*

- We already have 7.3 billion systems that work (as of July 2015).
- The burning question is:
  - ✦ ***How*** do they work? The goal is ***understanding***.

# The death of linguistic theory? A sketch from Cognitive Science

Why might CoNLLers care about a Cognitive Science sketch?

- Progress in the study of human cognition/language has anticipated corresponding progress* in NLP in some notable cases

  * conceptual, not just technical, progress

  - (Restricted) Boltzmann Machines [Hinton & Sejnowski 83; Smolensky 83]
  - Back-prop [Rumelhart, Hinton & Williams 86]
  - Bayes networks [Pearl's cognitive theory of causation 88]
  - Tensor methods [Dolan & Dyer 87; Smolensky 87, 90]
  - Maxent
    - general knowledge [Smolensky 83, 86]
    - grammar [Harmonic Grammar: Legendre, Miyata & Smolensky 90]

# Outline

## The death of linguistic theory? A sketch from Cognitive Science$^{++}$

II. Ling$^{++}$: how to understand linguistic cognition in terms of *structure processing*

A. Representations

1. Ling: Structural roles
2. Ling$^{++}$: Gradient structural roles

B. Grammars

1. Ling: Universal wellformedness constraints, *Con*
2. Ling$^{++}$: *SoftCon*, Bayesian universal biases
3. Limiting the power of language learning; innate linguistic knowledge?
   1. Argument from Universals* [*not* from Poverty of the Stimulus]
   2. New experimental evidence supporting *

I. NN$^{++}$: Understandably implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Understanding structure computation: (gradient) symbolic functions

C. Understanding encoding of structural wellformedness: Ling$^{++}$ grammars

*Upshot:* Upgraded to Ling$^{++}$, implemented in NN$^{++}$, linguistics may lead us to deeper understanding of how to compute natural language (learning)

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

II. Ling$^{++}$: how to understand linguistic cognition in terms of *structure processing*

A. Representations

1. Ling: Structural roles
2. Ling$^{++}$: Gradient structural roles

B. Grammars

1. Ling: Universal wellformedness constraints, *Con*
2. Ling$^{++}$: *SoftCon*, Bayesian universal biases
3. Limiting the power of language learning; innate linguistic knowledge?
   i. Argument from Universals* [*not* from Poverty of the Stimulus]
   ii. New experimental evidence supporting *

I. NN$^{++}$: Understandably implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Undertanding structure computation: (gradient) symbolic functions

C. Understanding encoding of structural wellformedness: Ling$^{++}$ grammars

# Linguistic representations: The power of structure

Linguistic-theoretic semantics: elementary explanatory ingredients

1. Argument Structure:
   syntactic ↔ semantic roles: [L [loves M]] ↔ loves(L, M)
2. Embedding/recursion: said(J, thinks(K, loves(L, M)))
3. Semantic (LF) operators ¬ , ∃, if, ◇
   a. Operator/variable binding $\exists x.P(x)$
   b. Operator scope $(\neg P)\ \&\ Q$ $\neg\ (P\ \&\ Q)$
4. Compositionality ¬ , & $P, Q$
5. β-reduction in λ calculus $(\lambda x.\mathrm{B})\mathrm{A} = \mathrm{B}$ with every $x \rightarrow \mathrm{A}$
6. Tree adjoining in TAG

# Outline

## The death of linguistic theory? A sketch from Cognitive Science

I. NN$^{++}$: Implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Computing (gradient) symbolic functions

C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

A. Representations

1. Ling: Structural roles

2. Ling$^{++}$: Gradient structural roles

B. Grammars

1. Ling: Universal wellformedness constraints, *Con*

2. Ling$^{++}$: *SoftCon*, Bayesian universal biases

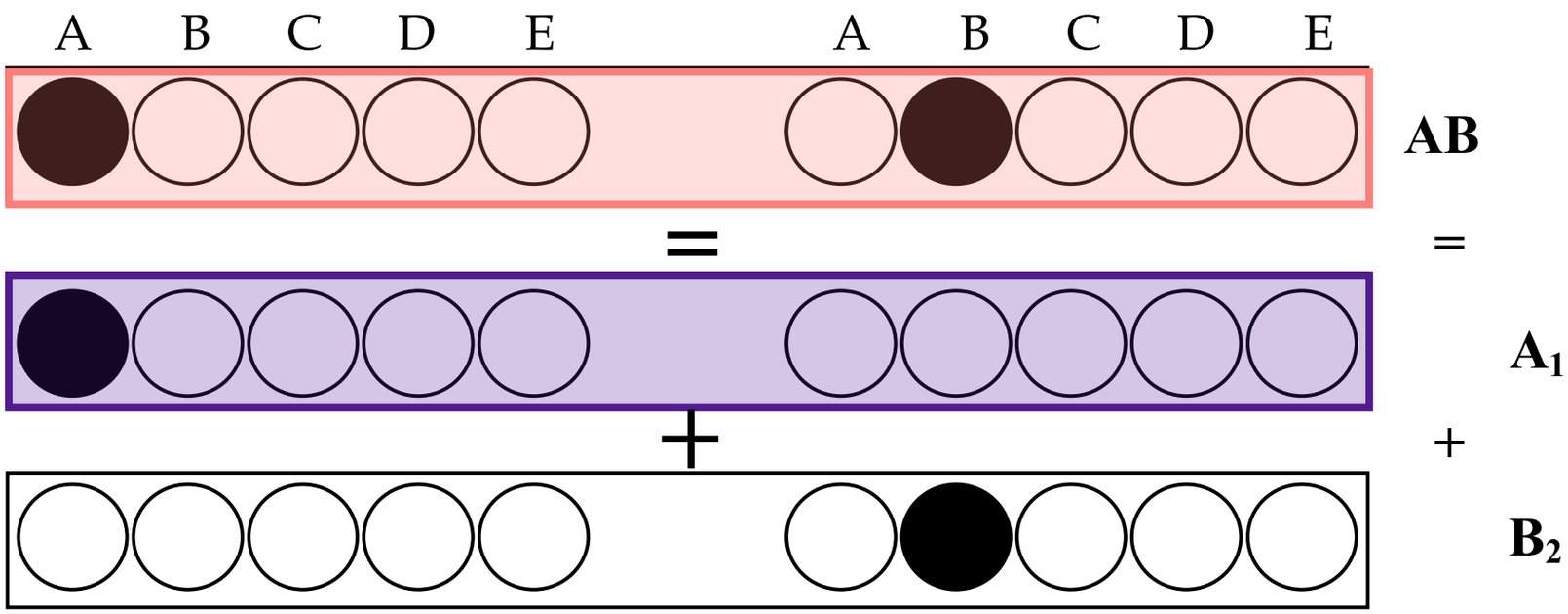
3. Limiting the power of language learning; innate linguistic knowledge?

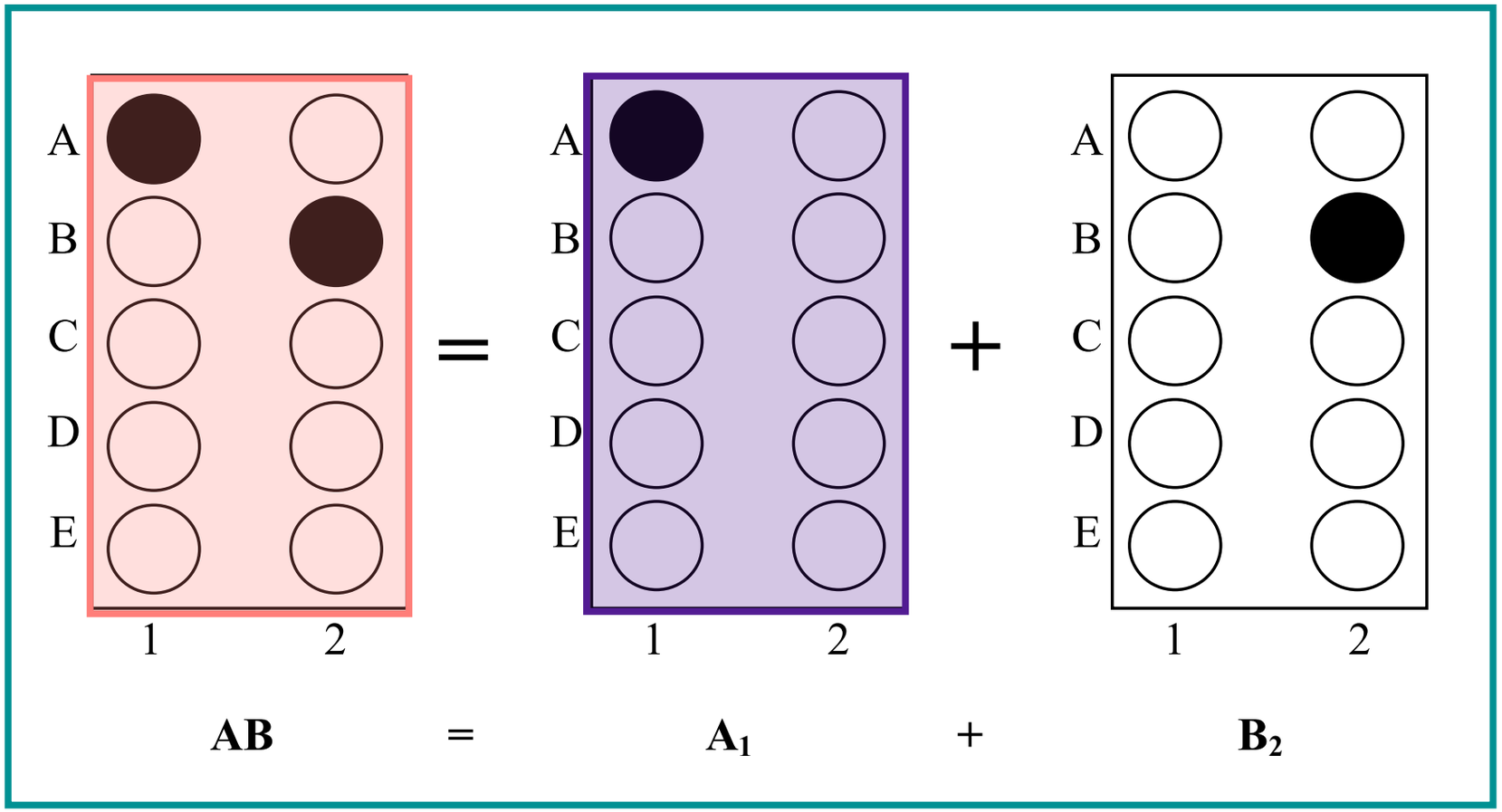
i. Argument from Universals* [*not* from Poverty of the Stimulus]

ii. New experimental evidence supporting *

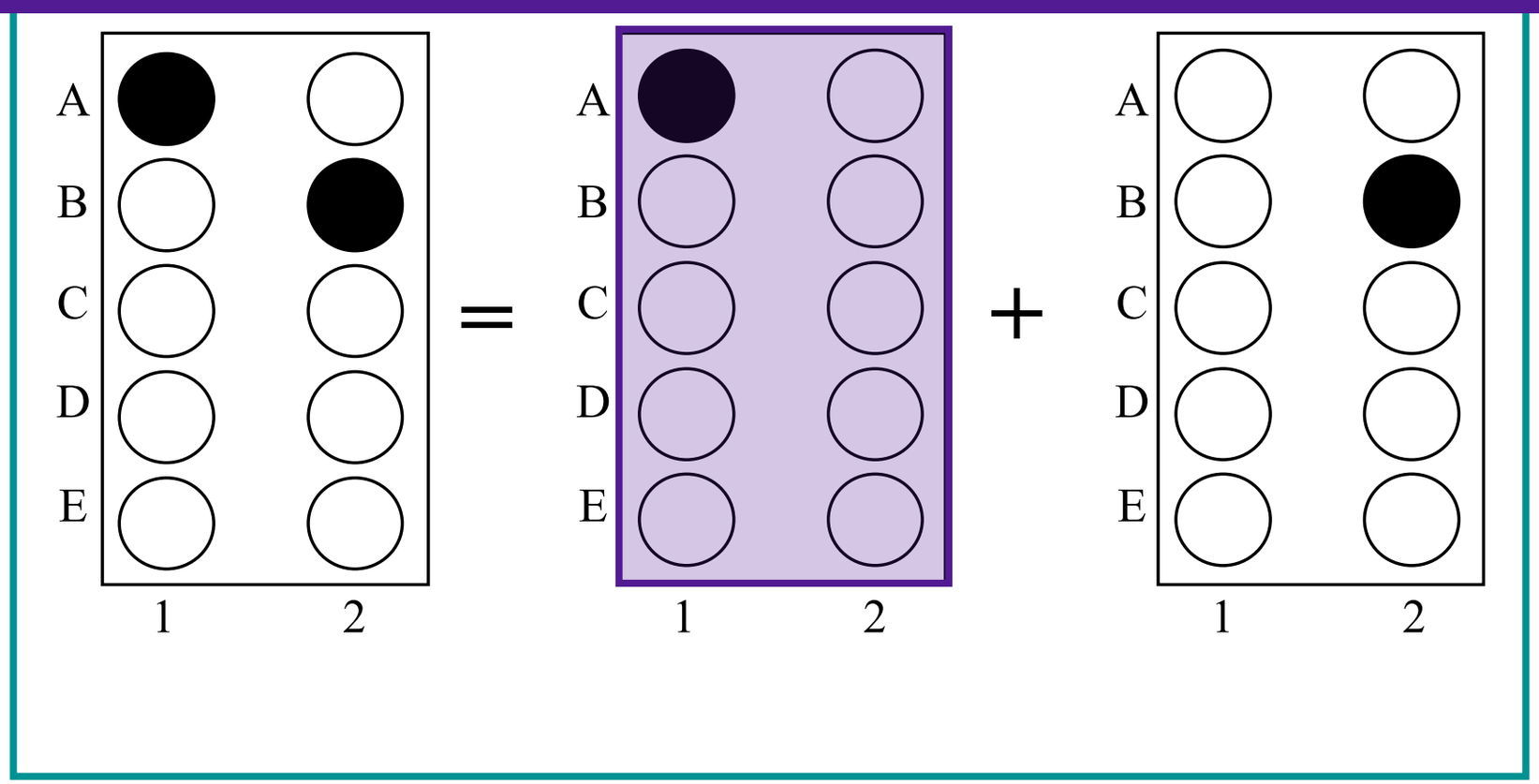
*Upshot:* Upgraded to Ling$^{++}$, implemented in NN$^{++}$, linguistics may lead us to deeper understanding of how to compute natural language (learning)

# Fully local representation of letter strings

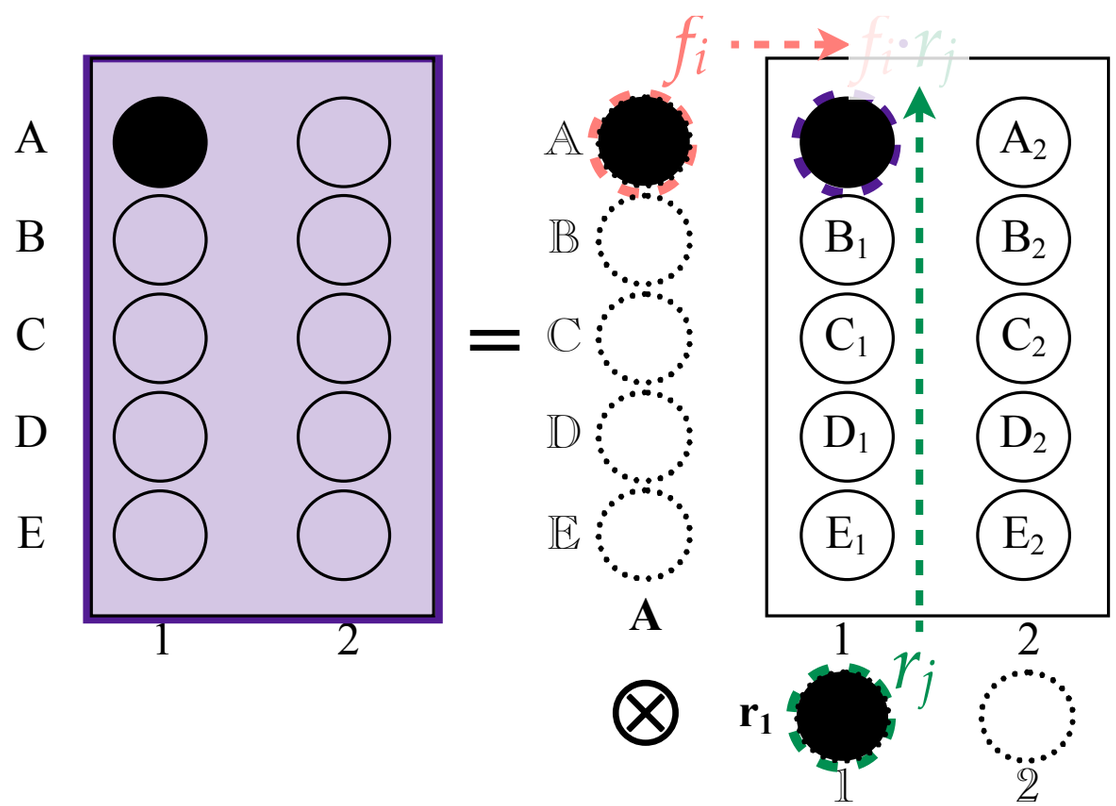


# Fully local representation of letter strings



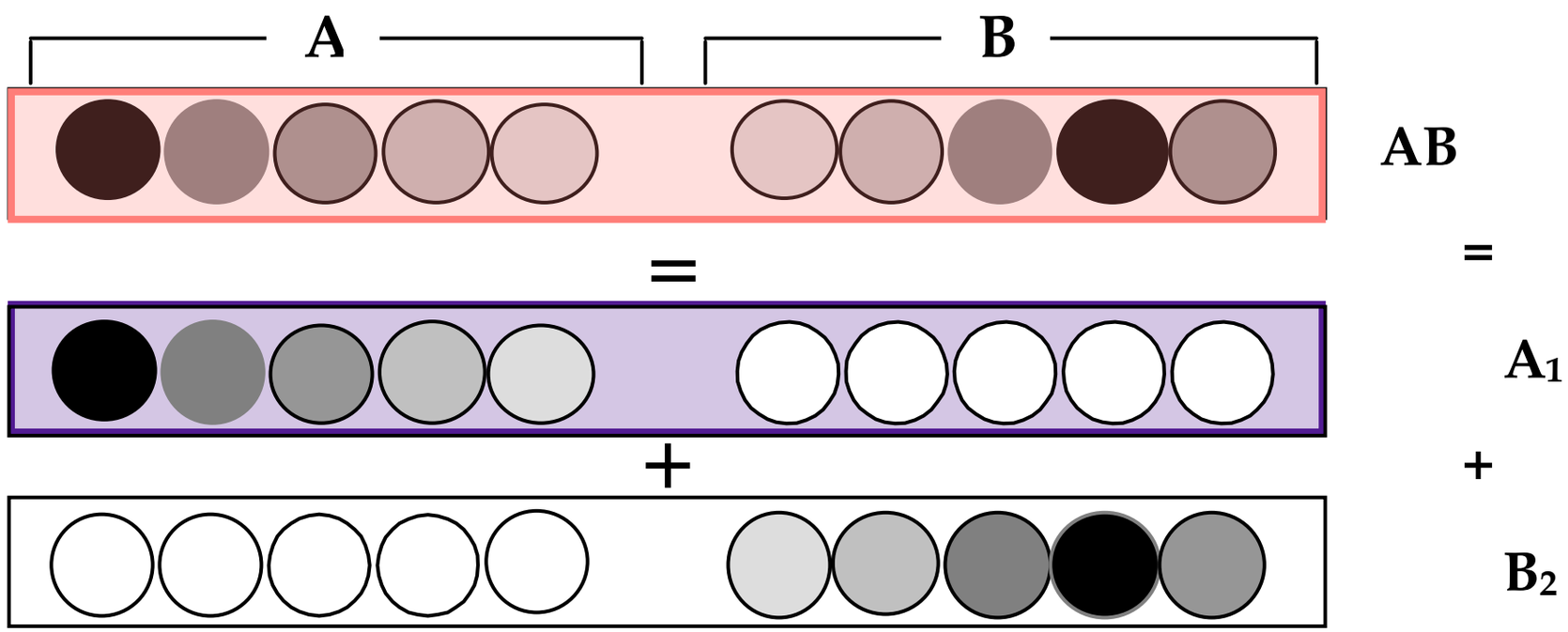

$$\mathbf{AB} = \mathbf{A} \otimes \mathbf{r}_1 + \mathbf{B} \otimes \mathbf{r}_2 = \sum_x \mathbf{a}_x \otimes \mathbf{r}_x$$

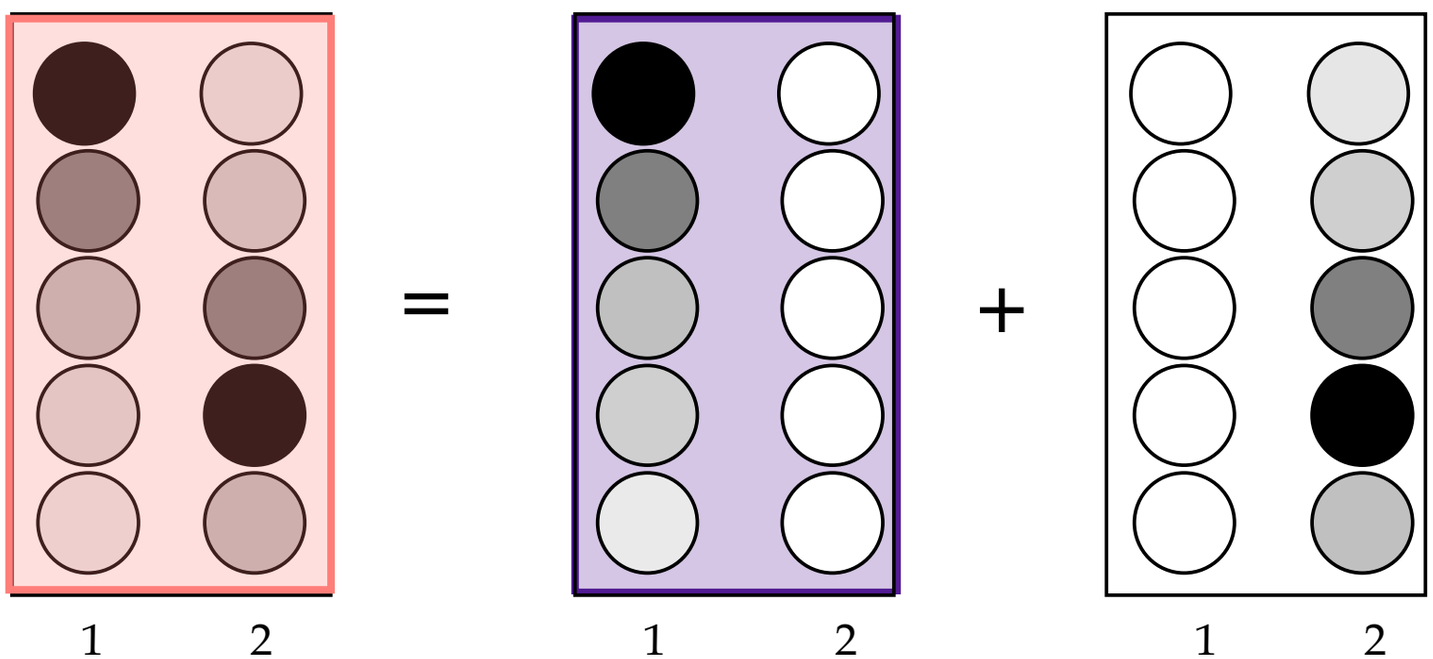



**Fully local case**
(both **A** and $\mathbf{r}_1$ are 1-hot)

$$\mathbf{A}_1 = \mathbf{A} \otimes \mathbf{r}_1$$

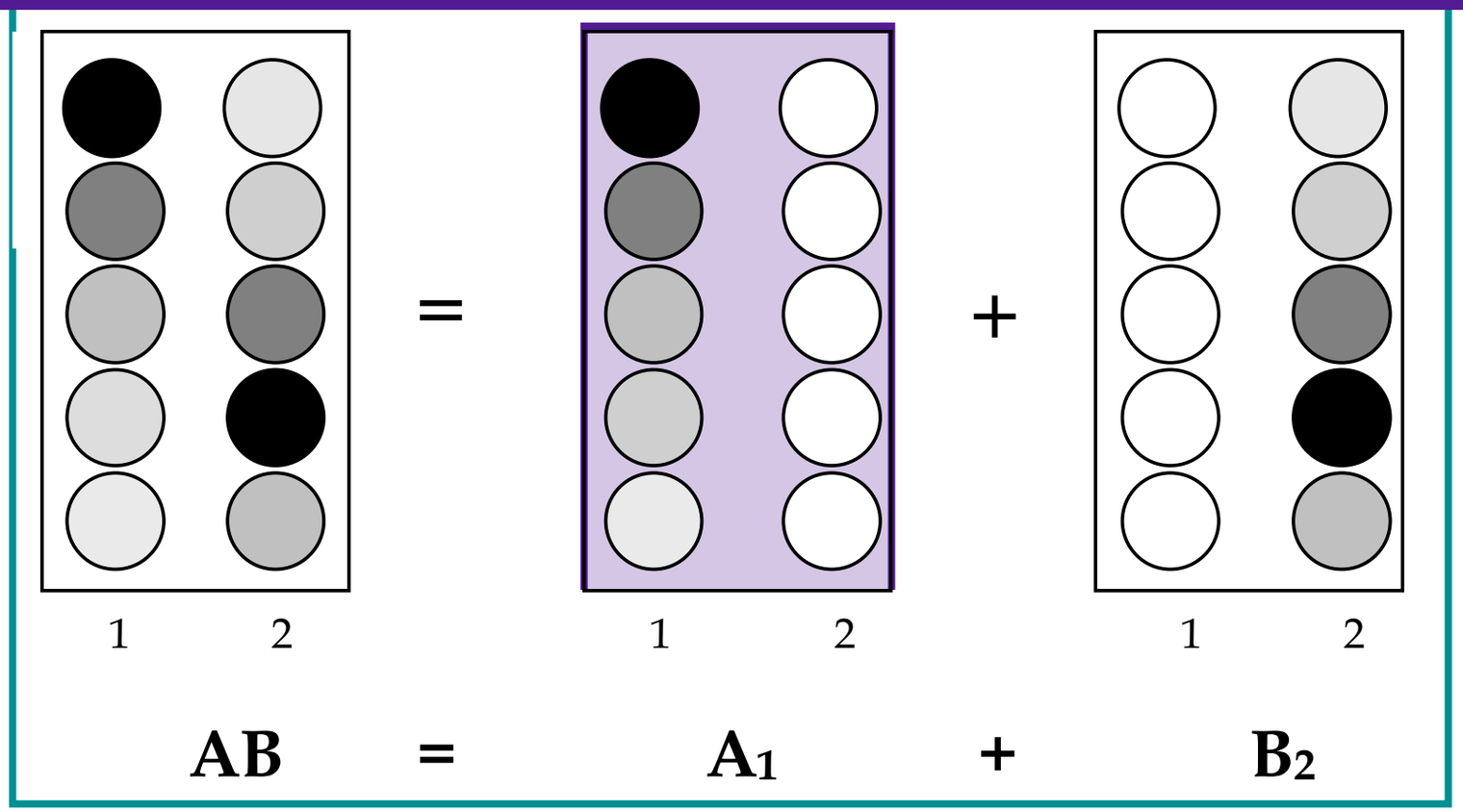
# Fully local representation of letter strings



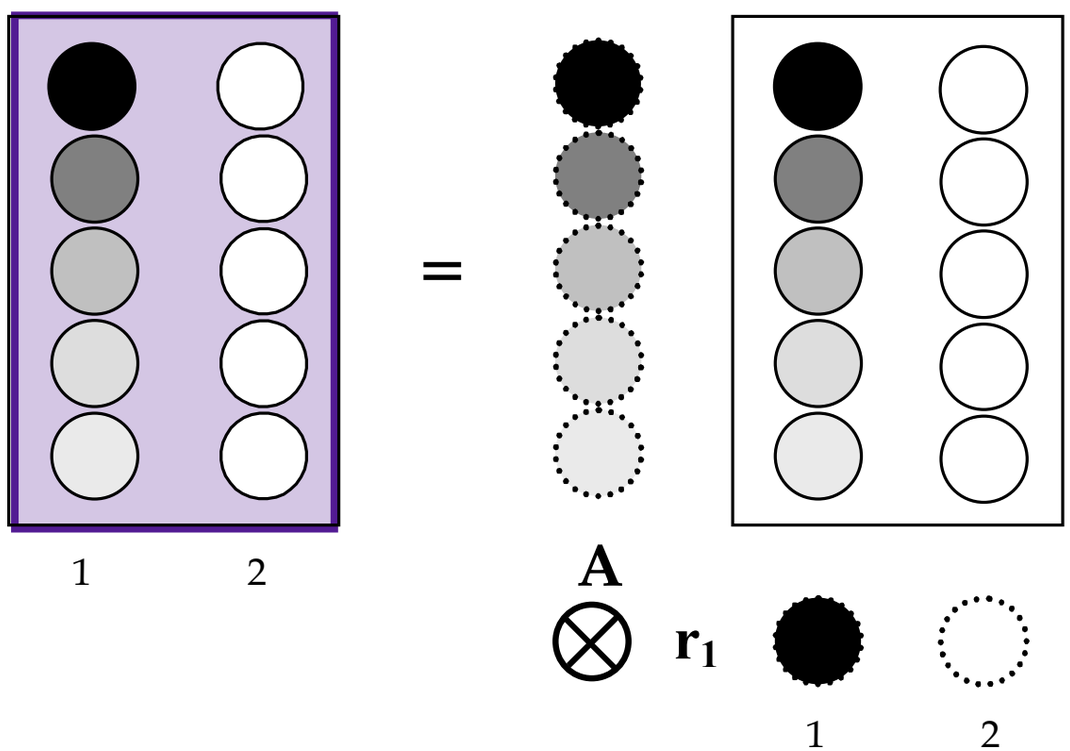




**AB = $A_1$ + $B_2$**

# Semi-local representation of letter strings



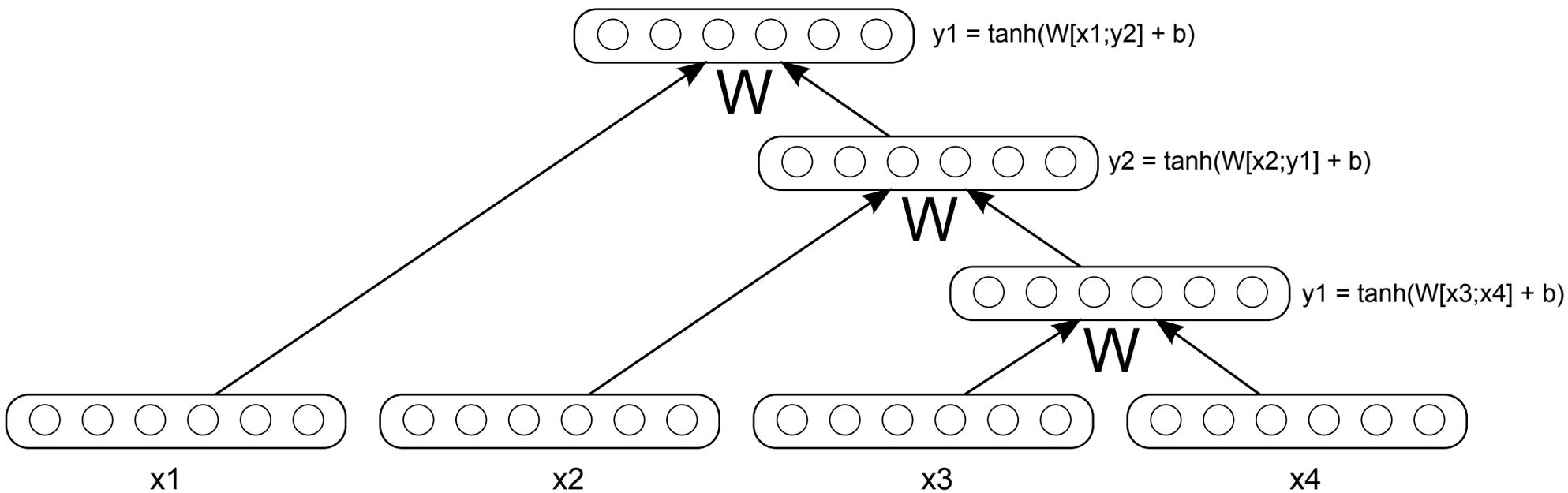

$$\mathbf{AB} = \mathbf{A} \otimes \mathbf{r}_1 + \mathbf{B} \otimes \mathbf{r}_2$$
$$= \sum_x \mathbf{a}_x \otimes \mathbf{r}_x$$




**Semi-local case**
($\mathbf{r}_1$ but not **A** is 1-hot)

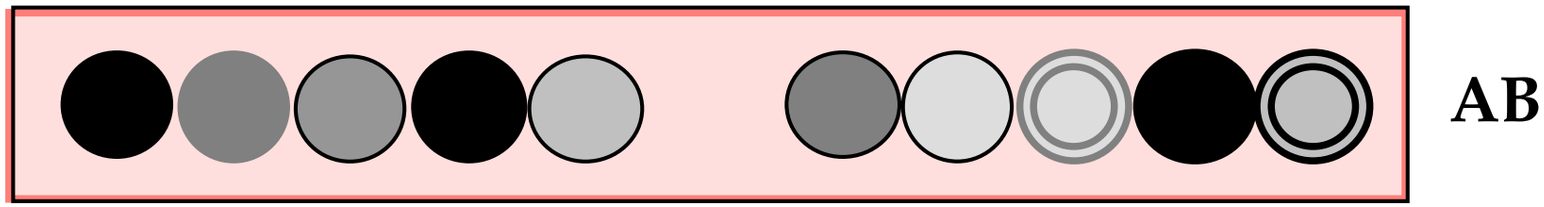
# Semi-local representation of trees

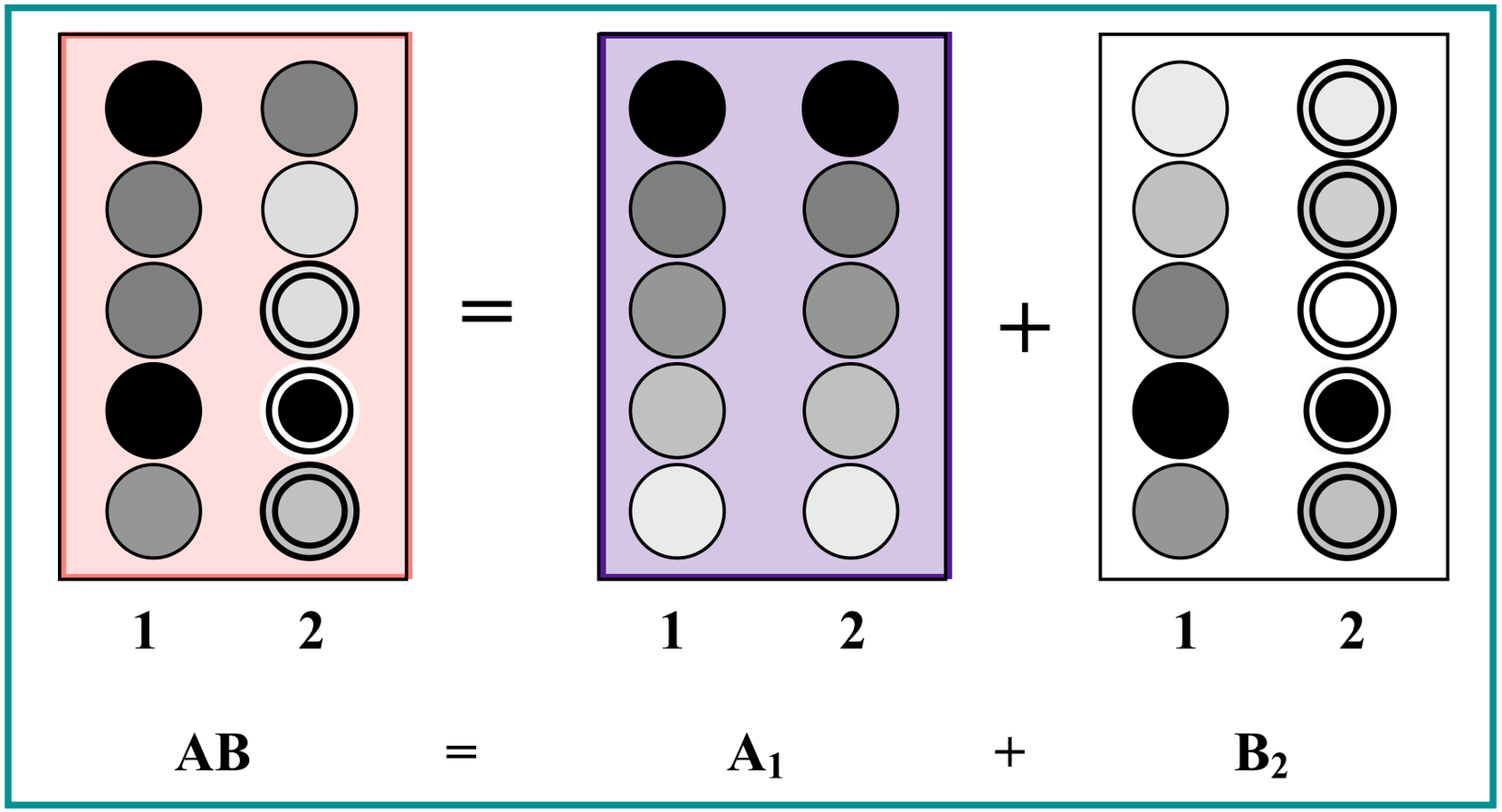
Recursive NNs:
Implicit role vectors are local



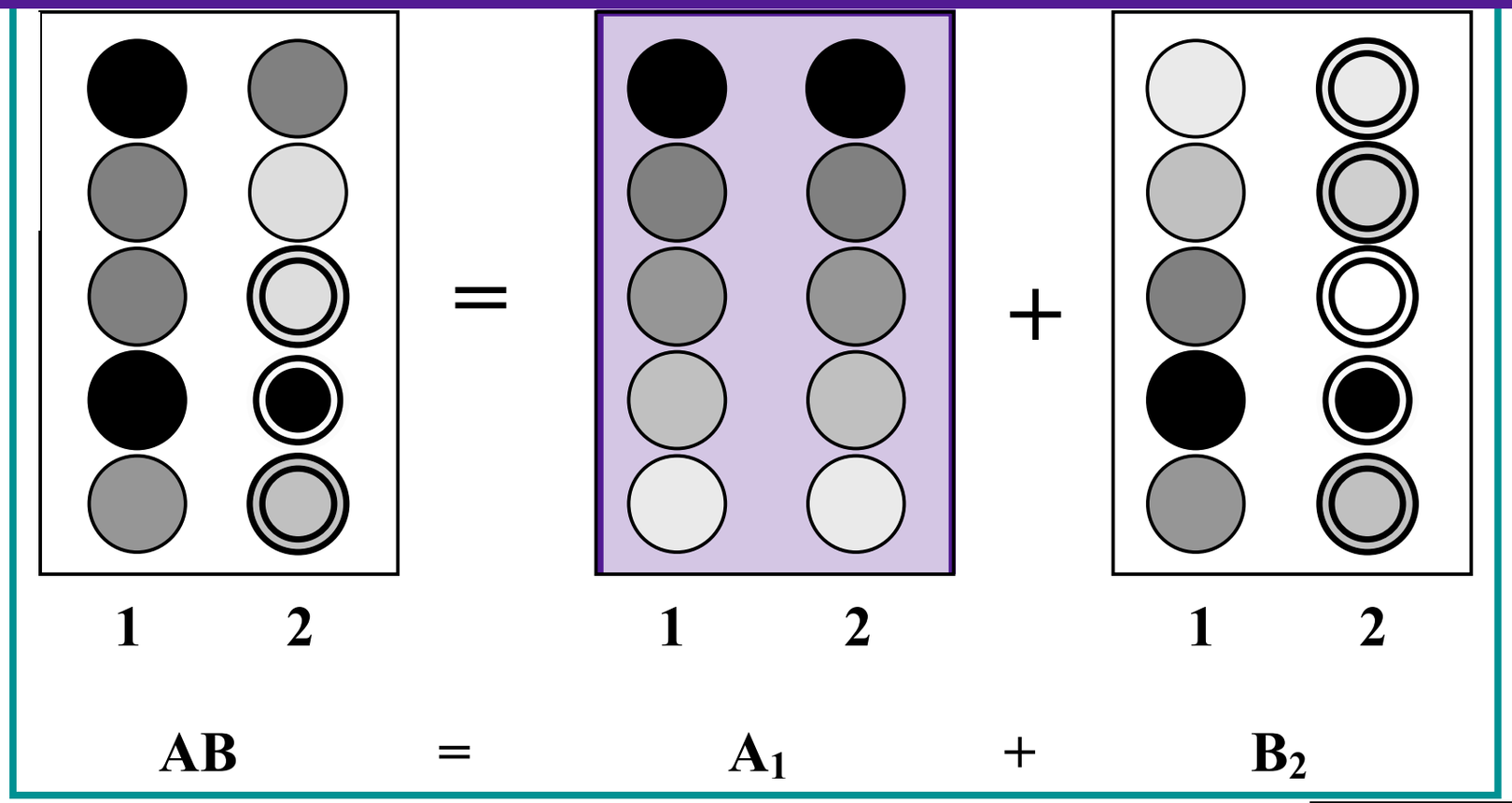

Socher, Manning & Ng 2010

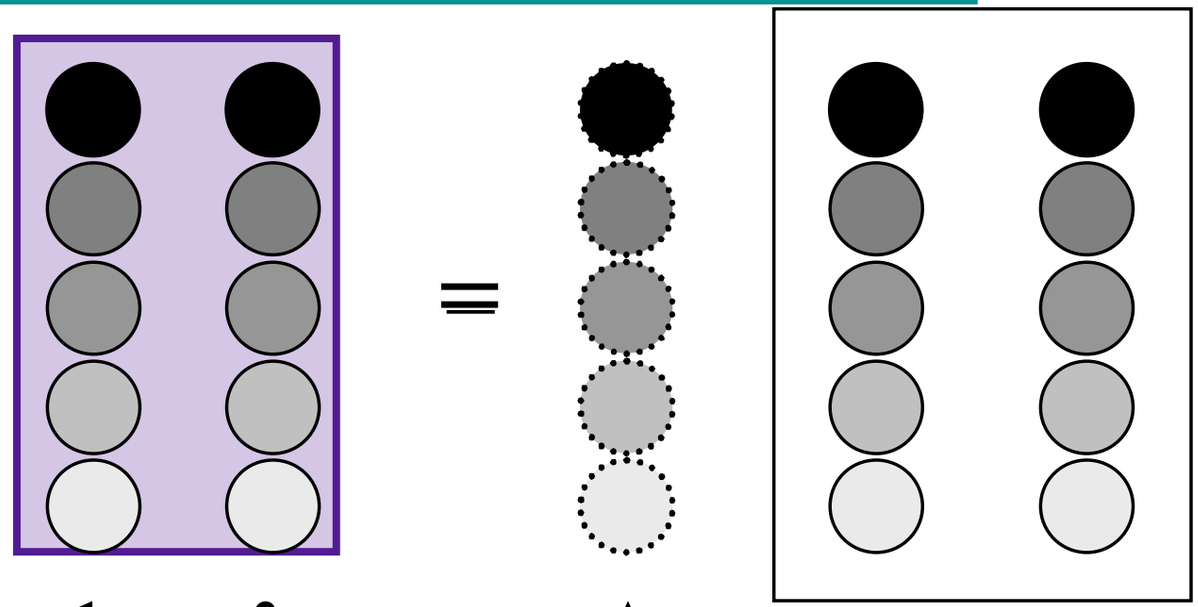
# Fully distributed representation of letter strings



# Fully distributed representation of letter strings



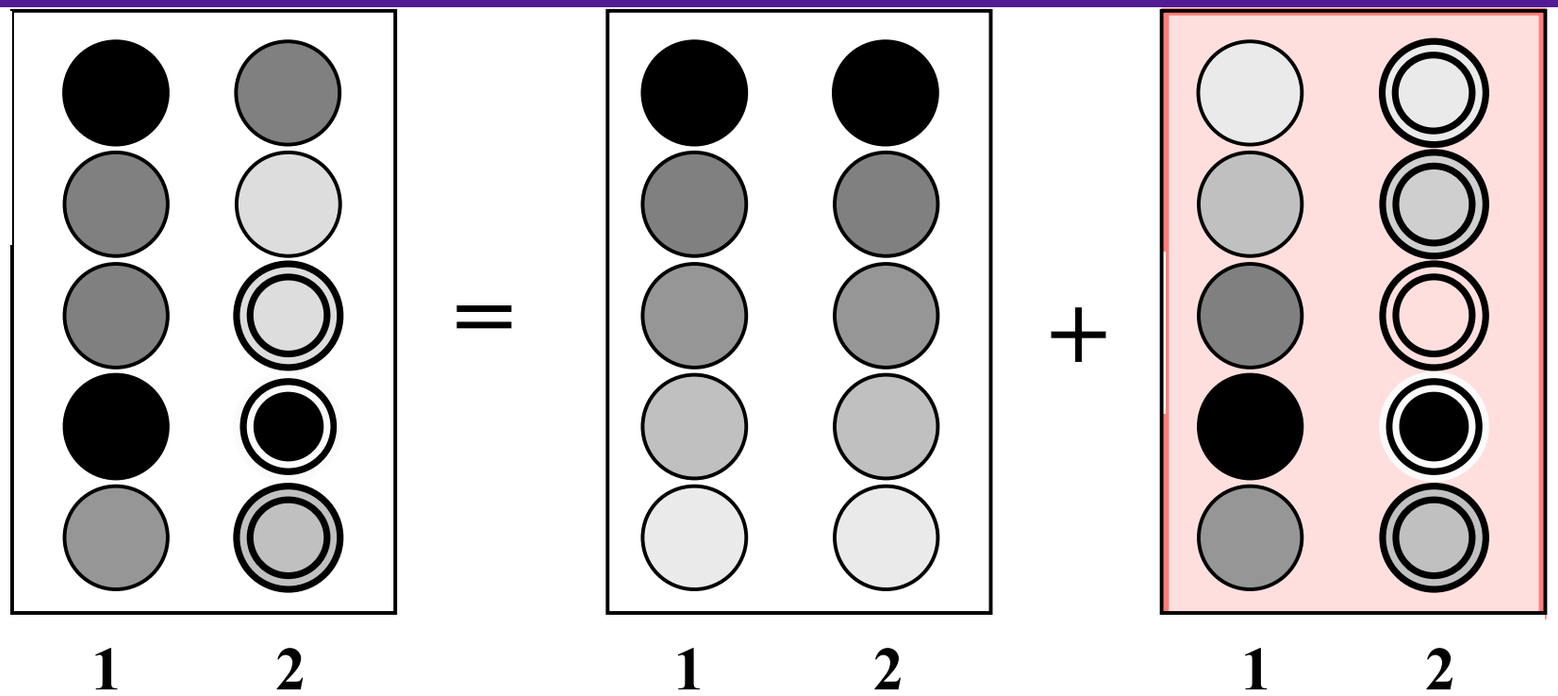




$\otimes$ $r_1$

**Distributed role vector**:
A key idea of this work

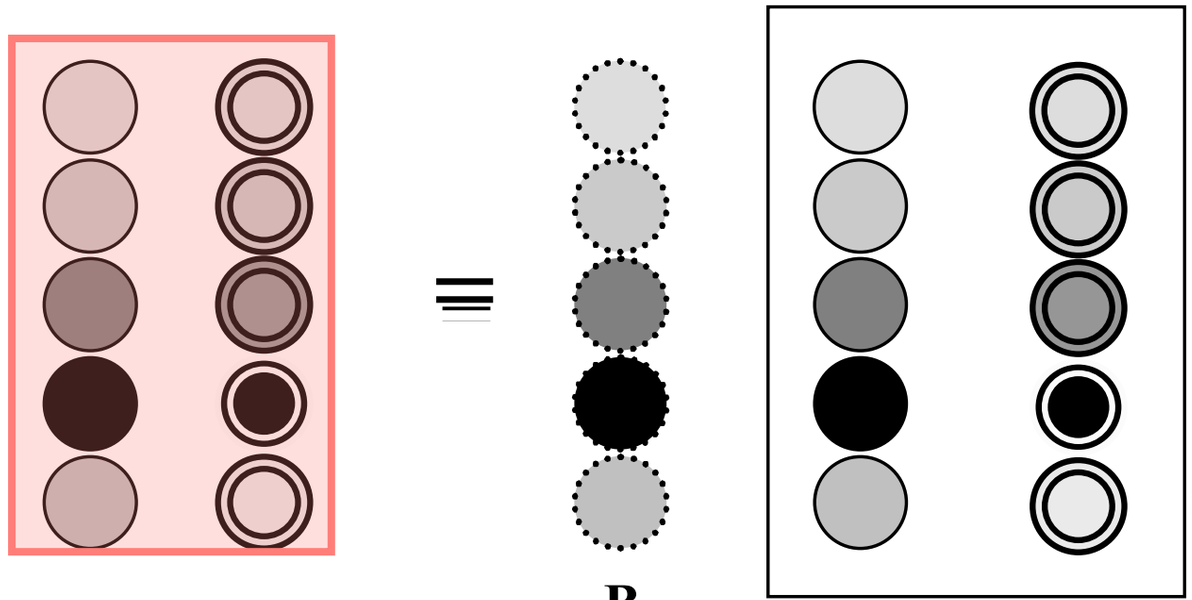
$$A_1 = A \otimes r_1$$

# Fully distributed representation of letter strings



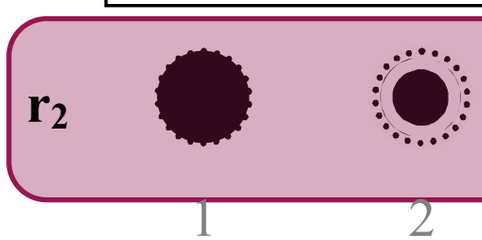

**AB** = $\mathbf{A}_1$ + $\mathbf{B}_2$

$$\mathbf{AB} = \mathbf{A} \otimes \mathbf{r}_1 + \mathbf{B} \otimes \mathbf{r}_2 = \sum_x \mathbf{a}_x \otimes \mathbf{r}_x$$




**Fully distributed case**
(neither $\mathbf{r}_1$ nor **A** is 1-hot)

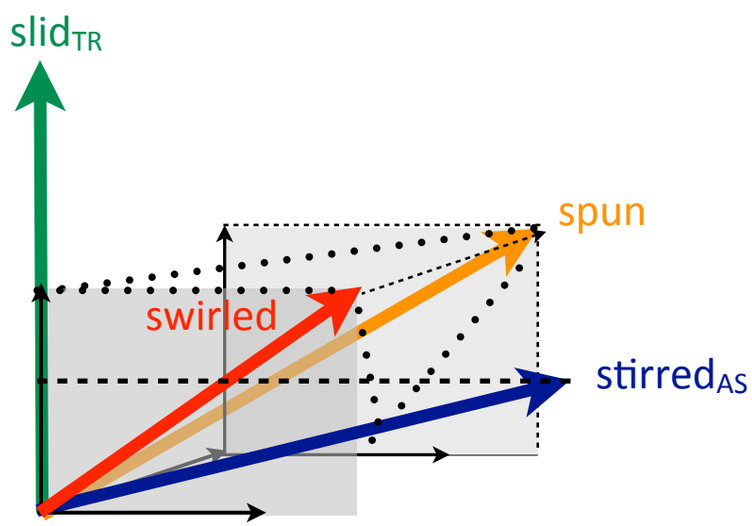
Every unit contributes to symbols in *both* positions.

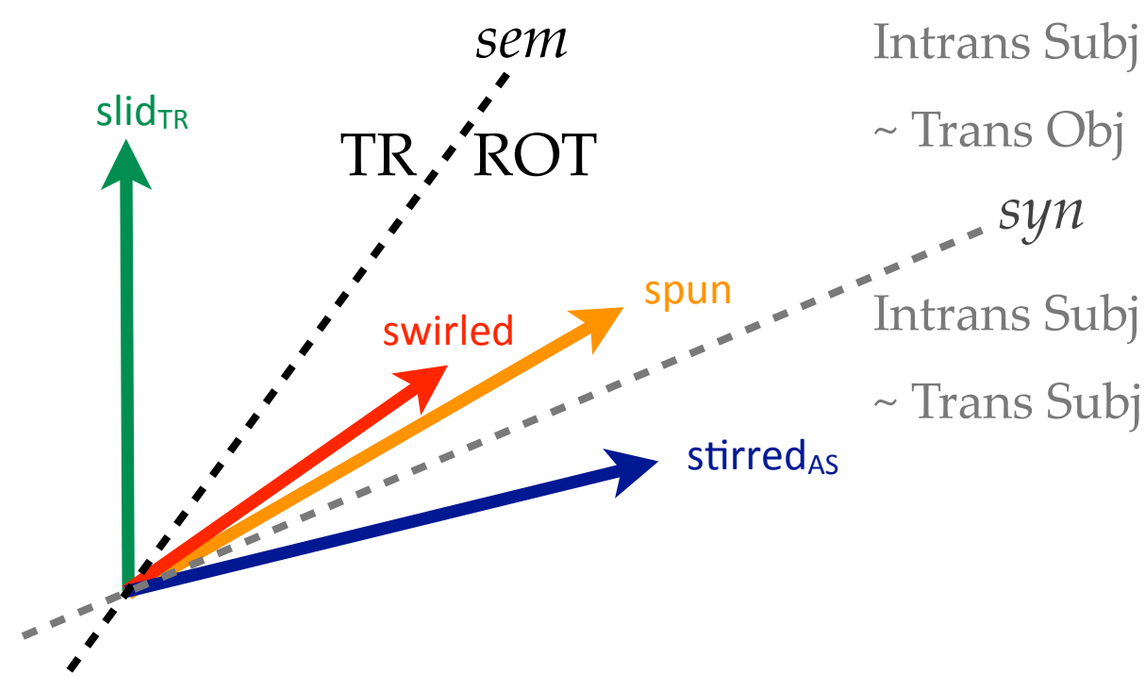



**Distributed role vector**:
A key idea of this work

$\mathbf{B}_2$ = $\mathbf{B} \otimes \mathbf{r}_2$

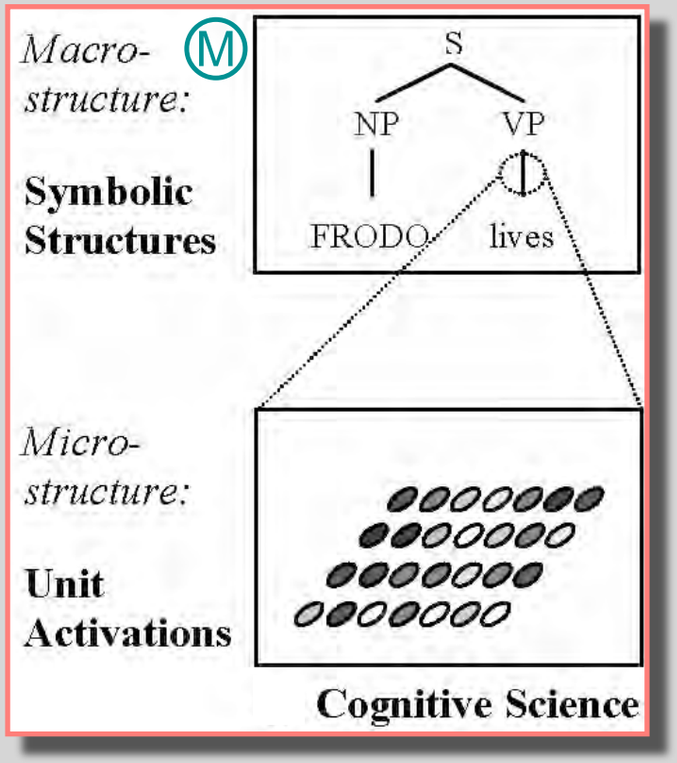
# Arg-struc similarities among verbs' SYN/SEM-role vectors

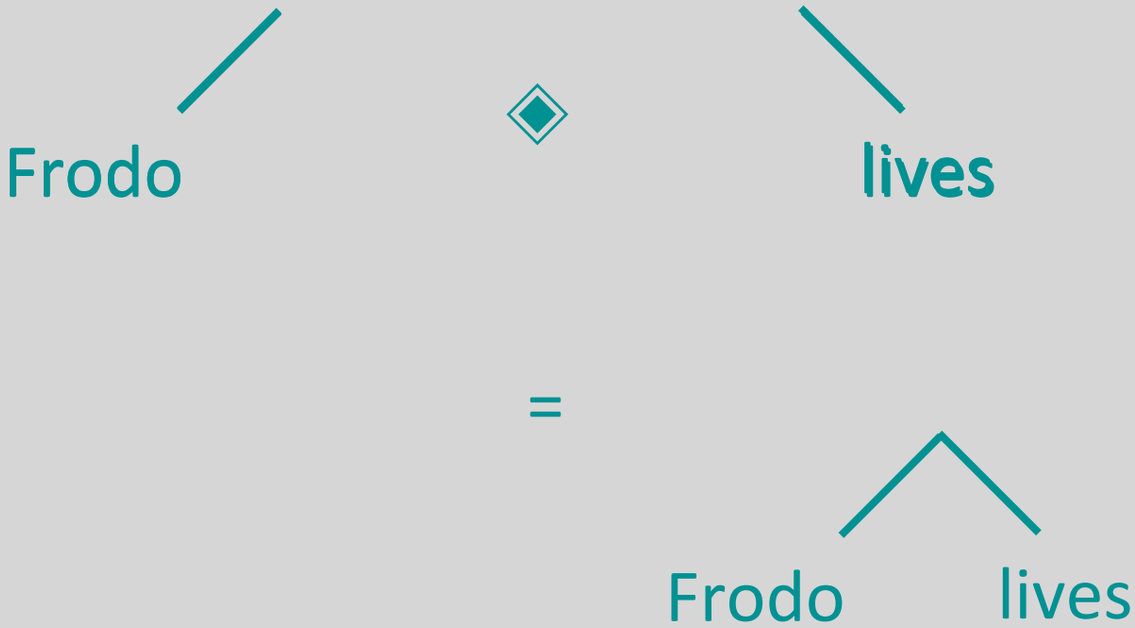


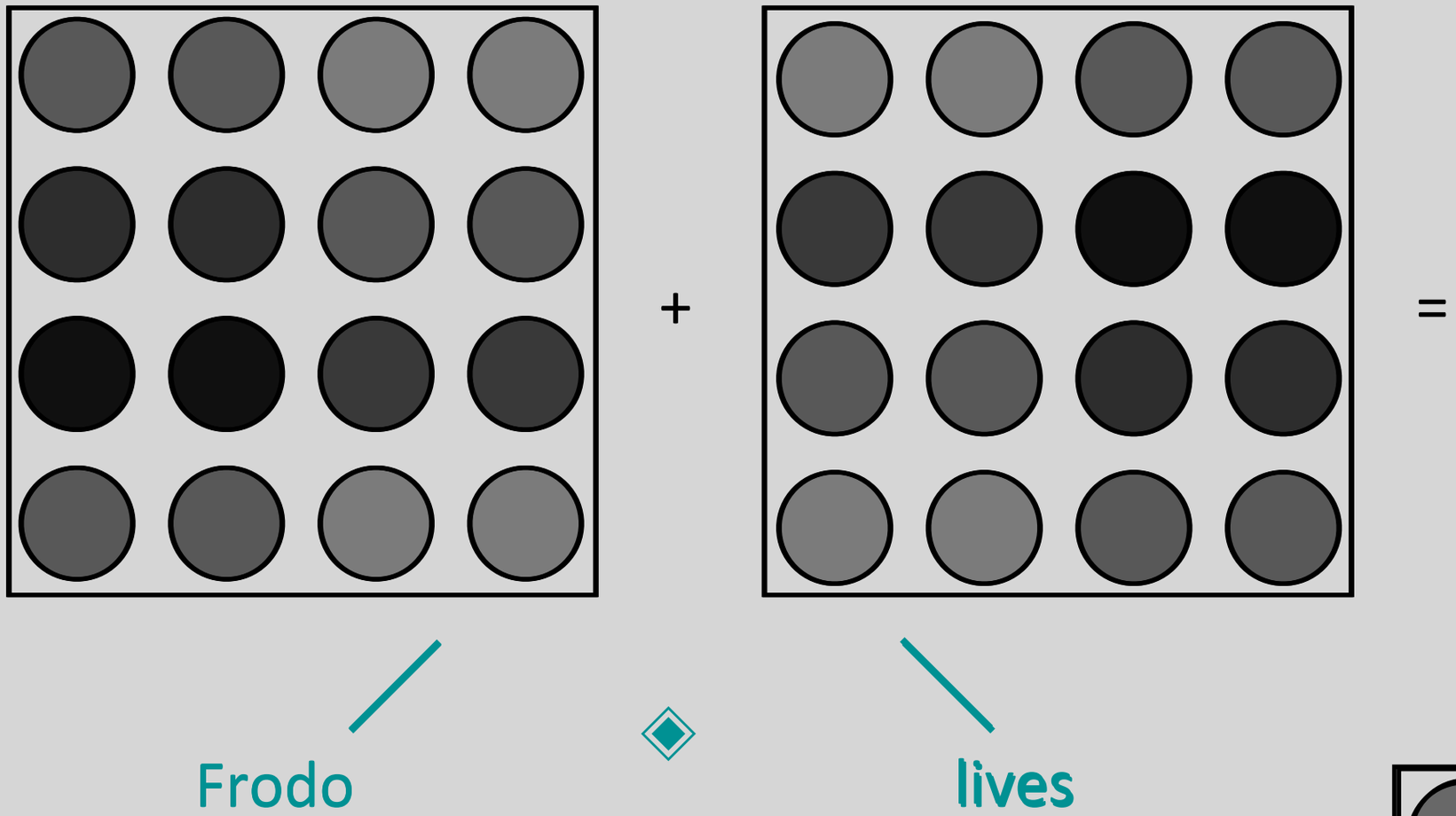

All Vs are *sem* = ROT, *syn* = [Intrans Subj ~ Trans Obj] ("unaccusative")
except where noted TR or [Intrans Subj ~ Trans Subj] ("unergative")

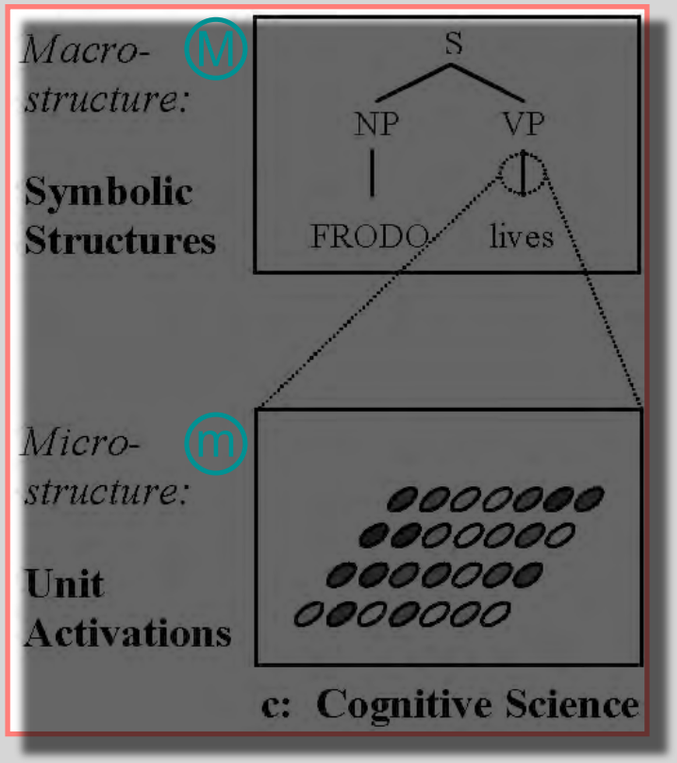
# *Reprise* — Combinatorial structure in activation patterns



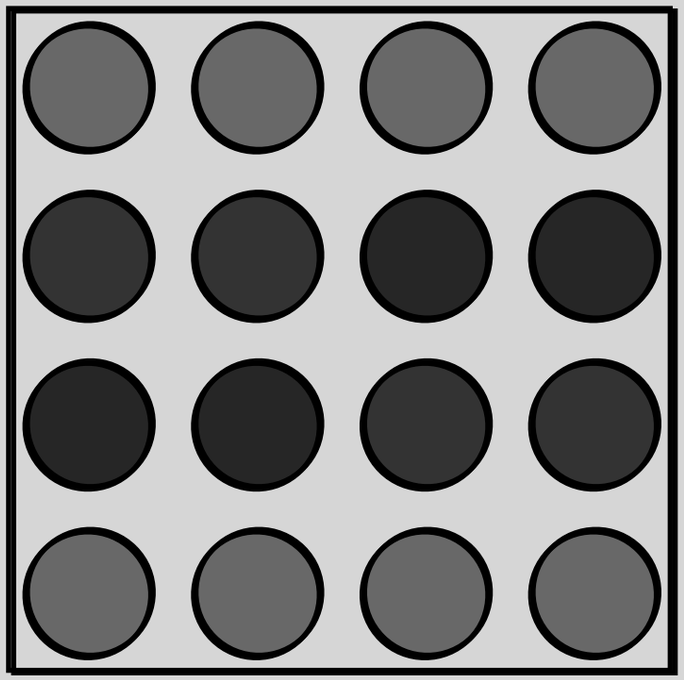



# *Reprise* — Combinatorial structure in activation patterns

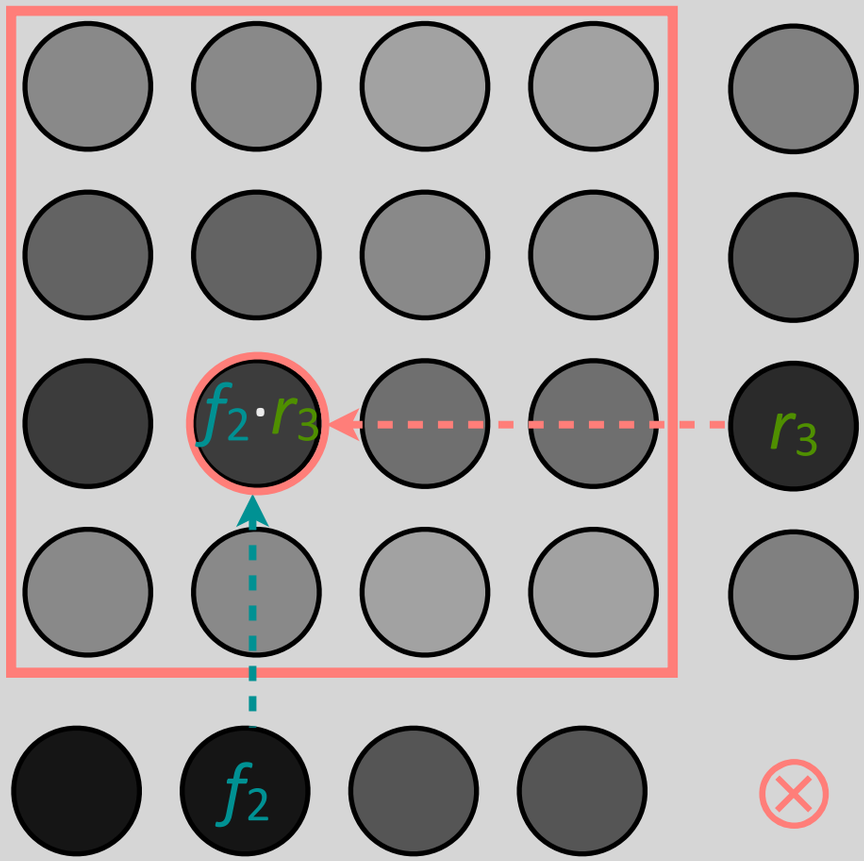






# *Reprise* — Constituent construction



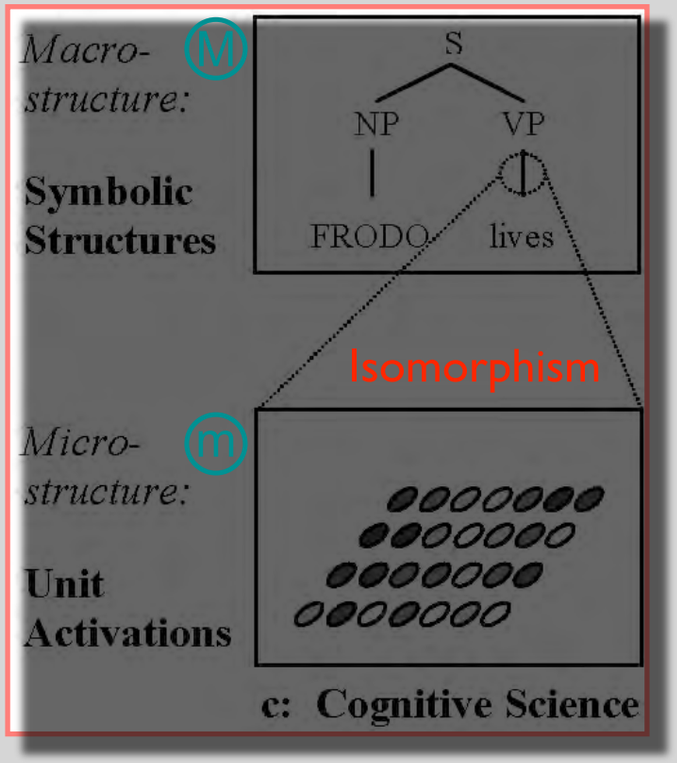

*Role:*

*Filler:*
Frodo

⊗

Frodo

Filler pattern
⊗
Role pattern
=
Filler/role
constituent pattern




**Tensor product representations**

Ⓜ $\mathbf{s} = \{\mathbf{f}_k/\mathbf{r}_k\}$

ⓜ $\mathbf{s} = \sum_k \mathbf{f}_k \otimes \mathbf{r}_k$

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Computing (gradient) symbolic functions

C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

A. Representations

1. Ling: Structural roles

2. Ling$^{++}$: Gradient structural roles

B. Grammars

1. Ling: Universal wellformedness constraints, *Con*

2. Ling$^{++}$: *SoftCon*, Bayesian universal biases

3. Limiting the power of language learning; innate linguistic knowledge?

i. Argument from Universals* [*not* from Poverty of the Stimulus]

ii. New experimental evidence supporting *

Can we understand what structure processing NN$^{++}$ can do?

*Yes.*

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Computing (gradient) symbolic functions

C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

A. Representations

1. Ling: Structural roles

2. Ling$^{++}$: Gradient structural roles

B. Grammars

1. Ling: Universal wellformedness constraints, *Con*

2. Ling$^{++}$: *SoftCon*, Bayesian universal biases

3. Limiting the power of language learning, innate linguistic knowledge?

i. Argument from Universals* [*not* from Poverty of the Stimulus]

ii. New experimental evidence supporting *

Single-step operation, massively parallel, from distributed encoding of I to distributed encoding of O. Implemented in the simplest NN: linear I$\longrightarrow$ O net

Basic observation: Primitive tree-constructing and tree-constituent-accessing functions are implementable as linear transformations — matrix multiplication.

An arbitrarily complex composition of these is still just a single linear transformation.

So a linear network can compute the set of recursive functions that is the closure under composition of the primitive tree operations.

To implement a recursive function defined by primitive-recursion equations, solve the corresponding recursion equations for *matrices*.

# Outline

## The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Computing (gradient) symbolic functions

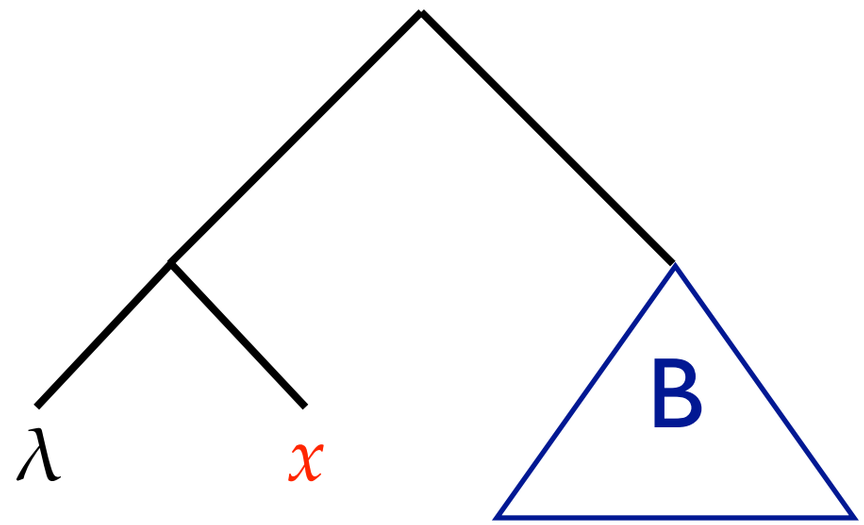
llformedness constraints, *Con*

2. Ling$^{++}$: *SoftCon*, Bayesian universal biases
3. Limiting the power of language learning; innate linguistic knowledge?
   i. Argument from Universals* [*not* from Poverty of the Stimulus]
   ii. New experimental evidence supporting *

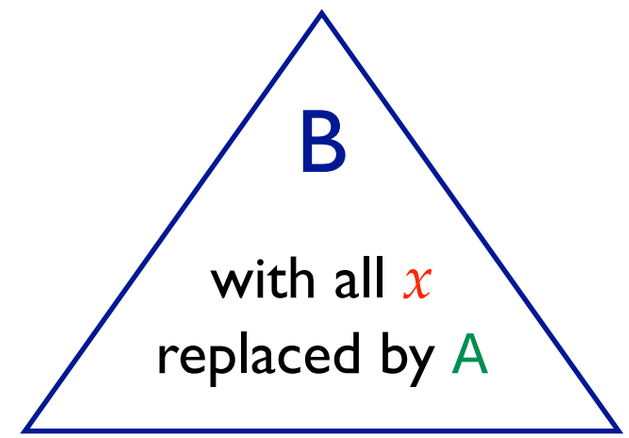
# β-Reduction

Basic operation of $\lambda$-calculus [function application]: $(\lambda x.\mathsf{B})\mathsf{A}$

$\Lambda =$




$\mapsto$



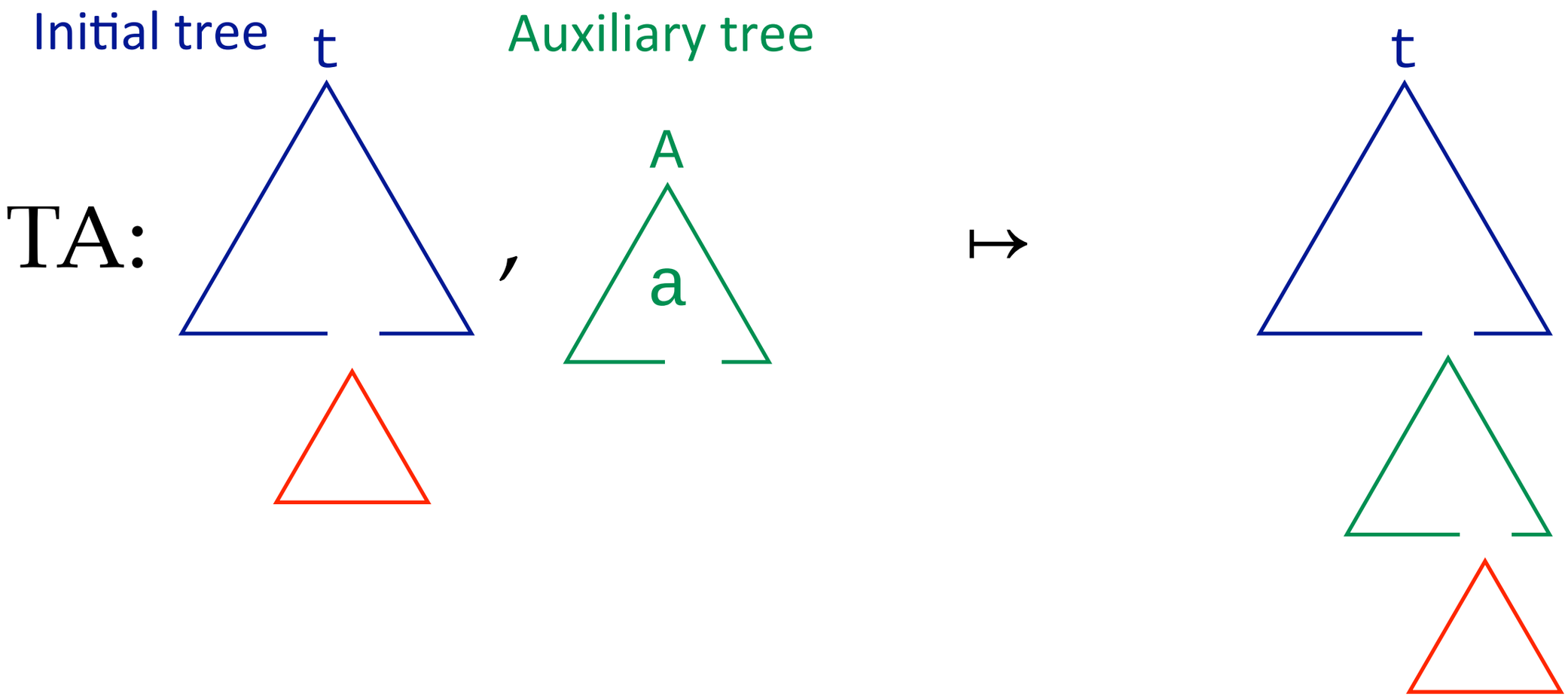

$\mathbf{\Lambda \cdot r_{10}^{+} \equiv x}$

$\mathbf{\Lambda \cdot r_{1}^{+} \equiv B}$

$$\beta\text{-reduce}(\Lambda, \mathsf{A}) = \left[\, \mathbb{1} + (\mathbf{A} - \mathbf{x}) \otimes \mathbf{x}^{+} \,\right] \mathbf{B}$$

*Next:* Substituting-in a structure rather than an atom.

# Tree Adjoining




**TA(t,a) =**

| | | | |
|---|---|---|---|
| $\mathbf{t} - \mathbf{A} \otimes \mathbf{R}_{\mathbf{A}}$ | retain all of **t** unaffected by adjoining | $\mathbf{a} \cdot \mathbf{r}_{\varepsilon}^{+} \equiv \mathbf{A}$ | root symbol of **a** ("**A**") |
| $+\mathbf{A} \otimes [\mathbf{R}_{\alpha} \otimes \mathbf{R}_{\mathbf{A}}]$ | reposition $[_{\mathbf{A}} \cdots]$ in **t** | $\mathbf{A}^{+} \cdot \mathbf{t} \equiv \mathbf{R}_{\mathbf{A}}$ | location of **A** in **t** |
| $+\mathbf{a} \otimes \mathbf{R}_{\mathbf{A}}$ | embed **a** in place of **A** | $\mathbf{t} \cdot \mathbf{R}_{\mathbf{A}}^{+} \equiv \mathbf{A}$ | subtree **A** in **t** |
| $-\alpha \otimes [\mathbf{R}_{\alpha} \otimes \mathbf{R}_{\mathbf{A}}]$ | remove $\alpha$ from embedded **a** | $\alpha^{+} \cdot \mathbf{a} \equiv \mathbf{R}_{\alpha}$ | location of $\alpha$ in **a** |

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

A. Neural networks with structured distributed vectorial representations

B. Computing (gradient) symbolic functions

C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

A. Representations

1. Ling: Structural roles

2. Ling$^{++}$: Gradient structural roles

B. Grammars

1. Ling: Universal wellformedness constraints: *Con*

2. Ling$^{++}$: *SoftCon*, Bayesian universal biases

3. Limiting the power of language learning; innate linguistic knowledge?

i. Argument from Universals* [*not* from Poverty of the Stimulus]

ii. New experimental evidence supporting *

Ling$^{++}$ grammars are Harmonic Grammars that reward or penalize the co-presence of 1–2 constituents within a single structure.

Key observation: These grammars can be straightforwardly implemented as connection matrices that generate, in the co-presence of the distributed representations of the target constituents, the specified contribution to the *Harmony* of the network state: this measure of wellformedness is maximized by activation-spreading dynamics in the network.

When appropriate stochastic dynamics are deployed, a Maxent distribution results, the log-probability of a state being proportional to its Harmony.

# Questions

Why do DNNs for language do as well as they do?

Are they approximating the kind of structural representation and processing characteristic of $NN^{++}$ systems?

If DNNs were constructed to incorporate $NN^{++}$ structural processing, would they do even better?

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

   A. Neural networks with structured distributed vectorial representations

   B. Computing (gradient) symbolic functions

   C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

   A. Representations

      1. Ling: Structural roles

      2. Ling$^{++}$: Gradient structural roles

ledge?

      i. Argument from Universals* [*not* from Poverty of the Stimulus]

      ii. New experimental evidence supporting *

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

   A. Neural networks with structured distributed vectorial representations

   B. Computing (gradient) symbolic functions

   C. Encoding Ling$^{++}$ grammars

II. [redacted]

   B. Grammars

      1. Ling: Universal wellformedness constraints, *Con*

[redacted]edge?

   i. Argument from Universals* [*not* from Poverty of the Stimulus]

   ii. New experimental evidence supporting *

# Outline

The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$
   A. Neural networks with structured distributed vectorial representations
   B. Computing (gradient) symbolic functions
   C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$
   A. Representations
      1. Ling: Structural roles
      2. Ling$^{++}$: Gradient structural roles
   B. Grammars
      1. Ling: Universal wellformedness constraints, *Con*
      2. Ling$^{++}$: *SoftCon*, Bayesian universal biases — *later*
      3. Limiting the power of language learning; innate linguistic knowledge?
         i. Argument from Universals* [*not* from Poverty of the Stimulus]
         ii. New experimental evidence supporting *

# Universal 18 [Greenberg 63]

N, Adj word order in the world's languages

| | N-Adj | Adj-N |
|---|---|---|
| Num-N | 17% | 27% |
| N-Num | 52% | *4% |

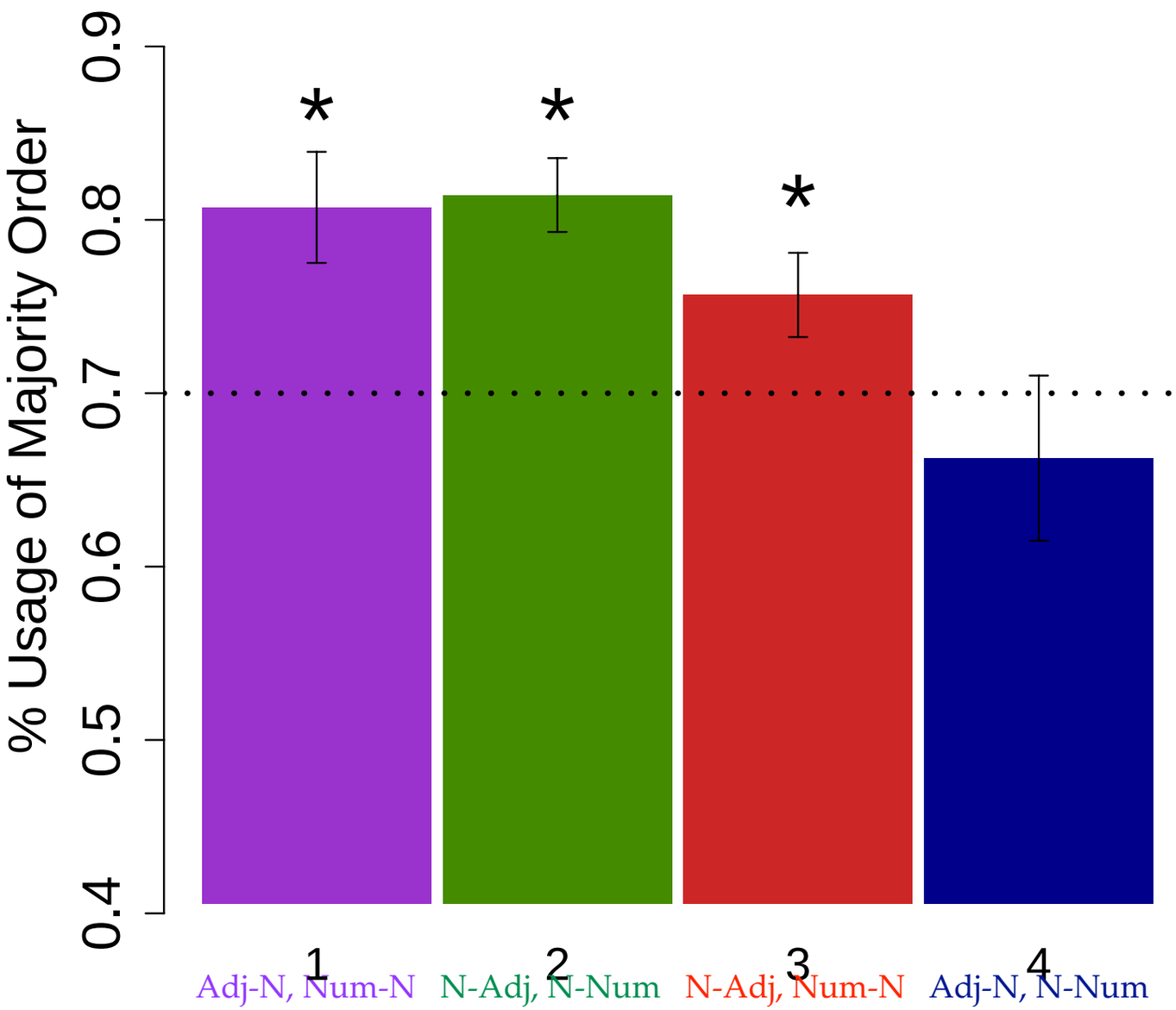
*Particular non-harmonic pattern Adj-N, N-Num is **disfavored**

*Substantive learning bias*

Dryer, M. (2008a). Order of adjective and noun. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The World Atlas of Language Structures Online*, Chapter 87. Munich, Germany: Max Planck Digital Library.

# Experimental results I: Adult group data

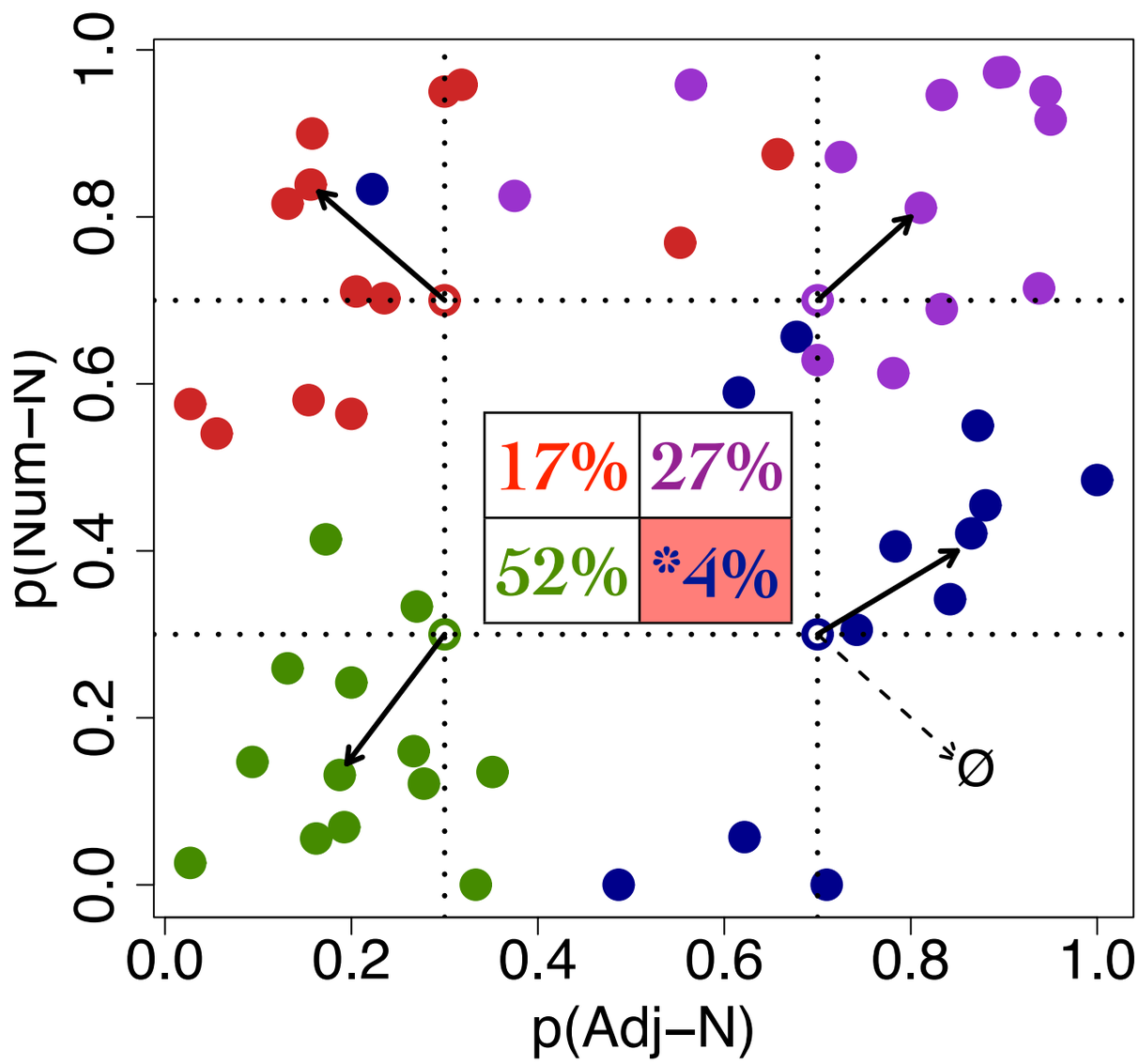
Do typological statistics correspond to any active on-line learning bias?




Learners of an artificial language, given two-word nonce-utterance examples (N with either Adj or Num), dominant order in each of 4 conditions = 70%.

# Experimental results I: Adult individual data

Do typological statistics correspond to any active on-line learning bias?




Learners of an artificial language, given two-word nonce-utterance examples (N with either Adj or Num), dominant order in each of 4 conditions = 70%.

# Partial replication in children

6-7-year-old kids:

Strongly adopt a harmonic order in all conditions

[tending to follow the dominant order of their condition's Adj and N]

Culbertson, Jennifer and Newport, Elissa L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition* 139: 71–82.

# Further evidence: U20

Greenberg (1968)'s U20: {Demonstrative Number Adjective Noun}:

Dem Num Adj N, N Adj Num Dem overwhelmingly most common:
"[Syntax/Semantics] Scope-isomorphic order"

Also N Dem Num Adj.

Trained with 2-word utterances, N + modifier, always N-initial.

Would learners favor English order among modifiers, or scope order?

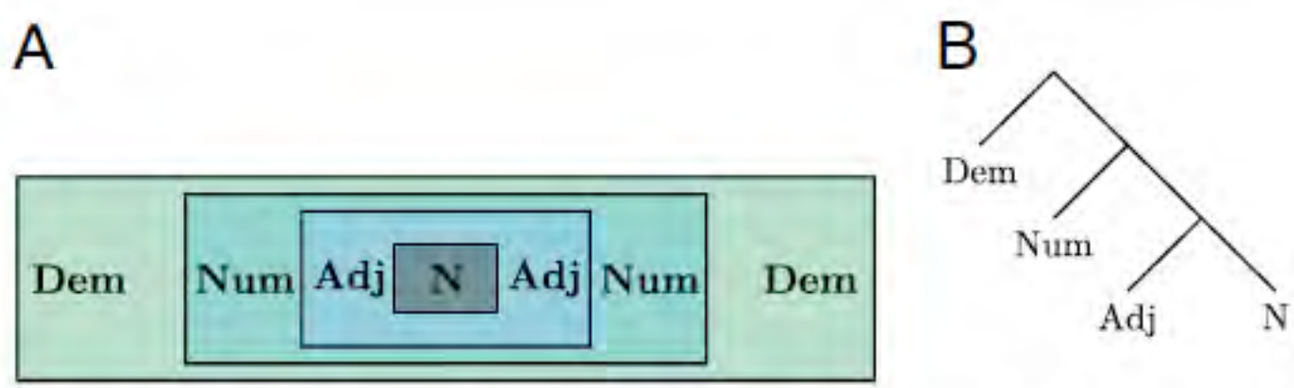



**Fig. 1.** (*A*) Scope relations within the nominal domain. (*B*) Hierarchical representation of scope. Dem(onstrative) takes outermost scope, over inner modifiers Num(eral) and Adj(ective).

Culbertson, Jennifer and Adger, David. (2014). Language learners privilege structured meaning over surface frequency. *PNAS* 111, 5842–5847.

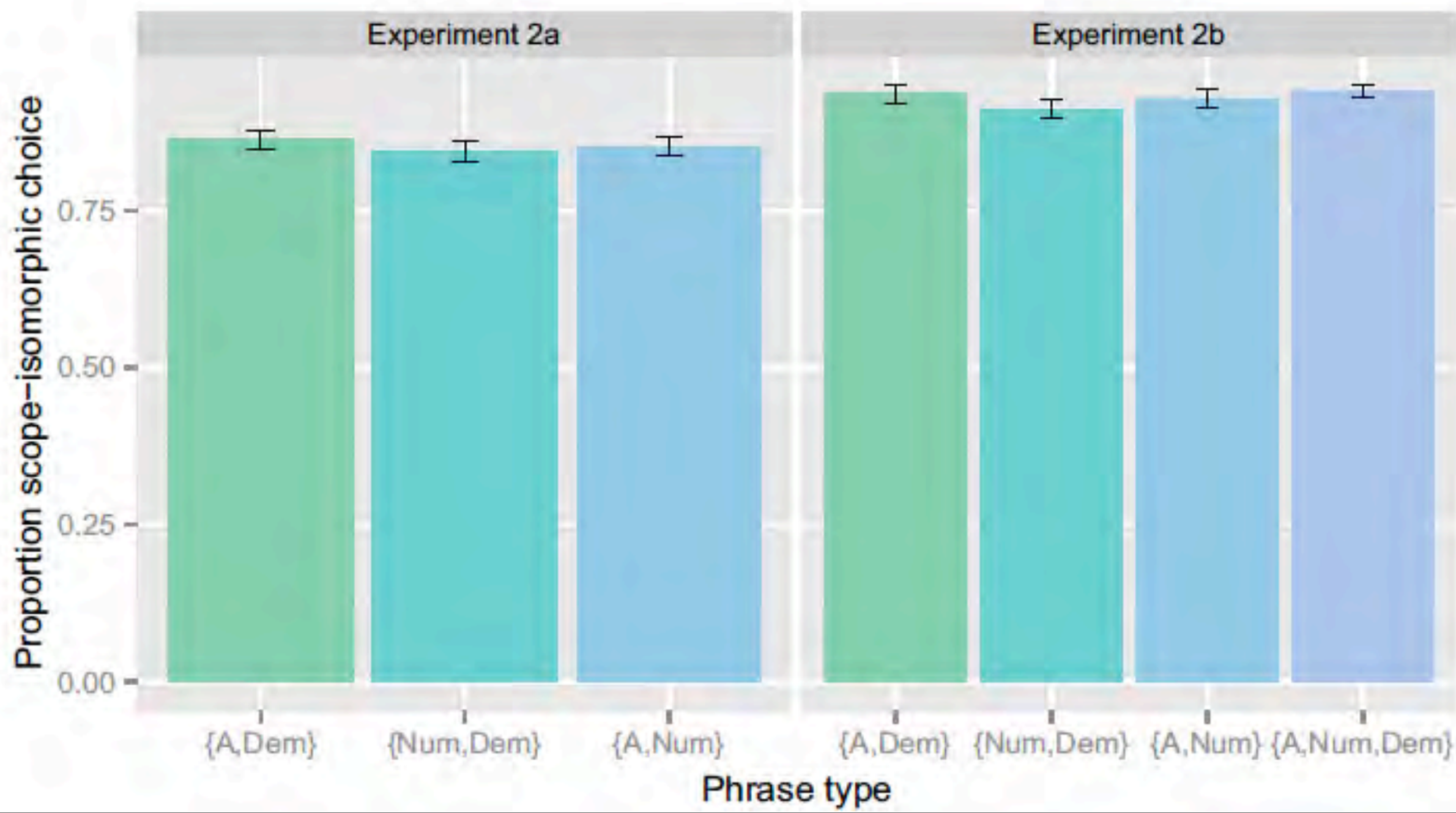
# Further evidence: U20

{Demonstrative Number Adjective Noun}:

Dem Num Adj N, N Adj Num Dem overwhelmingly most common:
"[Syntax/Semantics] Scope-isomorphic order"

Also N Dem Num Adj.

Trained with 2-word utterances, N + modifier, always N-initial



# Outline

The d

I.

ons

II.

B. Grammars

1. Ling: Universal wellformedness constraints, *Con*
2. Ling$^{++}$: *SoftCon,* Bayesian universal biases
3. Limiting the power of language learning; innate linguistic knowledge?
   i. Argument from Universals* [*not* from Poverty of the Stimulus]
   ii. New experimental evidence supporting *

# Constraint bias and statistical universal U18 [theoretical]

Culbertson, J., Smolensky, P. & Wilson, C. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, 5, 392–424.

To generate only harmonic orders (Num-Adj-N, N-Adj-Num), posit:

*Con* = {**HEAD-R**, **HEAD-L**}

To admit typologically less-preferred Num-N-Adj, add **NUM-L** to *SoftCon* but with a bias against 'use' (non-zero weight) of specific-head constraints: Gaussian prior, with 'strength' (precision) κ, that disfavors $G_{\{w_C\}}$ the more $w_{\text{NUM-L}}$ is > 0

To admit marginal Adj-N-Num, add **NUM-R** to *Con*, which incurs the specific-head penalty κ plus an additional penalty λ; result:

$$P_{\text{prior}}(G_{\{w_C\}}) \propto e^{-\kappa(w^2_{\text{NUM-L}} + w^2_{\text{NUM-R}})} e^{-\lambda w^2_{\text{NUM-R}}}$$

☞ **Not all constraints in *Con* are created equal: substantive bias**

Approach to explaining statistical typological patterns ('universals')

# Real-time learning bias in U18 [experimental]

Culbertson, J., Smolensky, P. & Wilson, C. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, 5, 392–424.

In addition to word-order bias, there is the regularization bias:

Culbertson, J., Smolensky, P. & Legendre, G. 2012. Learning biases predict a word order universal. *Cognition*, 122, 306–329.

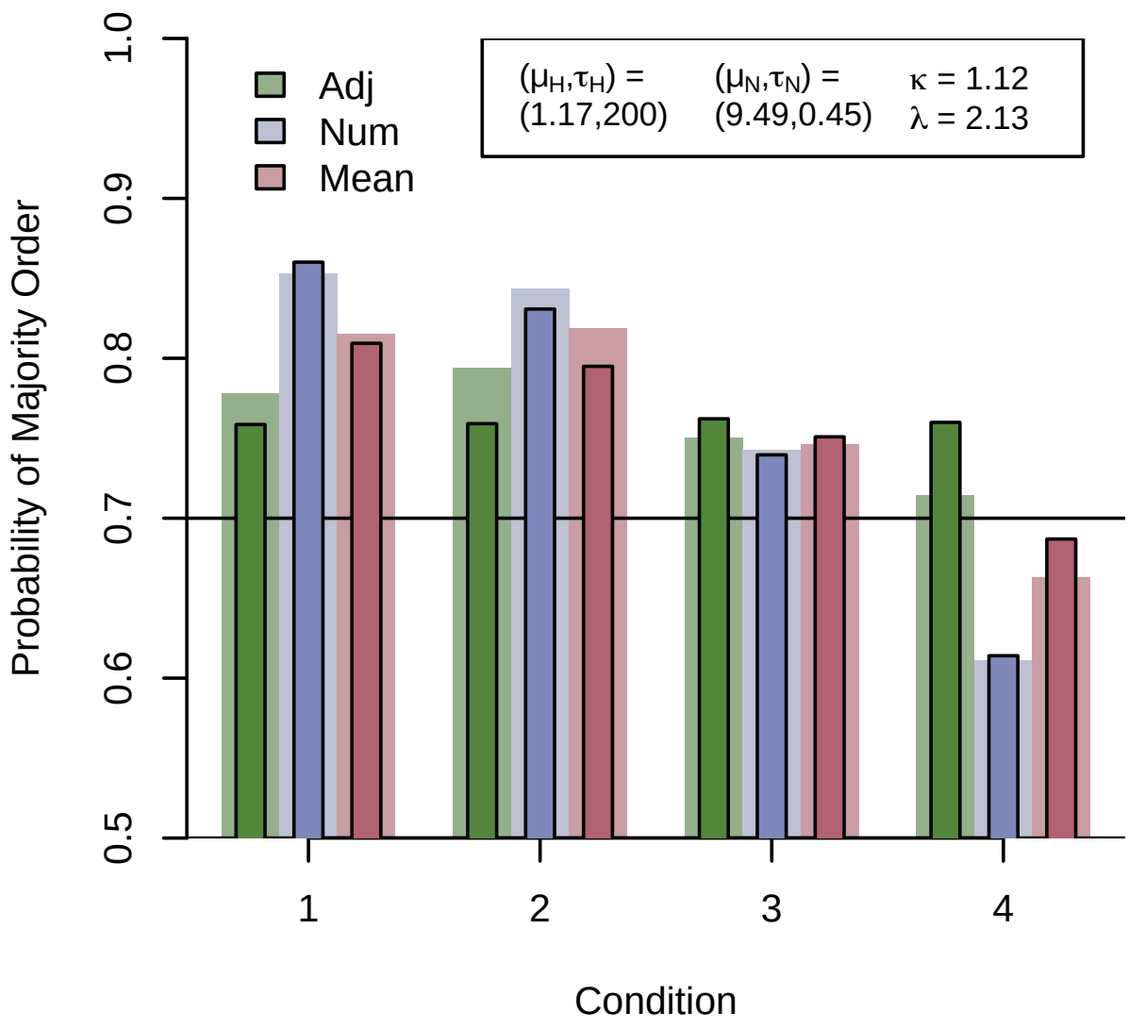
Prior also favors > 0 weight values (disprefers completely free variation resulting from all-0 weights)

- Gaussians with means greater than zero

Fitting the total prior distribution's parameters to the experimental data,
Bayes yields these predictions for the experiment:

# Real-time learning bias in U18 [experimental]

Culbertson, J., Smolensky, P. & Wilson, C. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, 5, 392–424.




**Lighter** bars (background) are **observed**; **darker** are **predicted**

- Preferred means pull weights further from zero than needed to match the input ⇒ regularization [line at 0.7 is rate in the input].
- Cost of doing so for specific-head constraint weights (NUM-L, NUM-R) impedes regularization.
- So get the most regularization for harmonic orders (conditions 1–2), then condition 3, and least for U18-violating condition 4 [see red 'Mean' bars]

# Outline

## The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

   A. Neural networks with structured distributed vectorial representations

   B. Computing (gradient) symbolic functions

   C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

   A. Representations

      1. Ling: Structural roles

      2. Ling$^{++}$: Gradient structural roles

   B. Grammars

      1. Ling: Universal wellformedness constraints, *Con*

      2. Ling$^{++}$: *SoftCon*, Bayesian universal biases

      3. Limiting the power of language learning; innate linguistic knowledge?

         i. Argument from Universals* [*not* from Poverty of the Stimulus]

         ii. New experimental evidence supporting *

# That's all folks!

Thanks to the organizers
and to you for your attention.

# Outline

## The death of linguistic theory? A sketch from Cognitive Science$^{++}$

I. NN$^{++}$: Implements Ling$^{++}$

   A. Neural networks with structured distributed vectorial representations

   B. Computing (gradient) symbolic functions

   C. Encoding Ling$^{++}$ grammars

II. Ling$^{++}$

   A. Representations

      1. Ling: Structural roles

      2. Ling$^{++}$: Gradient structural roles

   B. Grammars

      1. Ling: Universal wellformedness constraints, *Con*

      2. Ling$^{++}$: *SoftCon*, Bayesian universal biases

      3. Limiting the power of language learning; innate linguistic knowledge?

         i. Argument from Universals* [*not* from Poverty of the Stimulus]

         ii. New experimental evidence supporting *