Language Acquisition
Fall 2010/Winter 2011

Lexical Categories

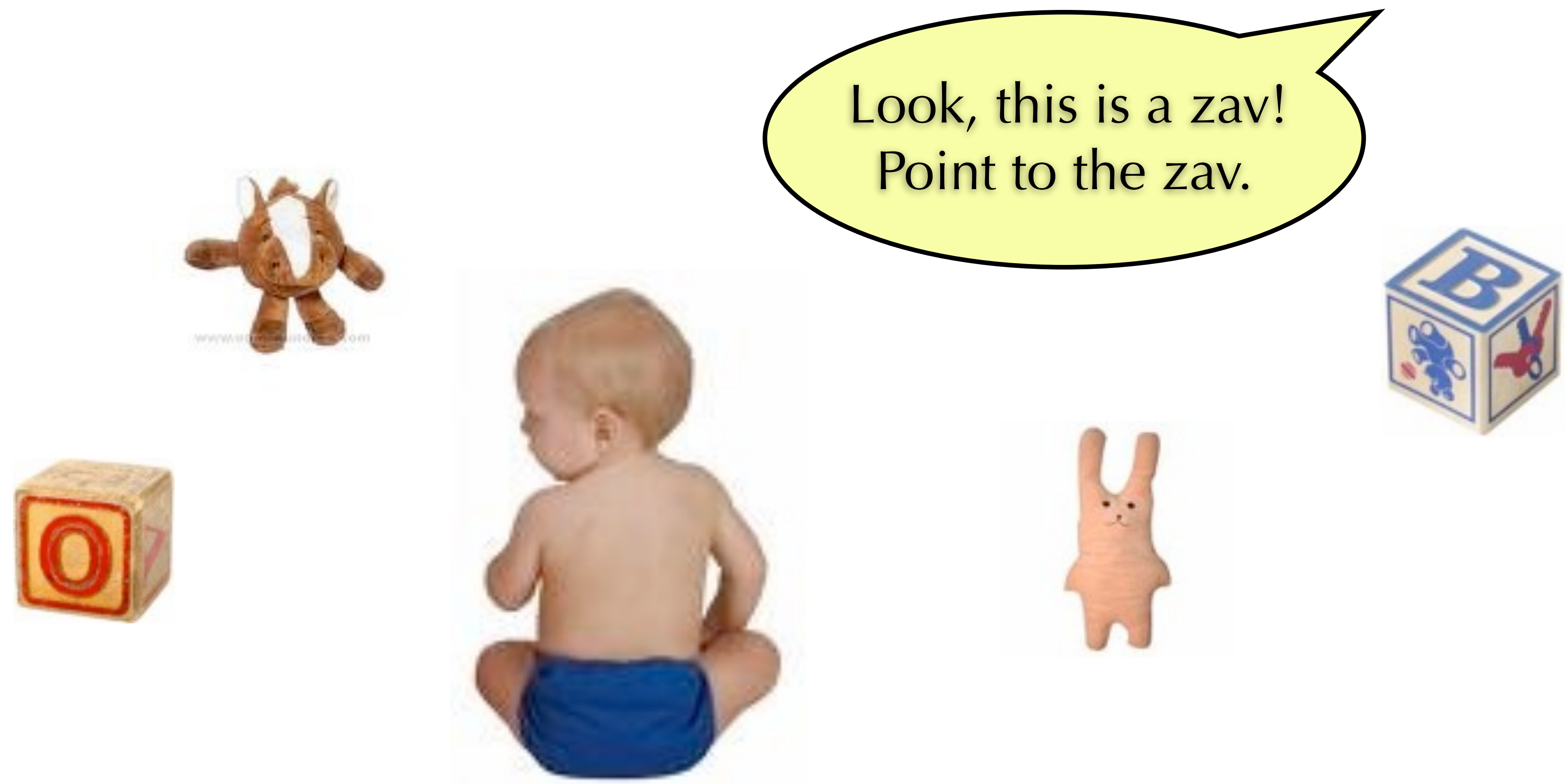
Afra Alishahi, Heiner Drenhaus

Computational Linguistics and Phonetics
Saarland University

Children’s Sensitivity to Lexical Categories

- Gelman & Taylor’84: 2-year-olds treat names not followed by a determiner (e.g. “Zav”) as a proper name, and interpret them as individuals (e.g., the animal-like toy).

Children's Sensitivity to Lexical Categories



- Gelman & Taylor'84: 2-year-olds treat names followed by a determiner (e.g. "the zav") as a common name, and interpret them as category members (e.g., the block-like toy).

Challenges of Learning Lexical Categories

- Children form lexical categories gradually and over time
  - Nouns and verb categories are learned by age two, but adjectives are not learned until age six
- Child language acquisition is bounded by memory and processing limitations
  - Child category learning is unsupervised and incremental
  - Highly extensive processing of data is cognitively implausible
- Natural language categories are not clear cut
  - Many words are ambiguous and belong to more than one category
  - Many words appear in the input very rarely

Information Cues

- Children might use different information cues for learning lexical categories
  - perceptual cues (phonological and morphological features)
  - semantic properties of the words
  - distributional properties of the local context each word appears in
- Distributional context is a reliable cue
  - Analysis of child-directed speech shows abundance of consistent contextual patterns (Redington et al., 1998; Mintz, 2003)
  - Several computational models have used distributional context to induce intuitive lexical categories (e.g. Schutze 1993, Clark 2000)

Computational Models of Lexical Category Induction

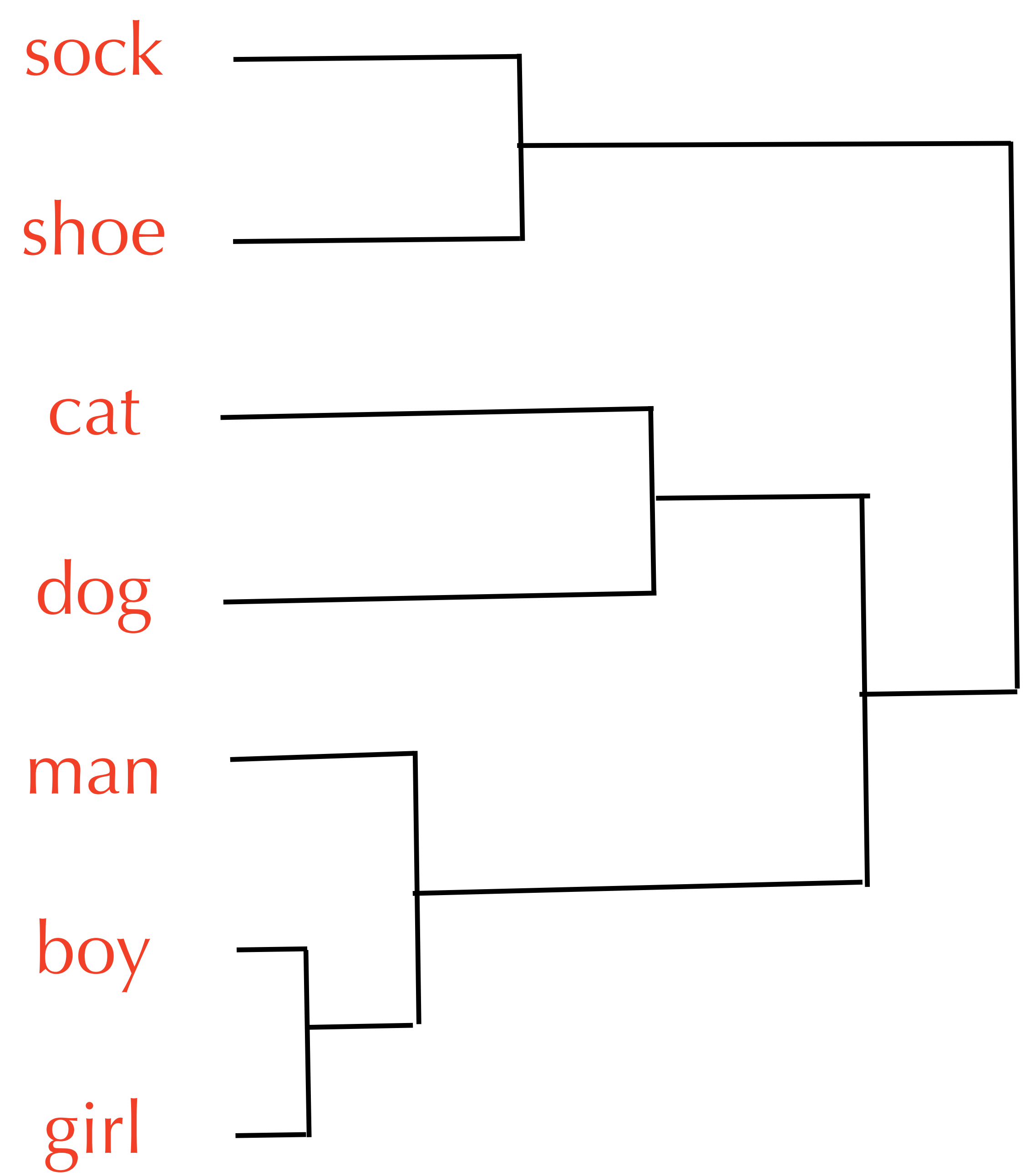
- The majority of the existing models categorize word types in an iterative, batch process
  - E.g. Brown'92, Schütze'93, Redington et al'98
- Incremental clustering models
  - Cartwright & Brent'97
    - Use word groups to extract templates from sentences, then use a MDL approach to merge word groups together
    - Evaluated on artificially generated input
  - Parisien et al'08
    - A Bayesian clustering model with a bootstrapping module; categories are revised periodically
    - Very sensitive to context features, and computationally extensive

Computational Models of Lexical Category Induction

- Hierarchical clustering

  [e.g., Schutze'93, Redington et al'98]

  - Start from a cluster per word
  - merge two most similar clusters in each iteration

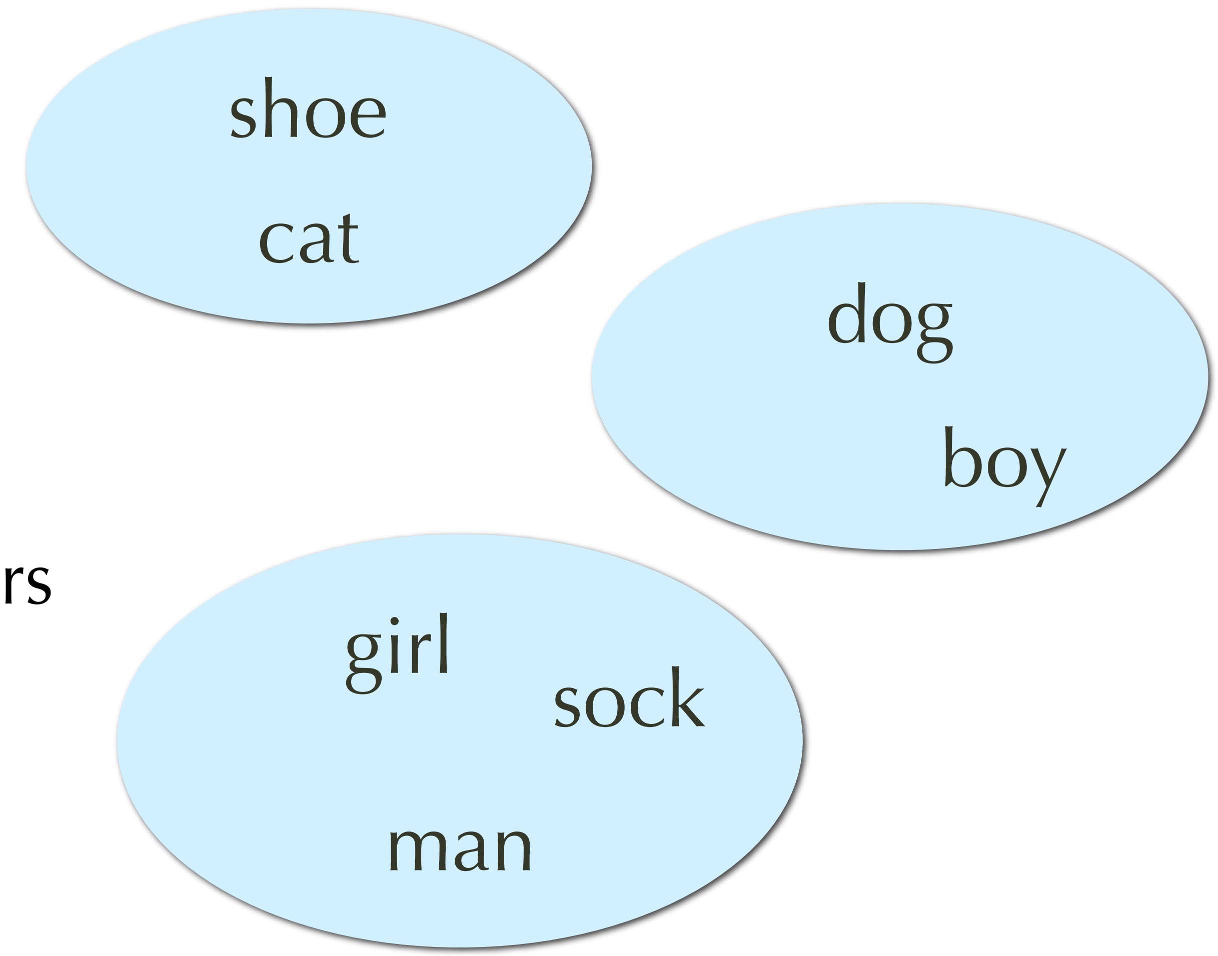


Computational Models of Lexical Category Induction

- Cluster optimization

  [e.g., Brown'92, Clark'00]

  - partition vocabulary into non-overlapping clusters
  - optimize clusters according to an information theoretic measure

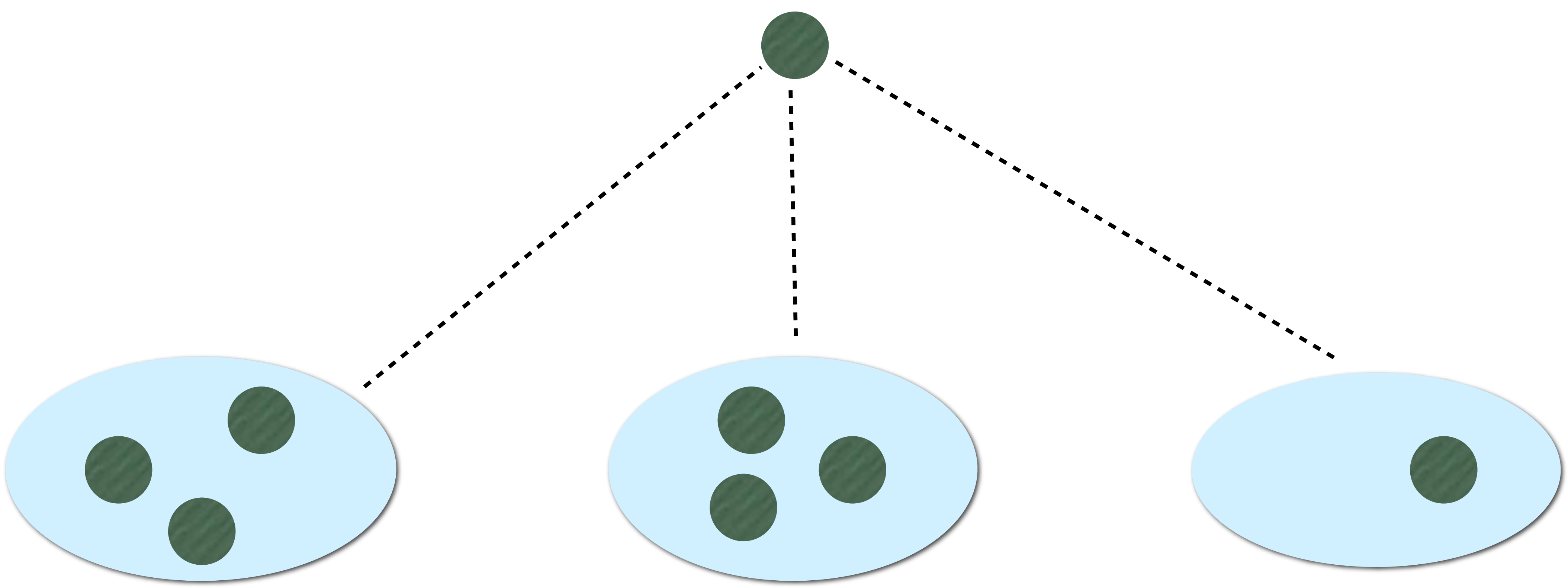


Computational Models of Lexical Category Induction

- Incremental clustering models

  (Cartwright & Brent'97, Parisien et al'08, Chrupala & Alishahi'10)

  - Each word usage is processed one at a time
  - It is added to the most similar existing cluster, or a new cluster is created



Case Study: Parisien et al. (2008)

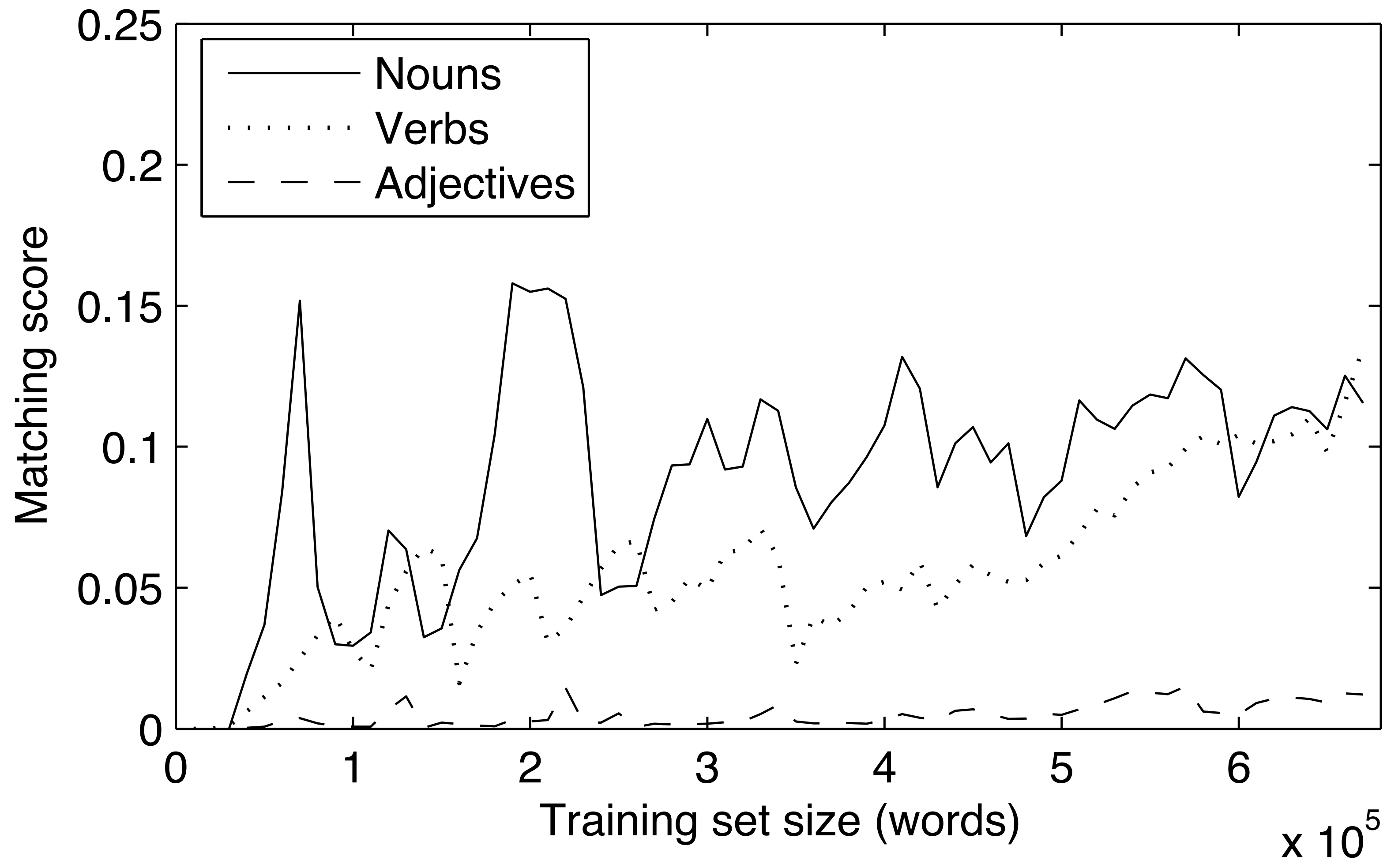
- A Bayesian model of lexical category induction
  - Word usages are categorized based on similarity of their content and context to the existing categories

-2	-1	0	1	2
"want	*to*	***put***	*them*	*on"*

  - Best cluster is selected by maximizing the conditional probability of each cluster for the current usage:

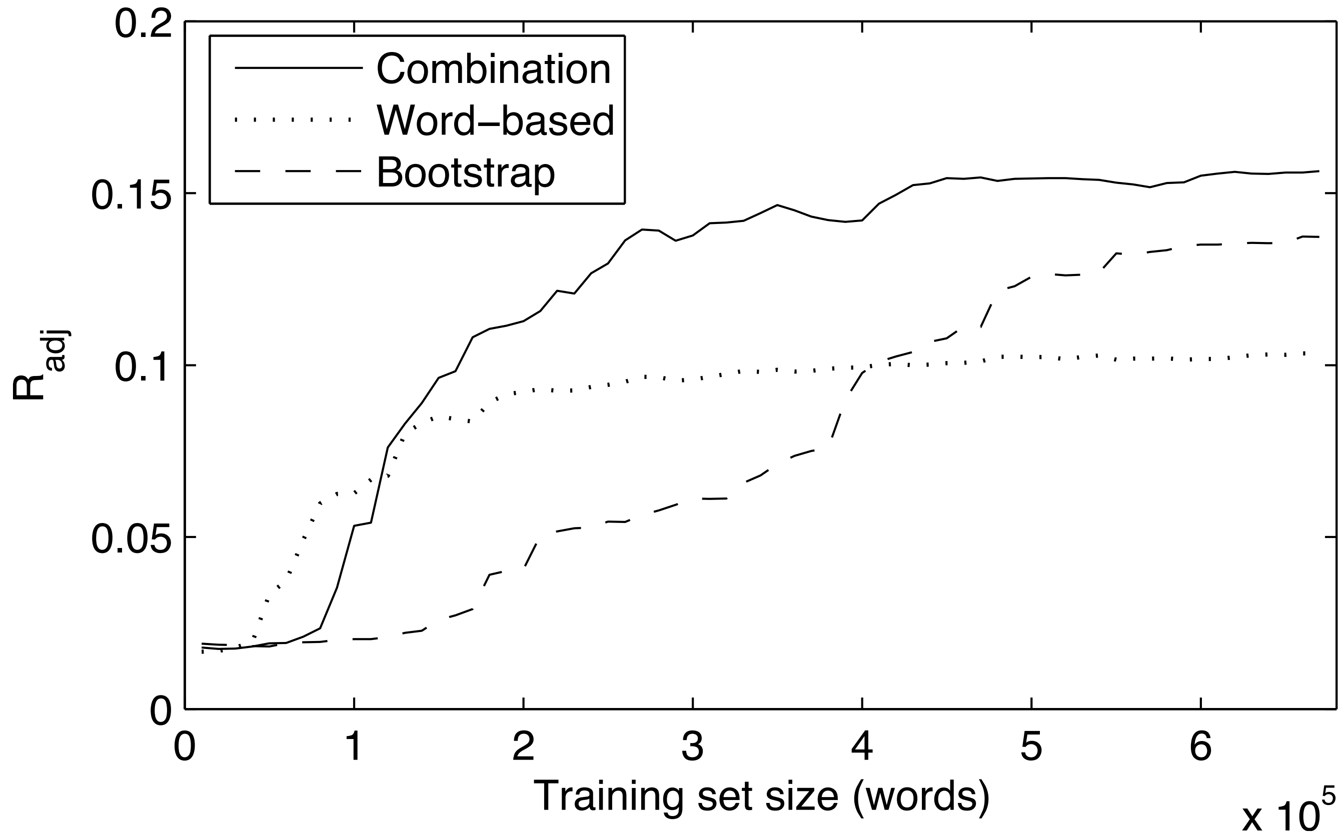
$$\mathrm{BestCluster}(F) = \operatorname*{argmax}_k P(k|F) = \frac{P(k)P(F|k)}{P(F)} \propto P(k)P(F|k)$$

Case Study: Parisien et al. (2008)



- The model replicates the order of acquisition of different categories as observed in children

Case Study: Parisien et al. (2008)



- The model predicts that using previous category labels will improve the overall performance

Case Study: Alishahi & Chrupala (2009)

- An incremental clustering algorithm:

```
1. Each word usage is put into a new category

2. The most similar category to the new one is found

   I. If the similarity is above a certain threshold
      θw, the two clusters are merged

   II.The most similar category to the newly merged
      one is found

      i. If the similarity is above a certain
         threshold θc, the two clusters are merged
```

Representation of Word Categories

- Word usage: a vector of content and context features:

-2 -1 0 1 2

"want to put them on"

-2=want	-1=to	0=put	1=them	2=on
1	1	1	1	1

- A lexical category is a cluster of word usages
  - Category: the mean of the distribution vectors of its members

-2=want	-2=have	-1=to	0=go	0=sit	0=show	0=send	1=it	...
0.25	0.75	1	0.25	0.25	0.25	0.25	0.5	...

  - The similarity between two categories: dot product of their vectors

Evaluation of the Acquired Categories

- Most of the models treat POS tags as gold-standard
  - Evaluate learned categories based on how well they match POS categories
- Instead, they use the categories in a variety of tasks
  - Word prediction from context
  - Inferring semantic properties of novel words based on the context they appear in
- They compare the performance in each task against a POS-based implementation of the same task

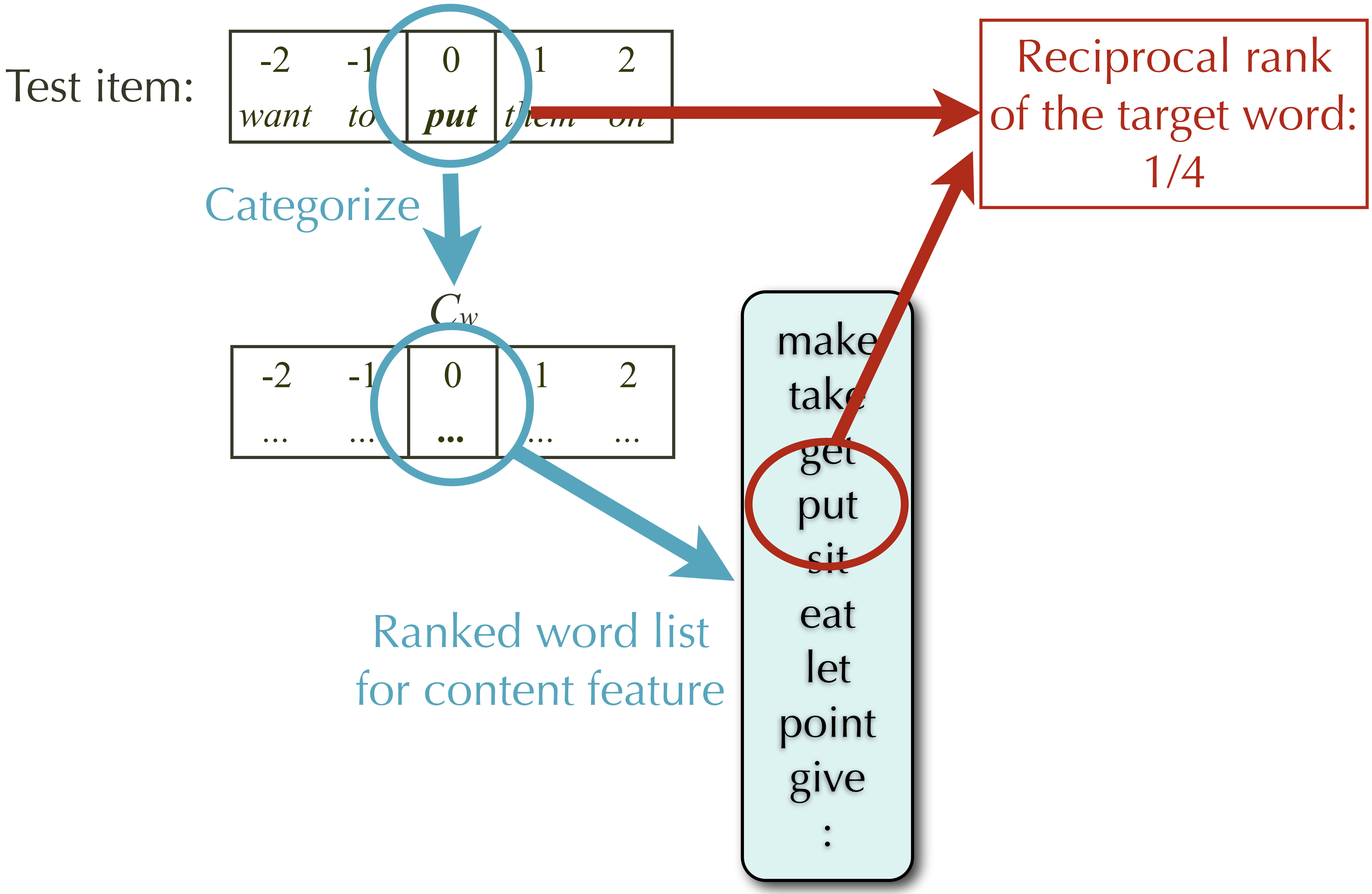
Word Prediction

She slowly --- the road

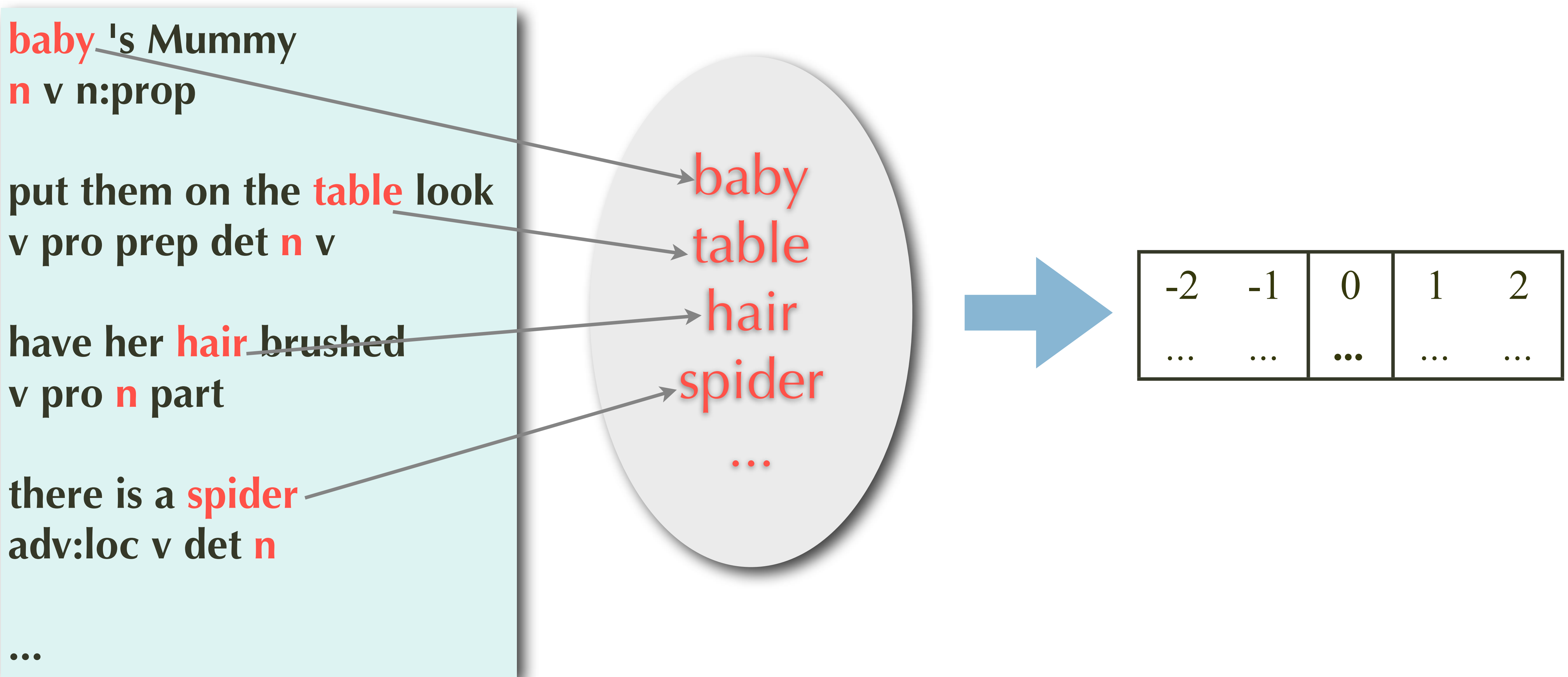
I had --- for lunch

- Task: predicting a missing (target) word based on its context
  - This task is non-deterministic (i.e. it can have many answers), but the context can significantly limit the choices
- Human subjects have shown to be remarkably accurate at using context for guessing target words (Gleitman'90, Lesher'02)

Word Prediction Using Categories



Word Prediction - POS Categories



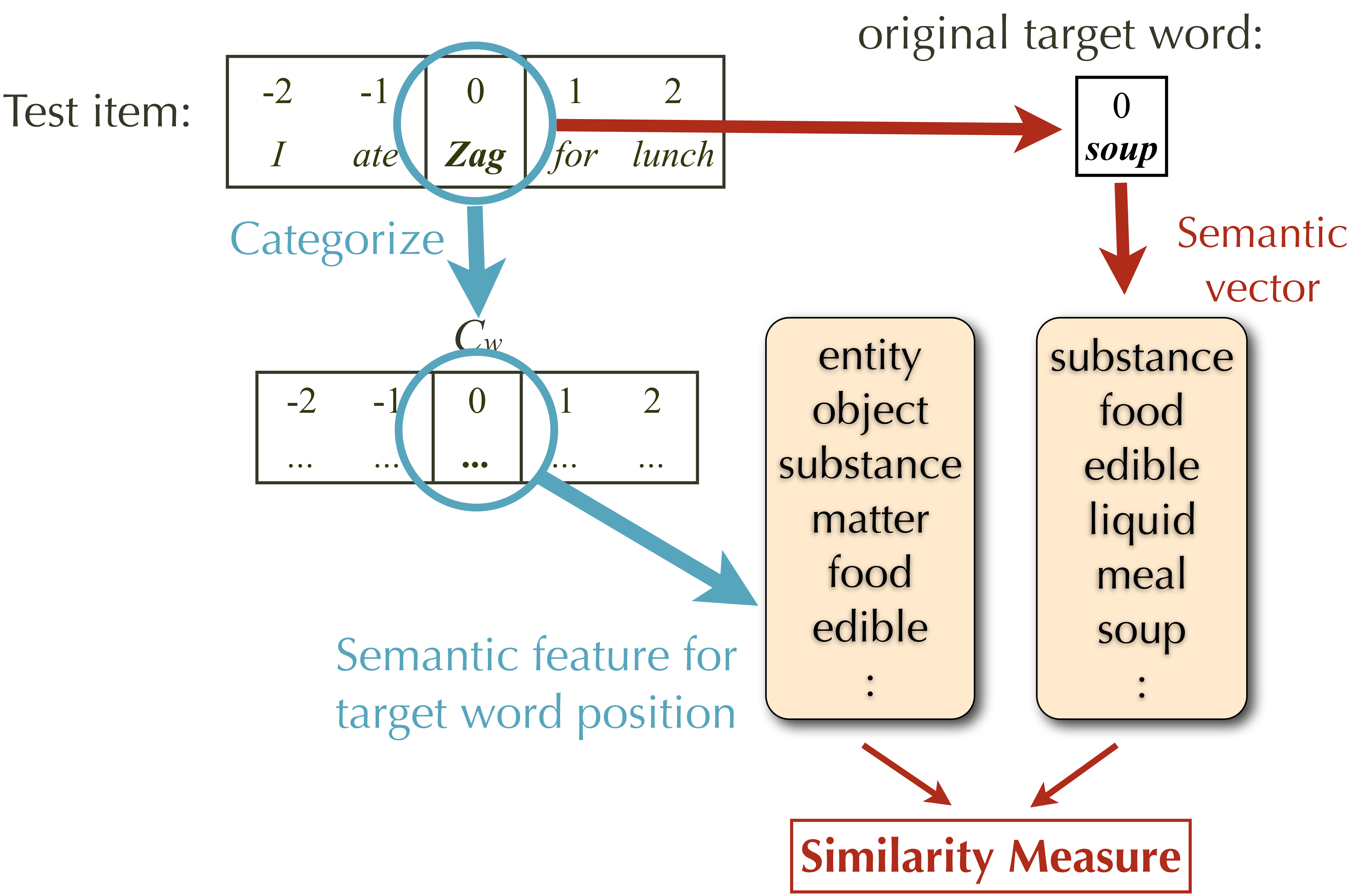
Labelled Data Noun Category Feature Representation

Inferring Word Semantic Properties

I had ZAV for lunch

- Task: guessing the semantic properties of a novel word based on its local context
- Children and adults can guess (some aspects of) the meaning of a novel word from context (Landau & Gleitman'85, Naigles & Hoff-Ginsberg'95)

Inferring Semantic Properties



Lexical Category Acquisition

- Finer-grained lexical categories seem more suitable for some tasks than traditional POS categories
  - Standardized applications are needed to evaluate and compare lexical categories induced by different unsupervised methods
- When categorizing words, do children pay attention to semantic cues as well?
  - Computational investigation: include the semantic features of words into a category learning model, and evaluate the performance
- What about other cues? (E.g., phonological and morphological features)

Rules that Govern Form

- Moving from fixed forms (e.g. '*apple*') to derivational forms

```
play → plays, played, playing
I, you, admire → "I admire you"
```

- Morphology and syntax
    - In all languages, the formation of words and sentences follows highly regular patterns
    - How are the regulations and exceptions represented?
- The study and analysis of language production in children reveals common and persistent patterns

U-shaped Learning Curves

- Observed U-shaped learning curves in children
  - Imitation: an early phase of conservative language use
  - Generalization: general regularities are applied to new forms
  - Overgeneralization: occasional misapplication of general patterns
  - Recovery: over time, overgeneralization errors cease to happen

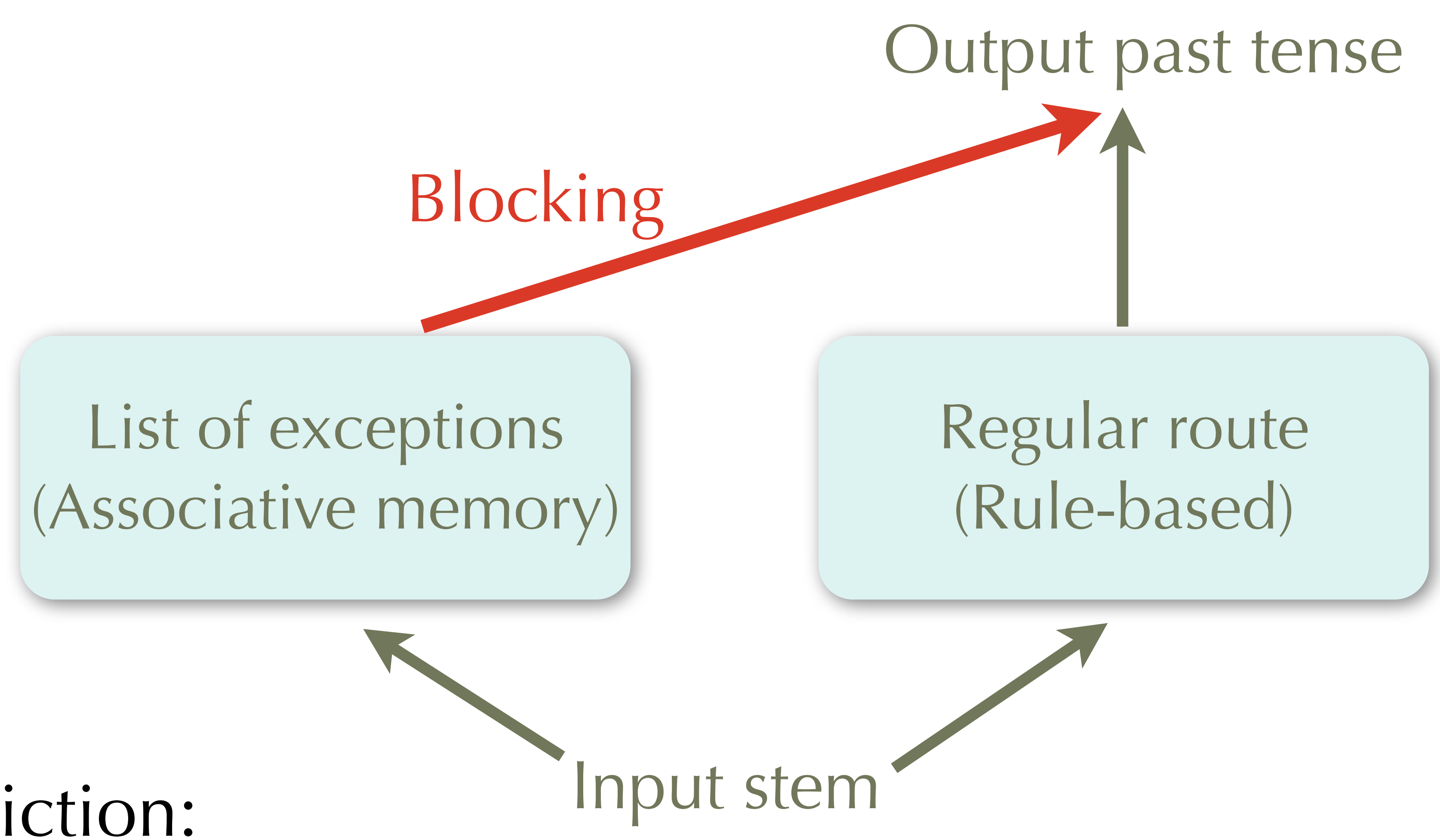
- Lack of Negative Evidence
  - Children do not receive reliable corrective feedback from parents to help them overcome their mistakes (Marcus, 1993)

Case Study: Learning English Past Tense

- The problem of English past tense formation:
  - Regular formation: `stem + 'ed'`
  - Irregulars do show some patterns
    - No-change: **hit → hit**
    - Vowel-change: **ring → rang, sing → sang**
- Over-regularizations are common: **goed**
  - These errors often occur after the child has already produced the correct irregular form: **went**
- What causes the U-shaped learning curve?

A Symbolic Account of English Past Tense

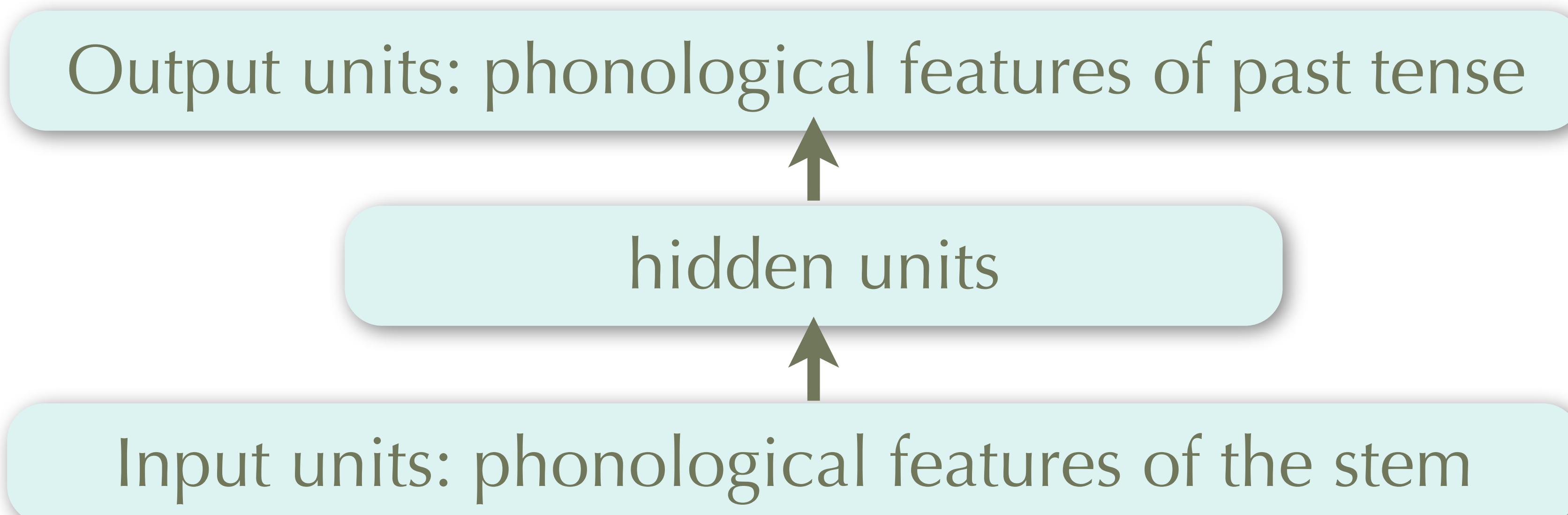
- Dual-Route Account (Pinker, 1991): two qualitatively different mechanisms



- Prediction:
  - Errors result from transition from rote learning to rule-governed
  - Recovery occurs after sufficient exposure to irregulars

A Connectionist Account of Learning English Past Tense

- A connectionist model (Plunkett & Marchman, 1993)



- Properties:
  - Early in training, the model shows tendency to overgeneralize; by the end of training, it exhibits near perfect performance
  - U-shaped performance is achieved using a single learning mechanism, but depends on sudden change in the training size

A Hybrid, Analogy-based Account

- A rational model of learning past tense based on the ACT-R architecture (Taatgen & Anderson, 2002)
  - Declarative memory chunks represent past tenses, both as a goal and as examples

```
PAST-TENSE-GOAL23
        ISA PAST
        OF WALK
        STEM NIL
        SUFFIX NIL
```

goal to determine
past tense of walk

```
PAST-TENSE-GOAL23
        ISA PAST
        OF WALK
        STEM WALK
        SUFFIX ED
```

accomplished goal,
stored in the memory

A Hybrid, Analogy-based Account

- The analogy strategy is implemented by two production rules, based on simple pattern matching:

RULE ANALOGY-FILL-SLOT
IF the goal has an empty <u>suffix</u> slot
AND there is an example in which <u>suffix</u> has a value
THEN set the <u>suffix</u> of the goal to the <u>suffix</u> value of the example

RULE ANALOGY-COPY-A-SLOT
IF the goal has an empty <u>stem</u> slot and the <u>of</u> slot has a certain value
AND in the example the values of the <u>of</u> and <u>stem</u> slots are equal
THEN set the <u>stem</u> to the value of the <u>of</u> slot

ACT-R Equations

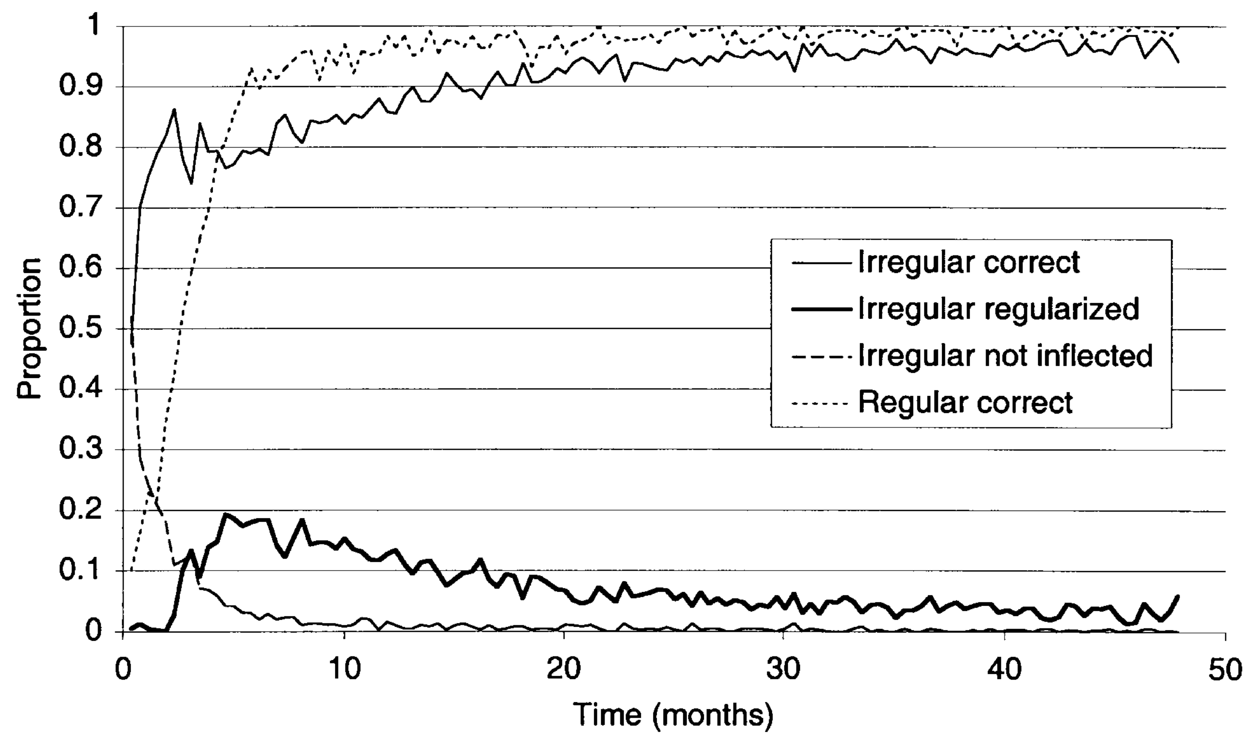
Equation	Description
Activation	
$A = B + \text{context} + \text{noise}$	The activation of a chunk has three parts: base-level activation, spreading activation from the current context and noise. Since spreading activation is a constant factor in the models discussed, we treat activation as if it were just base-level activation.
Base-level activation	
$B(t) = \log \sum_{j=1}^{n} (t - t_j)^{-d}$	n is the number of times a chunk has been retrieved from memory, and t_j represents the time at which each of these retrievals took place. So, the longer ago a retrieval was, the less it contributes to the activation. d is a fixed ACT-R parameter that represents the decay of base-level activation in declarative memory.
Retrieval time	
$\text{Time} = F\mathrm{e}^{-fA}$	Activation determines the time required to retrieve a chunk. A is the activation of the chunk that has to be retrieved, and F and f are fixed ACT-R parameters. Retrieval will only succeed as long as the activation is larger than retrieval threshold τ, which is also a fixed parameter.
Expected outcome	
$\text{Expected outcome} = P_{\mathrm{p}}G - C_{\mathrm{p}} + \text{noise}$	Expected outcome is based on three quantities, the estimated probability of success of a production rule (P), the estimated cost of the production rule (C), and the value of the goal (G).

A Hybrid, Analogy-based Account

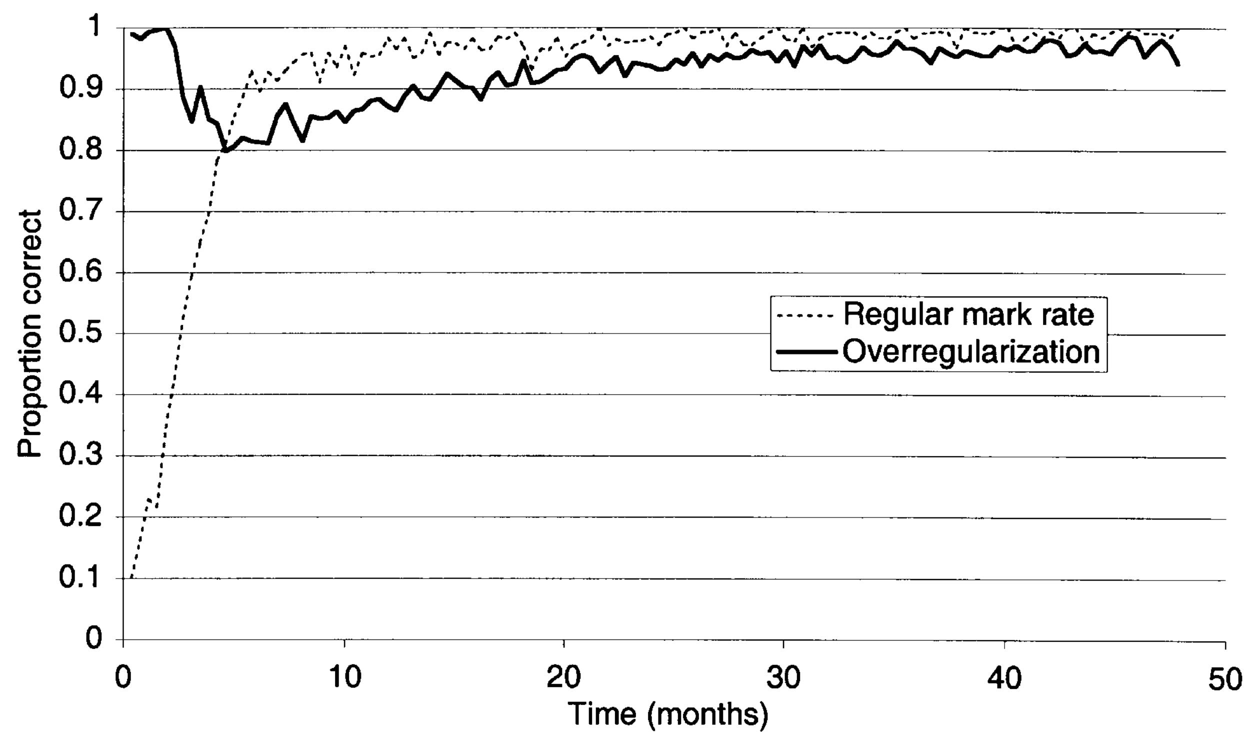
- ACT-R's production rule mechanism learns new rules by combining two rules that have fired consecutively into one:

```
RULE LEARNED-REGULAR-RULE

    IF the goal is to find the past tense of a
    word and slots stem and suffix are empty

    THEN set the suffix slot to ED and set the
    stem slot to the word of which you want the
    past tense
```

A Hybrid, Analogy-based Account



(a)



(b)

Innateness of Language

- Central claim: humans have innate knowledge of language
  - Assumption: all languages have a common structural basis
- Argument from the Poverty of the Stimulus (Chomsky 1965)
  - Linguistic experience of children is not sufficiently rich for learning the grammar of the language, hence they must have some innate specification of grammar
  - Assumption: knowing a language involves knowing a grammar
- Universal Grammar (UG)
  - A set of rules which organize language in the human brain