

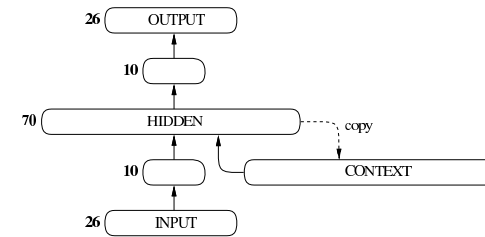
## Traditional view of language

- Language knowledge largely consists of an explicit **grammar** that determines what sentences are part of a language
  - Isolated from other types of knowledge—pragmatic, semantic, lexical(?)
- Language learning involves identifying the single, correct grammar of the language
- Grammar induction is underconstrained by the linguistic input given lack of explicit negative evidence
  - Impossible under near-arbitrary positive-only presentation (Gold, 1967)
- Language learning requires strong **innate linguistic constraints** to narrow the range of possible grammars considered

1

## A connectionist approach to sentence processing

Elman (1991, *Machine Learning*)



S → NP VI . | NP VT NP .  
 NP → N | N RC  
 RC → who VI | who VT NP | who NP VT  
 N → boy | girl | cat | dog | Mary | John |  
 boys | girls | cats | dogs  
 VI → barks | sings | walks | bites | eats |  
 bark | sing | walk | bite | eat  
 VT → chases | feeds | walks | bites | eats |  
 chase | feed | walk | bite | eat

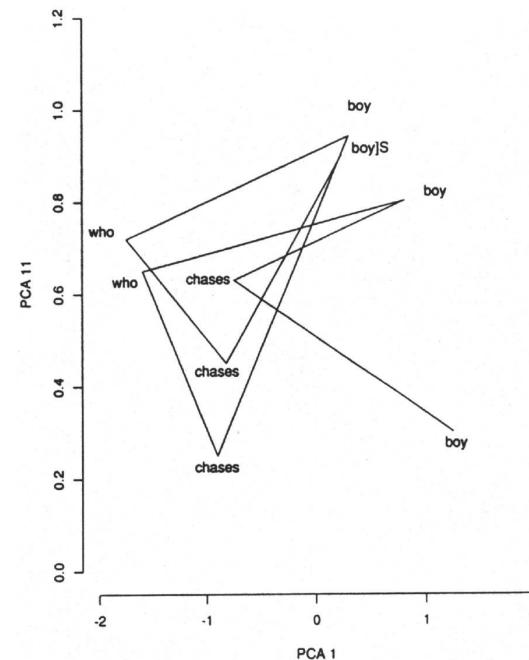
- Simple recurrent network trained to predict next word in English-like sentences
  - Context-free grammar, number agreement, variable verb argument structure, multiple levels of embedding
  - 75% of sentences had at least one relative clause; average length of 6 words.
  - e.g., Girls who cat who lives chases walk dog who feeds girl who cats walk .
- After 20 sweeps through 4 sets of 10,000 sentences, mean absolute error for new set of 10,000 sentences was **0.177** (cf. initial: 12.45; uniform: 1.92)

3

## Statistical view of language

- Language environment has rich **distributional regularities**
  - May not provide correction but is certainly not *adversarial* (cf. Gold, 1967)
- Language learning requires only that knowledge across speakers converges sufficiently to support effective communication
- No sharp division between linguistic vs. extra-linguistic knowledge
  - Effectiveness of learning depend both on the structure of the input and on **existing knowledge** (linguistic and extra-linguistic)
- Distributional information can provide *implicit* negative evidence
  - Example: **implicit prediction** of upcoming input
  - Sufficient for language learning when combined with *domain-general* biases

2



Boy chases boy who chases  
 boy who chases boy .

### Principal Components Analysis (PCA) of network's internal representations

- Largest amount of variance (PC-1) reflects **word class** (noun, verb, function word)
- Separate dimension of variation (PC-11) encodes **syntactic role** (agent/patient) for nouns and level of embedding for verbs

4

## The importance of “starting small”

### Elman (1993, *Cognition*)

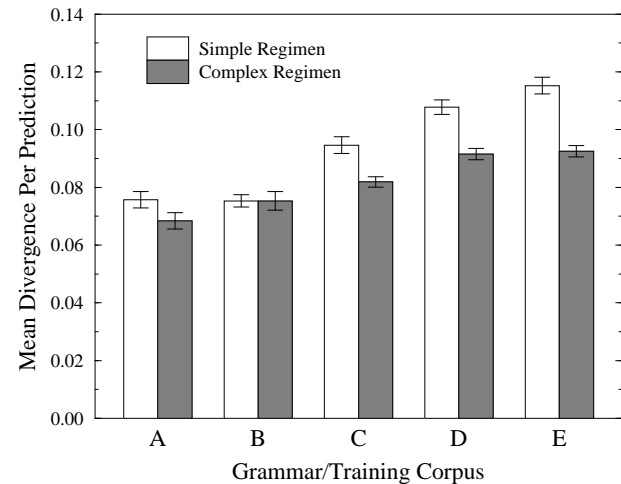
- Training was successful **only** when “starting small”
  - Trained on only simple sentences before gradually introducing embedded sentences
  - Trained on full language but with initially limited memory that gradually improved
- Consistent with Newport’s (1990, *Cog. Sci.*) “**less is more**” hypothesis
  - Child language acquisition is helped rather than hindered by maturational limits on cognitive resources

**Alternative Hypothesis:** Need to start small was exaggerated by lack of important *soft* constraints inherent in natural language

- SRN’s learn long-distant dependencies better when intervening material is partially correlated with distant information (Cleeremans et al., 1989, *Neural Comp.*)
- Soft semantic constraints—distributional biases on noun-verb co-occurrences across clauses—provide such correlations

5

## Results: Prediction error



- **Disadvantage** for “starting small” that increases with reliability of semantic constraints

7

## Simulation 1: Semantic constraints

### Rohde and Plaut (1999, *Cognition*)

- Replication of Elman (1993) simulation with addition of constraints on verb arguments
- Parametric variation of reliability of semantic constraints across clauses (A = none, ..., E = 100% reliable)
- Minor improvements in technical aspects of simulation (e.g., error function, initialization)
- Compared two **training regimens**:
  - Complex: Trained on full language throughout  
25 sweeps through 10,000 sentences (75% complex)
  - Simple: Trained incrementally
    - 5 sweeps on only simple sentences
    - 5 sweeps with 25% complex sentences
    - 5 sweeps with 50% complex sentences
    - 10 sweeps with 75% complex sentences

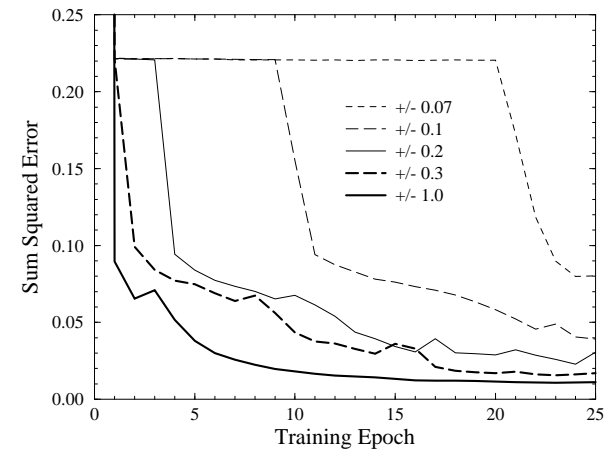
Verb	Intransitive Subjects	Transitive Subjects	Objects if Transitive
chase	–	any	any
feed	–	human	animal
bite	animal	animal	any
walk	any	human	only dog
eat	any	animal	human
bark	only dog	–	–
sing	human or cat	–	–

6

## Relation to Elman’s (1993) results

Exact replication, varying **magnitudes of initial random weights**

- Simulation 1 used  $\pm 1.0$ ; Elman used  $\pm 0.001$



- Very small initial weights prevent effective accumulation of error derivatives

8

## Simulation 2: Native vs. late bilingual acquisition

### Languages

- **English:** Analogous to language from Simulation 1
- **German:** German vocabulary (“hund” vs. “dog”), gender marking, case-marking in masculine, verb-final relative clauses
- Phoneme-based input and output representations

### Training Conditions

- **Monolingual:** Trained on either English or German
  - 6 million sentence presentations sampled from corpus of 50,000 sentences
- **Native Bilingual:** Trained on both English and German (50/50)
  - 6 million sentence presentations sampled from two corpora of 50,000 sentences each
  - Language selected randomly every 50 sentences
- **Late Bilingual:** Monolingual training followed by bilingual training
- Output unit error derivatives scaled by unit activation (“pseudo-Hebbian”)

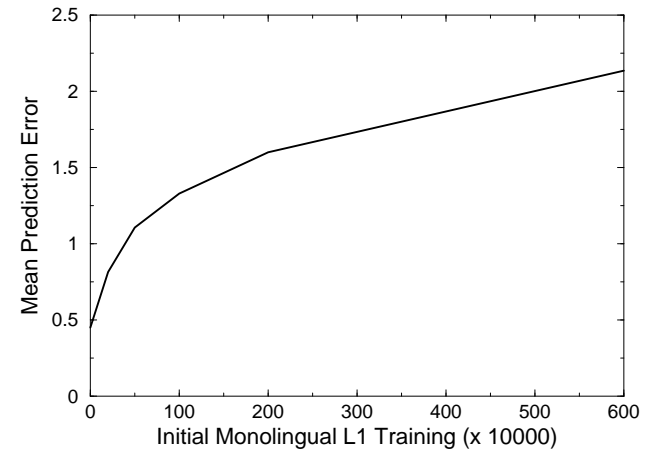
### Testing

- Late Bilingual tested on L2 (new sample of 5,000 sentences)
- All results counterbalanced for English vs. German

9

## Results: Early-bilingual acquisition

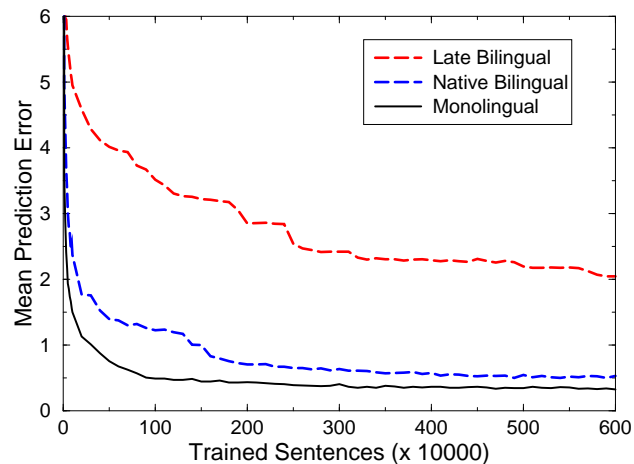
Final Late-Bilingual Performance on L2



- Even relatively brief exposure to monolingual L1 impacts subsequent L2 acquisition

11

## Results: Acquisition



- Initial monolingual training impedes subsequent bilingual acquisition
- Native bilingual acquisition is only slightly worse than monolingual acquisition

10

## Simulations 1 & 2: Conclusions

- Introducing soft semantic constraints aids learning of pseudo-natural languages by simple recurrent networks
  - No need to manipulate training environment or cognitive resources
  - Networks *inherently* learn local dependences before longer distance ones
- Critical-period effects may reflect **entrenchment** of representations that have learned to perform other tasks (including other languages)
  - No need to introduce additional maturational assumptions (e.g., “less is more”)

12

## Sentence comprehension

### Traditional perspective

- Linguistic knowledge as grammar, separate from semantic/pragmatic influences on performance (Chomsky, 1957)
- Psychological models with initial syntactic parse that is insensitive to lexical/semantic constraints (Ferreira & Clifton, 1986; Frazier, 1986)

### Problem: Interdependence of syntax and semantics

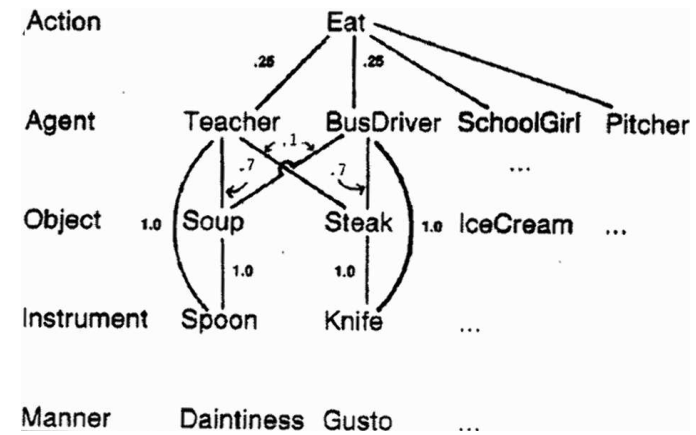
1. The spy saw the policeman with a revolver
2. The spy saw the policeman with binoculars
3. The bird saw the birdwatcher with binoculars
4. The pitcher threw the ball
5. The container held the apples/cola
7. The boy spread the jelly on the bread

### Alternative: Constraint satisfaction

Sentence comprehension involves integrating multiple sources of information (both semantic and syntactic) to construct the most plausible interpretation of a sentence (MacDonald et al., 1994; Seidenberg, 1997; Tanenhaus & Trueswell, 1995)

13

## Event structures

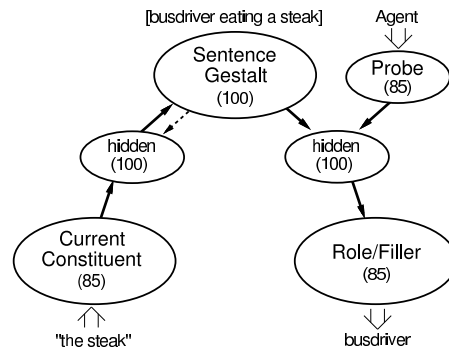


- 14 active frames, 4 passive frames, 9 thematic roles
- Total of 120 possible events (varying in likelihood)

15

## Sentence Gestalt Model (St. John & McClelland, 1990)

- Trained to generate thematic role assignments of event described by single-clause sentence
- Sentence constituents ( $\approx$  phrases) presented one at a time
- After each constituent, network updates internal representation of sentence meaning ("Sentence Gestalt")
- Current Sentence Gestalt trained to generate *full set* of role/filler pairs (by successive "probes")
  - Must **predict** information based on partial input and learned experience, but must **revise** if incorrect



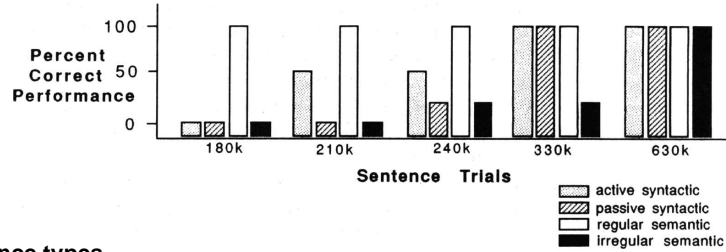
14

## Sentence generation

- Given a specific event, probabilistic choices of
  - Which thematic roles are explicitly mentioned
  - What word describes each constituent
  - Active/passive voice
- Example: busdriver eating steak with knife
  - THE-ADULT ATE THE-FOOD WITH-A-UTENSIL
  - THE-STEAK WAS-CONSUMED-BY THE-PERSON
  - SOMEONE ATE SOMETHING
- Total of 22,645 sentence-event pairs

16

## Acquisition



### Sentence types

- Active syntactic:** THE BUSDRIVER KISSED THE TEACHER
- Passive syntactic:** THE TEACHER WAS KISSED BY THE BUSDRIVER
- Regular semantic:** THE BUSDRIVER ATE THE STEAK
- Irregular semantic:** THE BUSDRIVER ATE THE SOUP

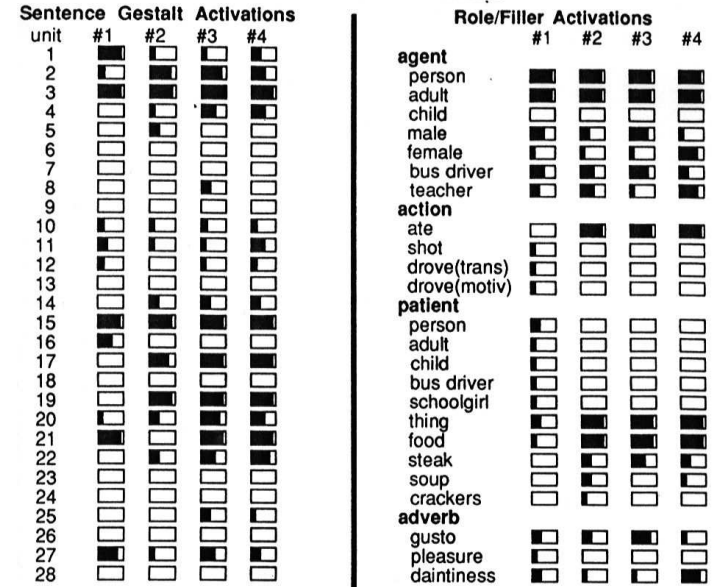
### Results

- Active voice learned before passive voice
- Syntactic constraints learned before semantic constraints
- Final network tested on 55 randomly generated unambiguous sentences
  - Correct on 1699/1710 (99.4%) of role/filler assignments

17

## Online updating and backtracking

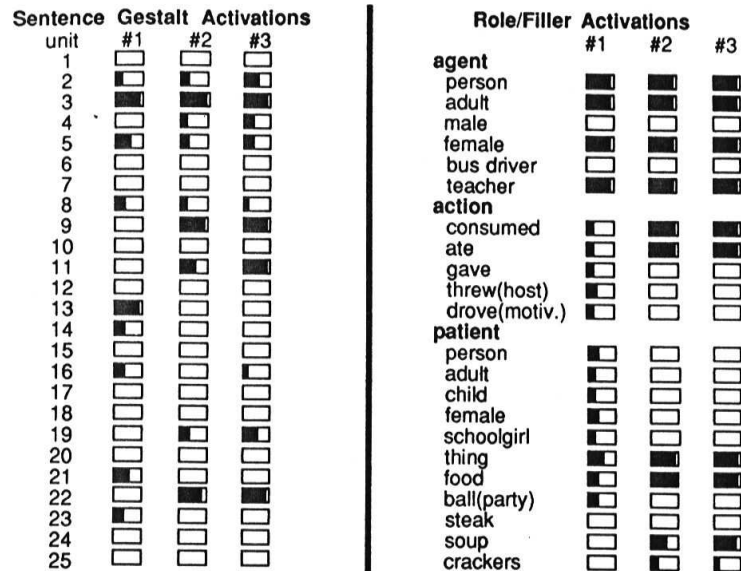
The adult ate the steak with daintiness.



19

## Implied constituents

The teacher ate the soup.

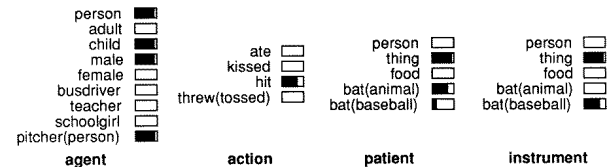


18

## Semantic-syntactic interactions

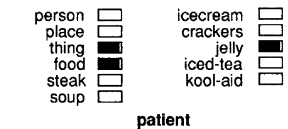
### Lexical ambiguity

The pitcher hit the bat with the bat.

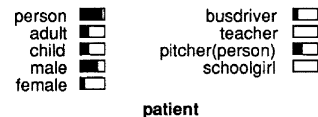


### Concept instantiation

The schoolgirl spread something with a knife.



The teacher kissed someone.

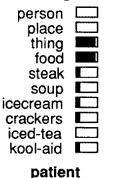


### Implied constituents

The teacher ate the soup.

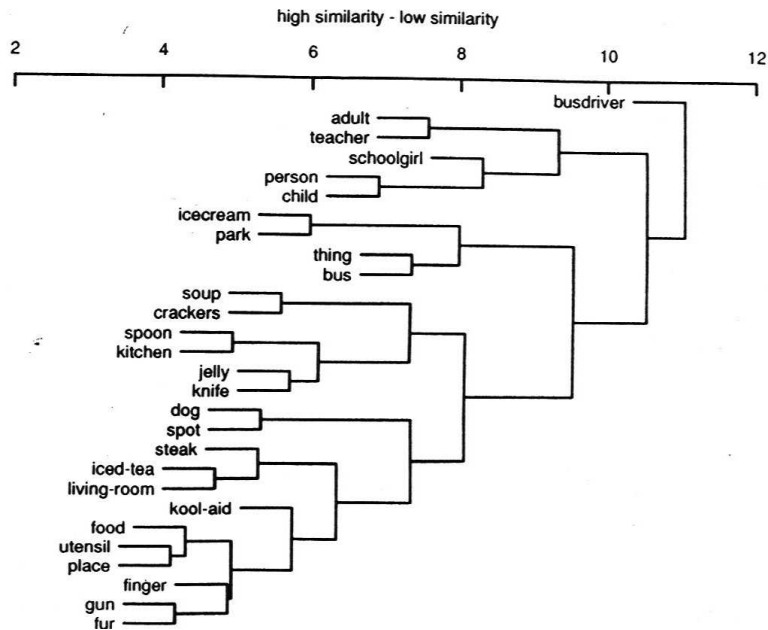


The schoolgirl ate.



20

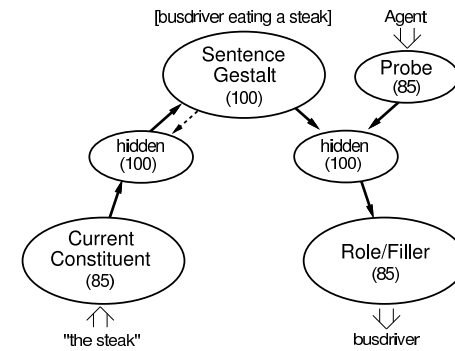
## Noun similarities



21

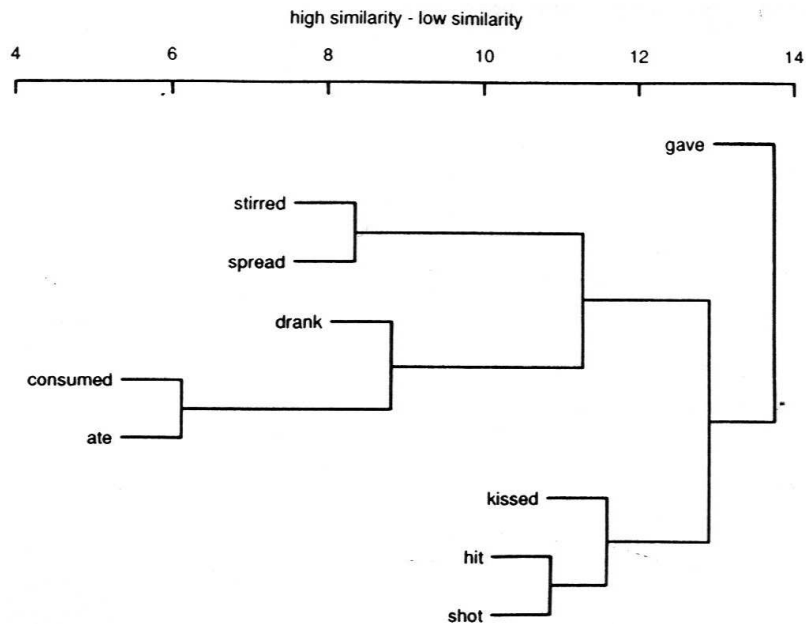
## Summary: St. John and McClelland (1990)

- Syntactic and semantic constraints can be learned and brought to bear in an integrated fashion to perform online sentence comprehension
- Approach stands in sharp contrast to linguistic and psycholinguistic theories espousing a clear separation of grammar from the rest of cognition



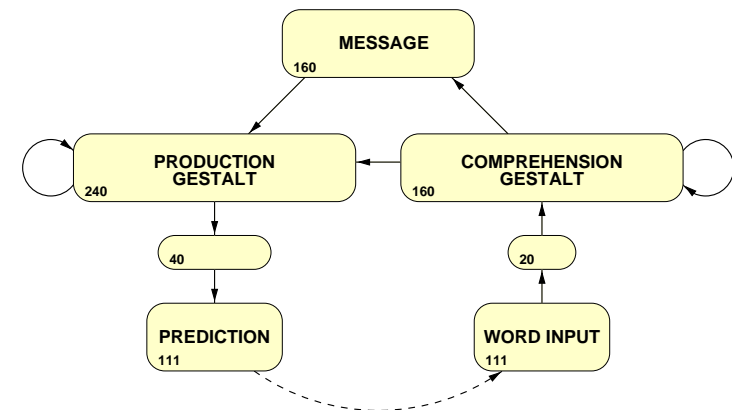
23

## Verb similarities



22

## Sentence comprehension and production (Rohde)



- Extends approach of Sentence Gestalt model to multi-clause sentences
- Trained to generate learned "message" representation and to predict successive words in sentences when given varying degrees of prior context

24

## Training language

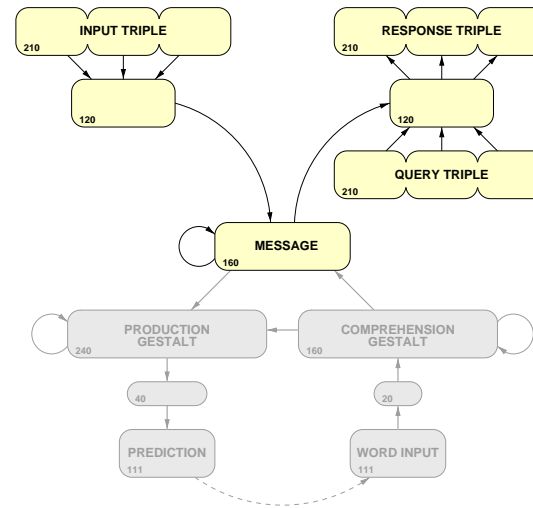
- Multiple verb tenses
  - e.g., ran, was running, runs, is running, will run, will be running
- Passives
- Relative clauses (normal and reduced)
- Prepositional phrases
- Dative shift
  - e.g., gave flowers to the girl, gave the girl flowers
- Singular, plural, and mass nouns
- 12 noun stems, 12 verb stems, 6 adjectives, 6 adverbs

### Examples

- The boy drove.
- An apple will be stolen by the dog.
- Mean cops give John the dog that was eating some food.
- John who is being chased by the fast cars is stealing an apple which was had with pleasure.

25

## Message encoder



### Methods

- Triples presented in sequence
- For each triple, all presented triples queried three ways (given two elements, generate third)
- Trained on 2 million sentence meanings

### Results

- Full language
  - Triples correct: 91.9%
  - Components correct: 97.2%
  - Units correct: 99.9%
- Reduced language ( $\leq 10$  words):
  - Triples correct: 99.9%

27

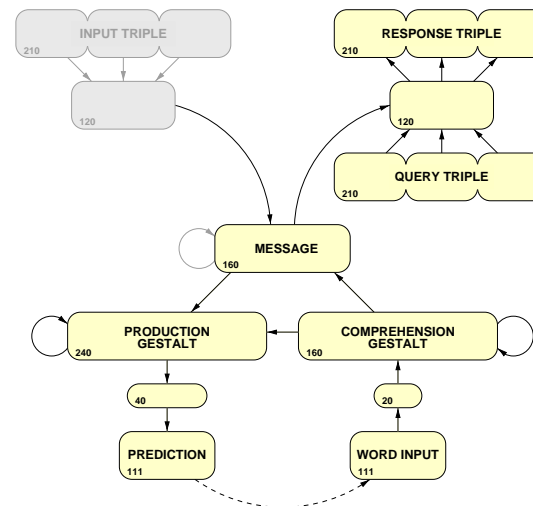
## Encoding messages with triples

The boy who is being chased by the fast dogs stole some apples in the park .

<b>steal</b> action transitive instantaneous past	<b>boy</b> entity singular definite animate human	<b>apple</b> entity plural indefinite
<b>chase</b> action passive ongoing present which	<b>boy</b> entity singular definite animate human	<b>dog</b> entity plural definite animate
<b>property</b>	<b>dog</b> entity plural definite animate	<b>fast</b> quality
<b>location</b>	<b>steal</b> action transitive instantaneous past	<b>park</b> entity singular definite

26

## Training: Comprehension (and prediction)



### Methods

- No context on half of the trials
- Context was *weak clamped* (25% strength) on other half
- Initial state of message layer clamped with varying strength

### Results

- Correct query responses with comprehended message:
  - Without context: 96.1%
  - With context: 97.9%

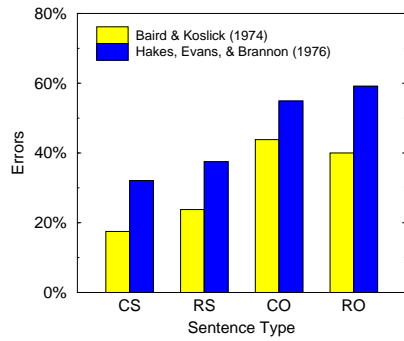
28

## Testing: Comprehension of relative clauses

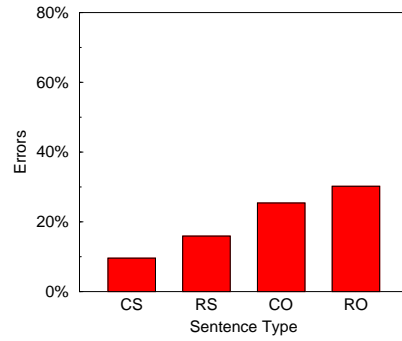
Single embedding: Center- vs. Right-branching; Subject- vs. Object-relative

- CS: A dog [who chased John] ate apples.
- RS: John chased a dog [who ate apples].
- CO: A dog [who John chased] ate apples.
- RO: John ate a dog [who the apples chased].

Empirical Data

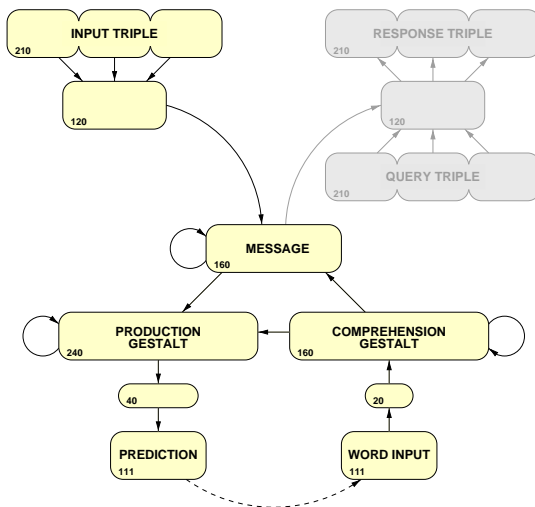


Model



29

## Testing: Production



### Methods

- Message initialized to correct value and weak clamped (25% strength)
- Most actively predicted word selected for production
- **No explicit training**

### Results

- 86.5% of sentences correctly produced.

30