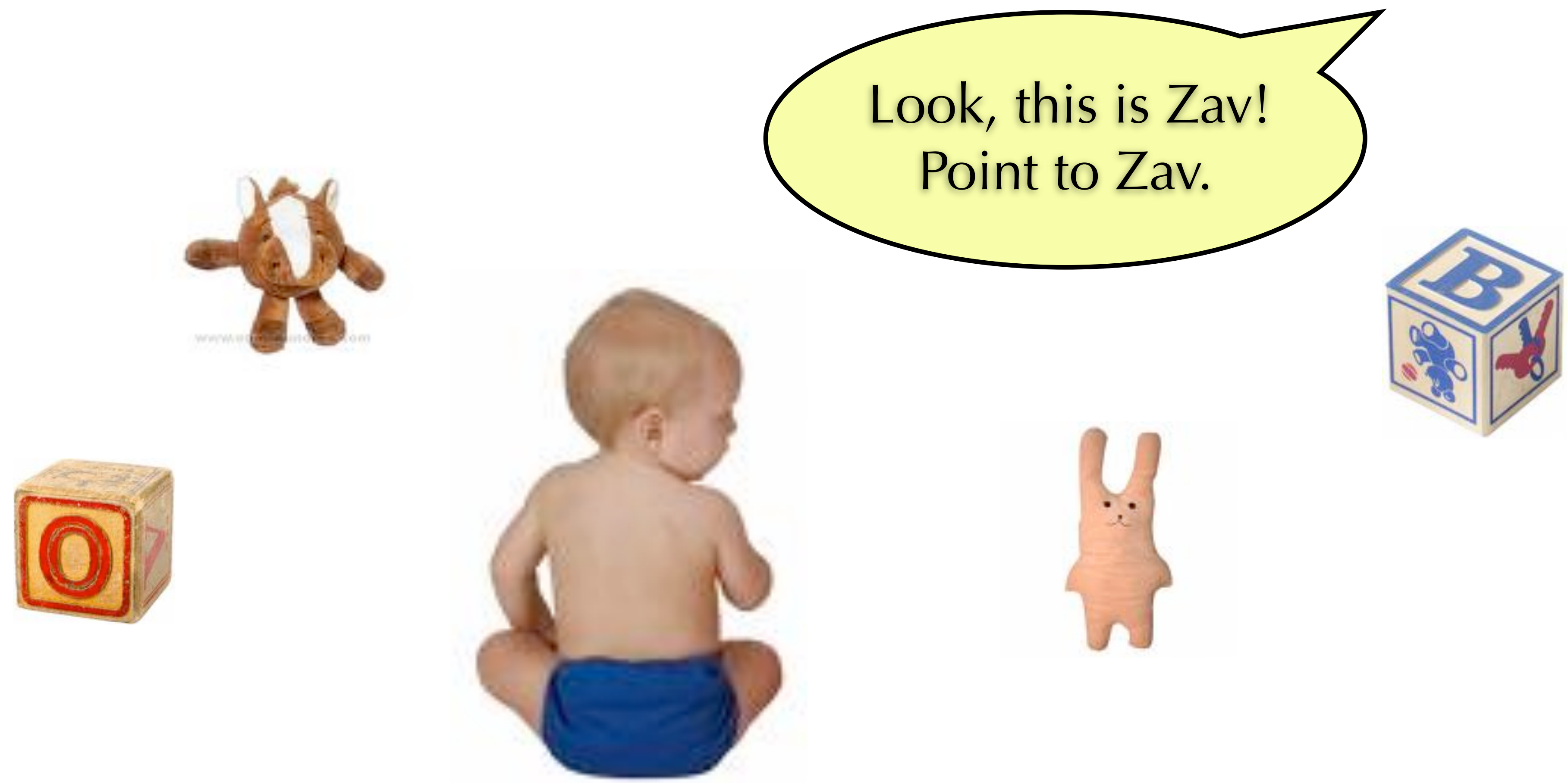
Lexical Category Acquisition as an Incremental Process

Afra Alishahi, Grzegorz Chrupała

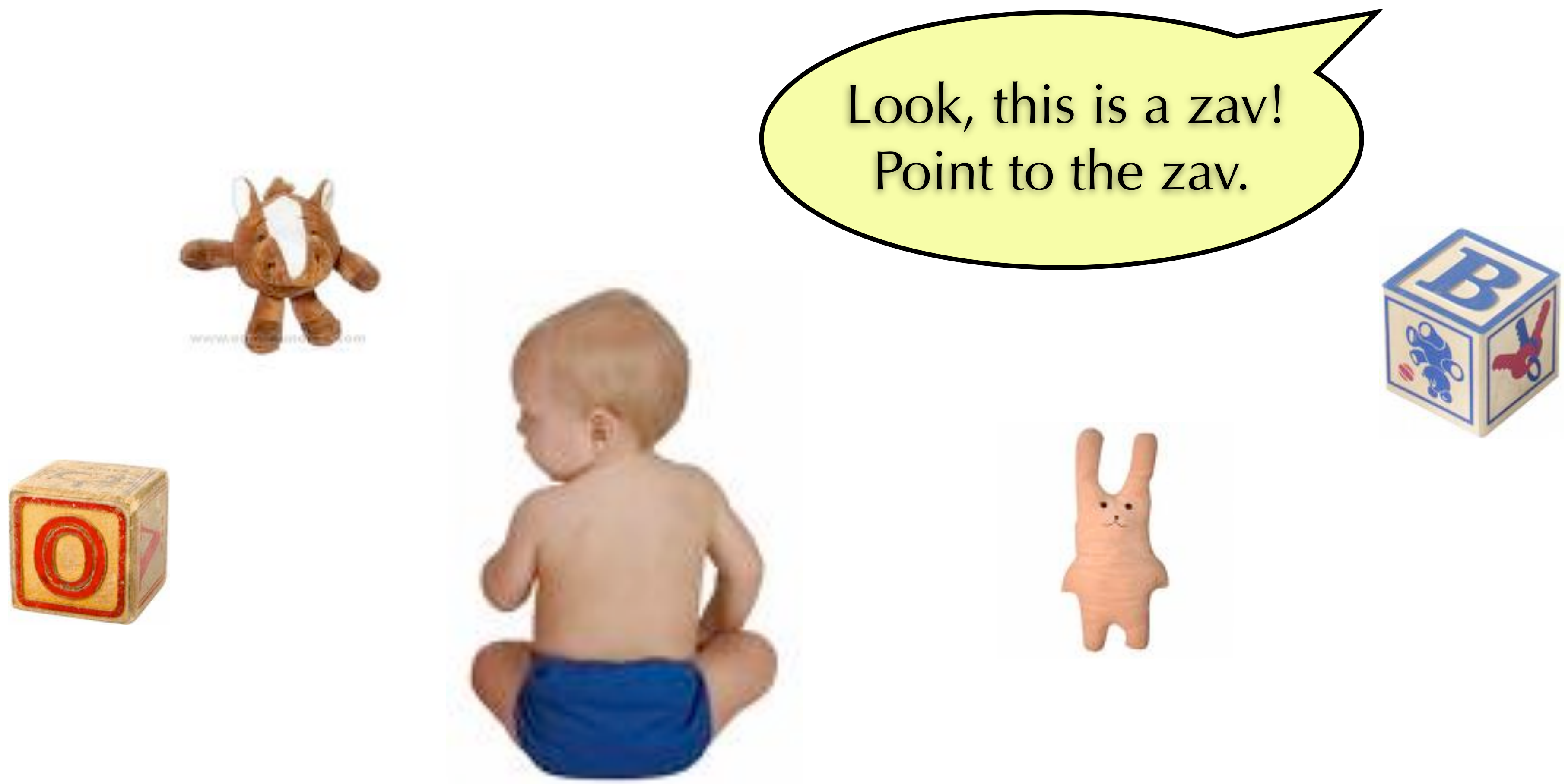
FEAST, July 21, 2009

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

Goals

- Propose a cognitively plausible algorithm for inducing categories from child-directed speech

- Suggest a novel way of evaluating the learned categories via a variety of language tasks

Part I: Category Induction

Information Sources

- 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

- Hierarchical clustering models
  - Starting from a cluster per each word type, the two most similar clusters are merged in each iteration (Schutze'93, Redington et al'98)
- Cluster optimization models
  - Vocabulary is partitioned into non-overlapping clusters, which are optimized according to an information theoretic measure (Brown'92, Clark'00)
- Incremental clustering models
  - Each word usage is added to the most similar existing cluster, or a new cluster is created (e.g. Cartwright & Brent'97, Parisien et al'08)
- Existing models rely on optimizing techniques, demanding high computational load for processing data

Our Model

- We propose an efficient incremental model for lexical category induction from unannotated text
  - 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"*

  - Each usage is represented as a vector:

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

Representation of Word Categories

- A lexical category is a cluster of word usages
  - The distributional context of a category is represented as 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 clusters is measured by the dot product of their vectors

Online Clustering Algorithm

For every word usage w:

- Create new cluster C_{new}
- Add $\Phi(w)$ to C_{new}
- $C_w = \text{argmax}_{C \in \text{Clusters}} \textbf{Similarity}(C_{new}, C)$
- If $\textbf{Similarity}(C_{new}, C_w) \geq \theta_w$
  - merge C_w and C_{new}
  - $C_{next} = \text{argmax}_{C \in \text{Clusters} - \{C_w\}} \textbf{Similarity}(C_w, C)$
  - If $\textbf{Similarity}(C_w, C_{next}) \geq \theta_c$
    - merge C_w and C_{next}

where $\textbf{Similarity}(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ and the vector $\Phi(w)$ represents the context features of the current word usage w.

Experimental Data

- Manchester corpus from CHILDES database (Theakston et al.'01, MacWhinney'00)

```
what about that
pro:wh prep pro:dem

make Mummy push her
v n:prop v pro

push her then
v pro adv:tem
```

Data Set	Corpus	#Sentences	#Words
Develop	Anne	857	3,318
Train	Anne	13,772	73,032
Test	Becky	1,116	5,431

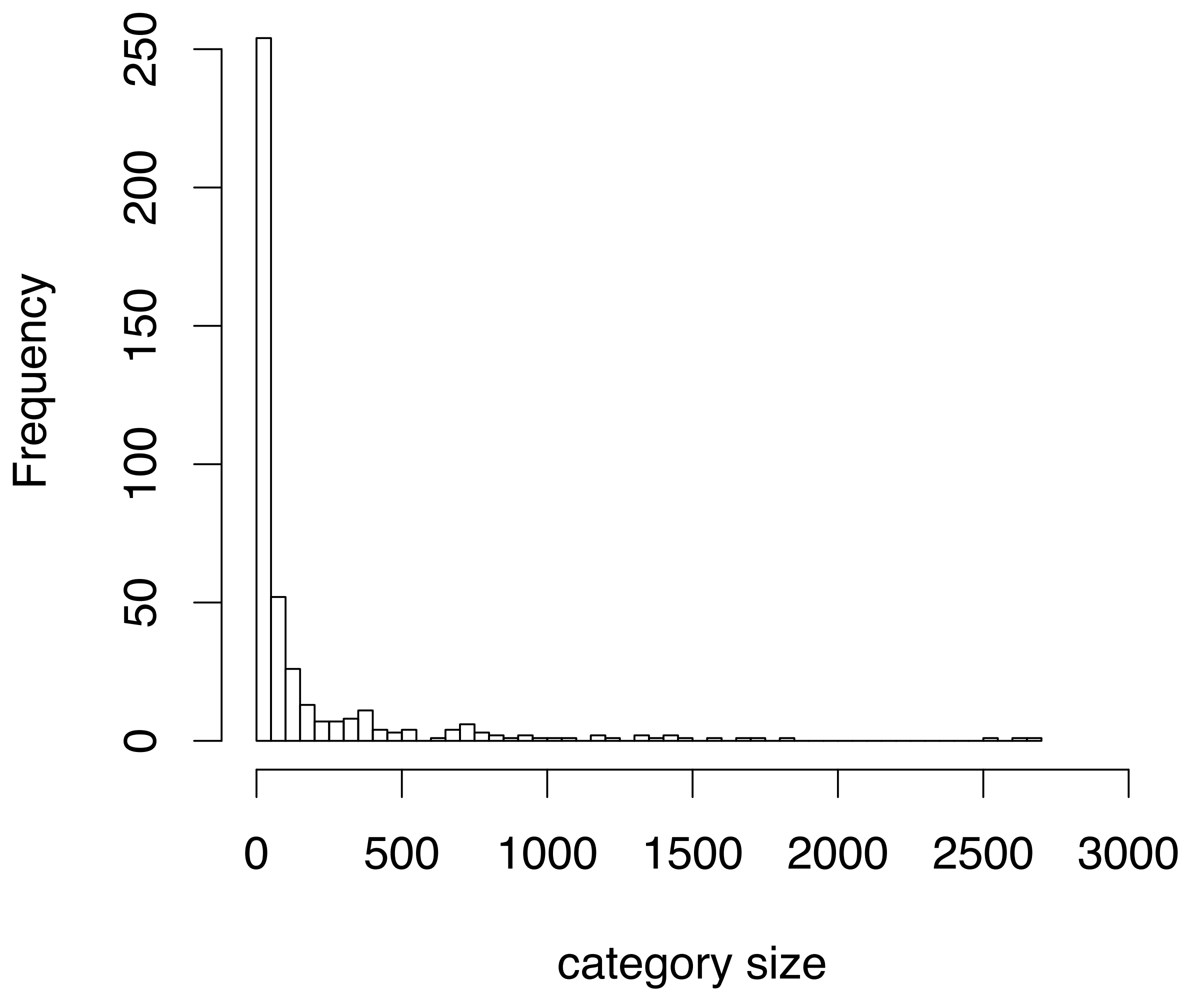
(One-word sentences are excluded from training and test data)

- Threshold values are set based on development data:

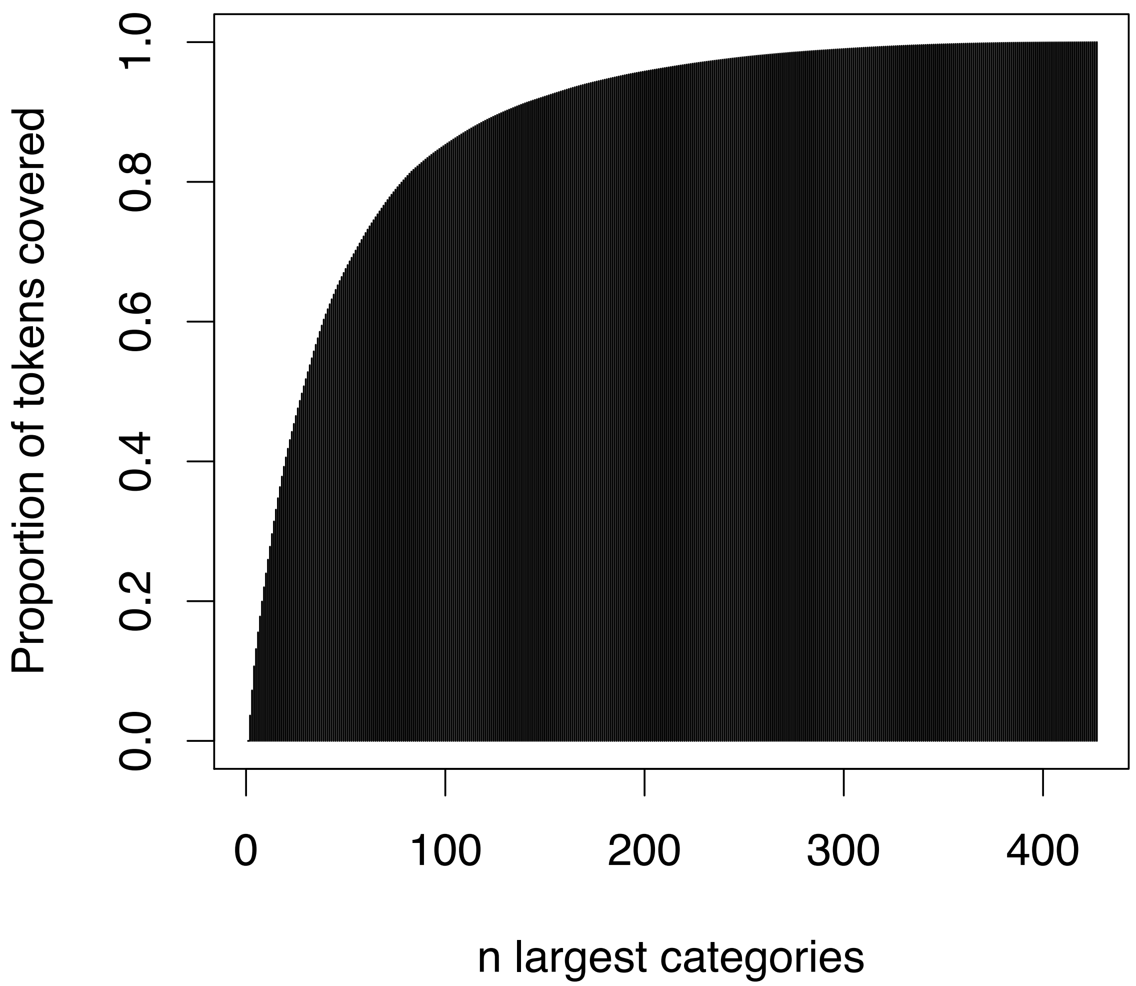
$\theta_w = 2^7 \times 10^{-3}$ and $\theta_c = 2^{10} \times 10^{-3}$

Category Size

Distribution of the size of categories



Coverage of tokens by categories



Processing the training data yielded a total of 427 categories.

Sample Induced Categories

do	train	‘s
are	cover	is
will	one	was
have	tunnel	in
can	hole	then
has	king	goes
does	door	on
had	fire-	
were	engine	
:	:	:

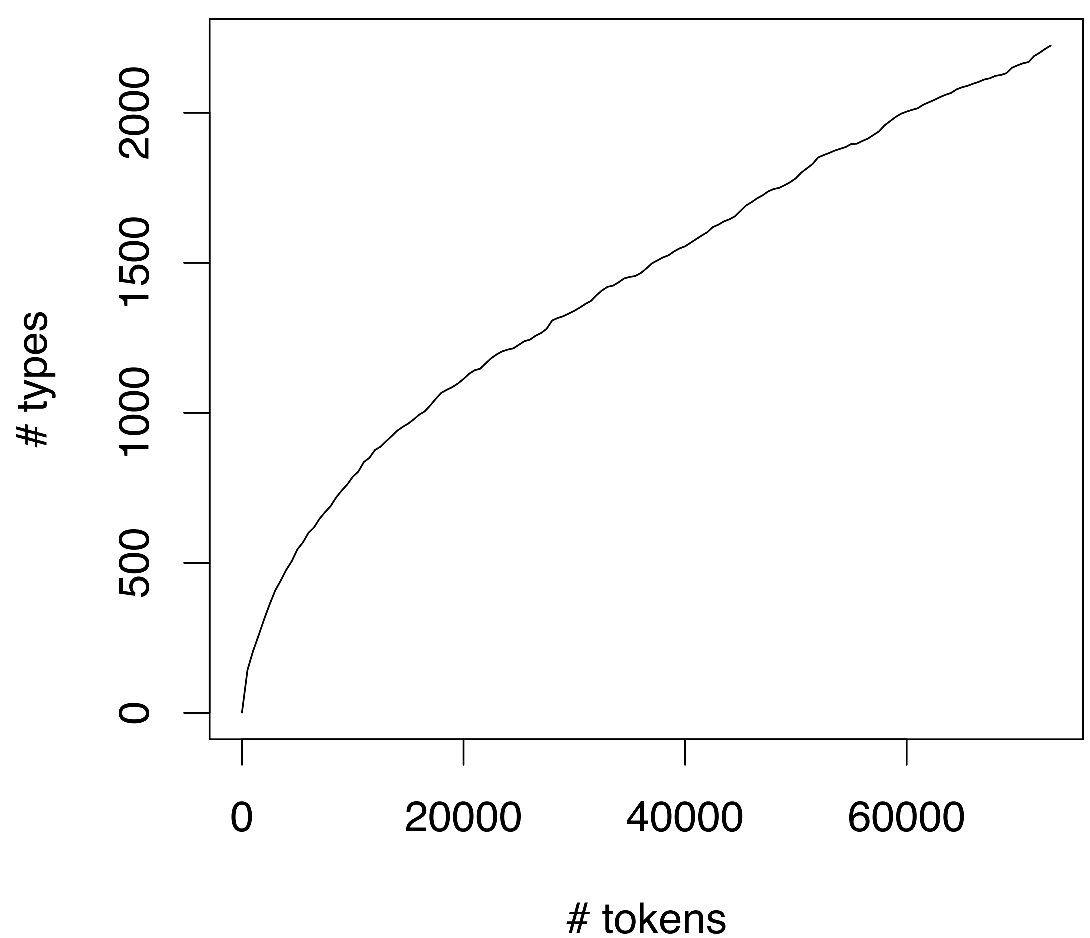
Most frequent values for the content word feature

bit	the	‘re
little	a	‘ve
good	this	want
big	that	got
very	her	see
long	there	were
few	their	do
drink	our	find
funny	another	going
:	:	:

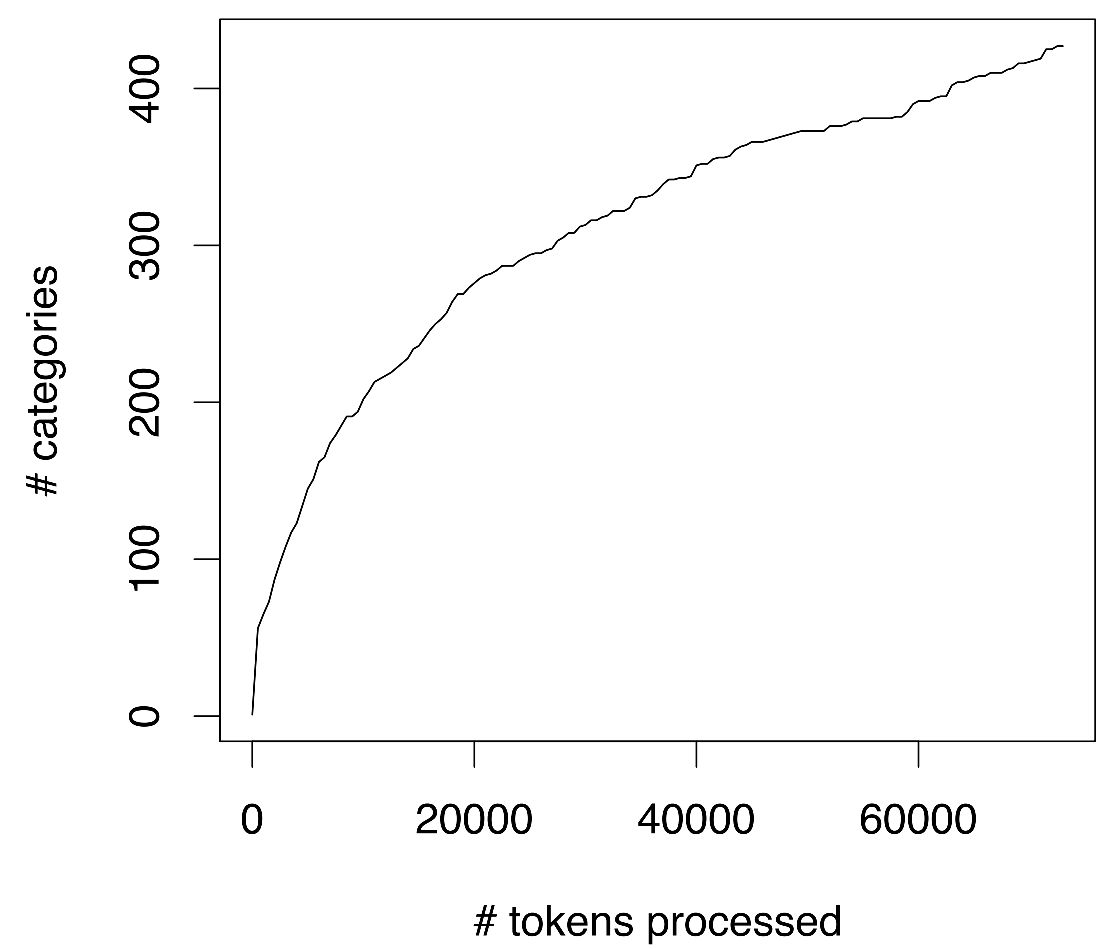
Most frequent values for the previous word feature

Vocabulary and Category Growth

Vocabulary growth



Category growth



- The growth of the size of the vocabulary (i.e. word types), as well as the number of lexical categories, slows down over time

Part 2: Evaluation

Common Evaluation Approach

- POS tags as gold-standard: evaluate their categories based on how well they match POS categories
  - Accuracy and Recall: every pair or words in an induced category should belong to the same POS category (Redington et al.'98)
  - Order of category formation: categories that resemble POS categories show the same developmental trend (Parisien et al'08)
- Alternative evaluation techniques
  - Substitutability of category members in training sentences (Frank et al.'09)
  - Perplexity of a finite state model based on two sets of categories (Clark'01)

Our Proposal: Measuring ‘Usefulness’ instead of ‘Correctness’

- Instead of using a gold-standard to compare our categories against, we use the categories in a variety of applications
    - Word prediction from context
    - Inferring semantic properties of novel words based on the context they appear in
- We 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 - Methodology

Test item:

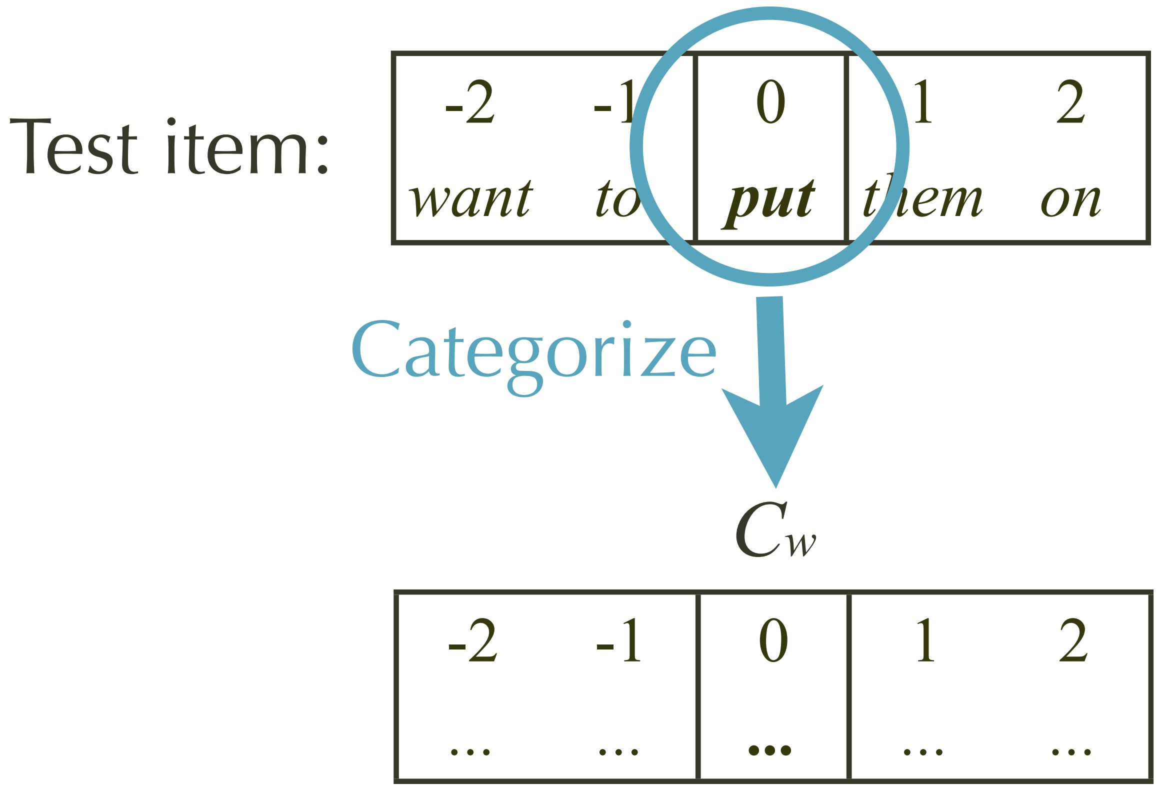
-2	-1	0	1	2
want	*to*	***put***	*them*	*on*

Word Prediction - Methodology

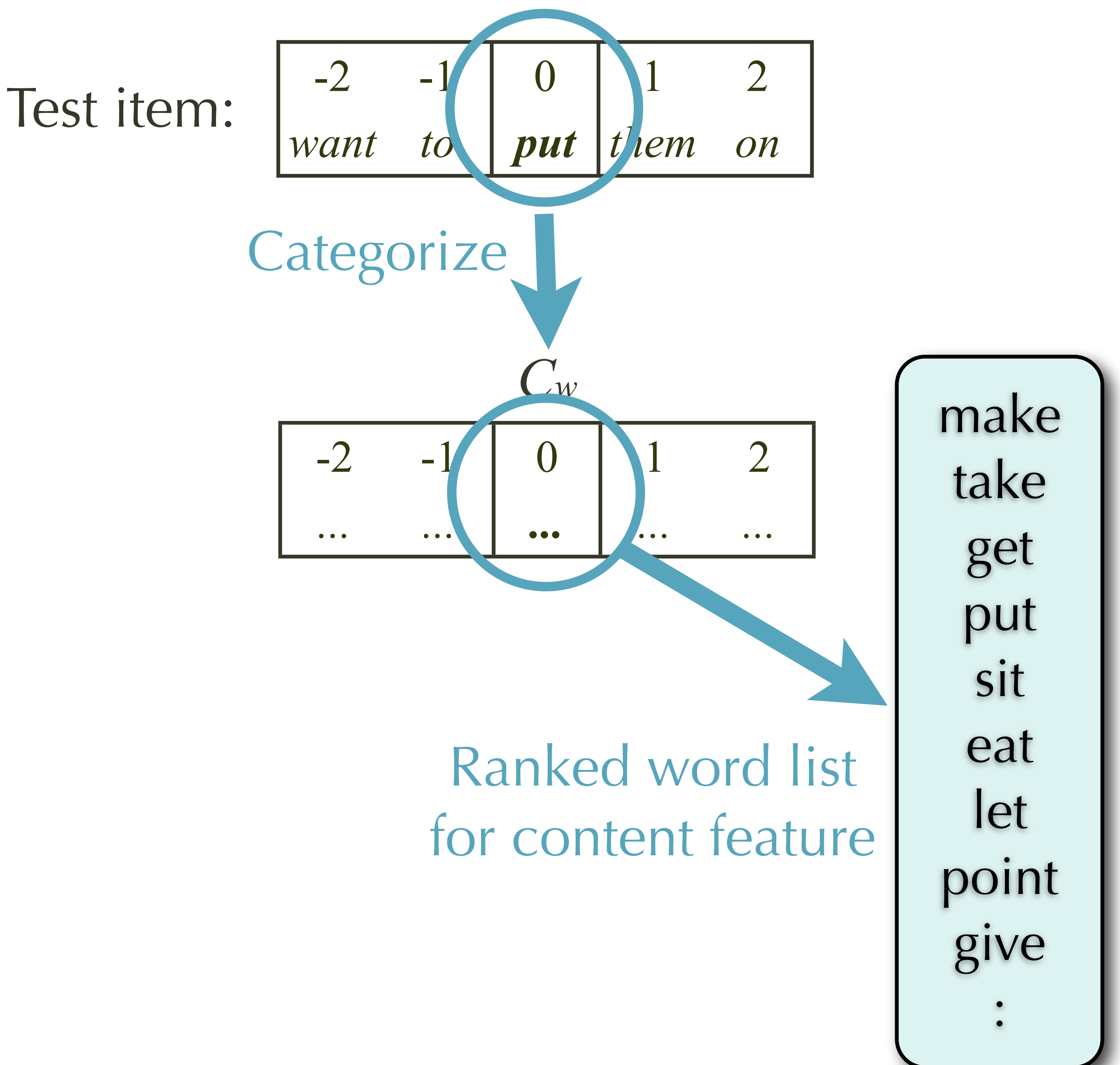
Test item:

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

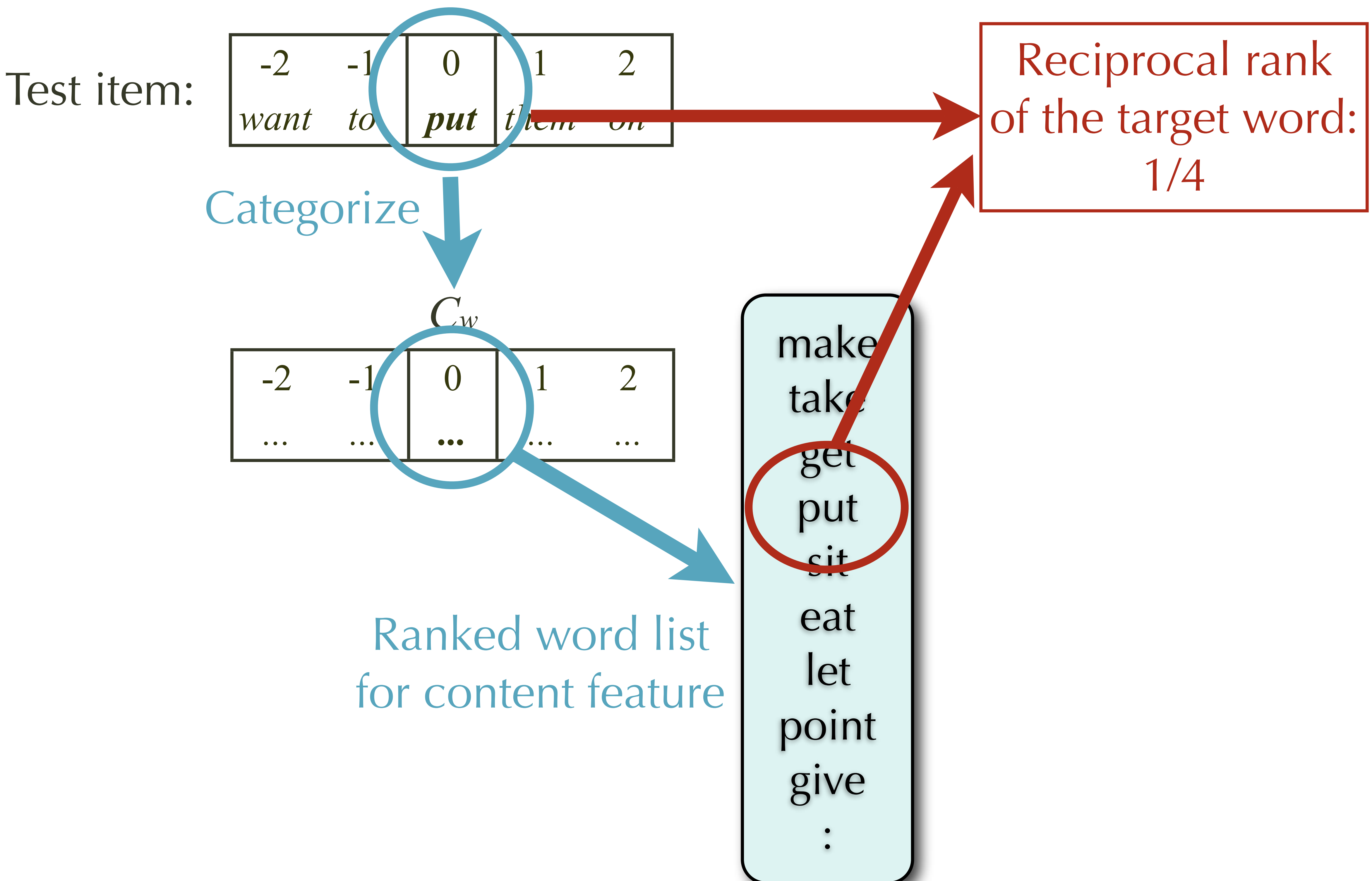
Word Prediction - Methodology



Word Prediction - Methodology



Word Prediction - Methodology



Word Prediction - POS Categories

baby 's Mummy
n v n:prop

put them on the table look
v pro prep det n v

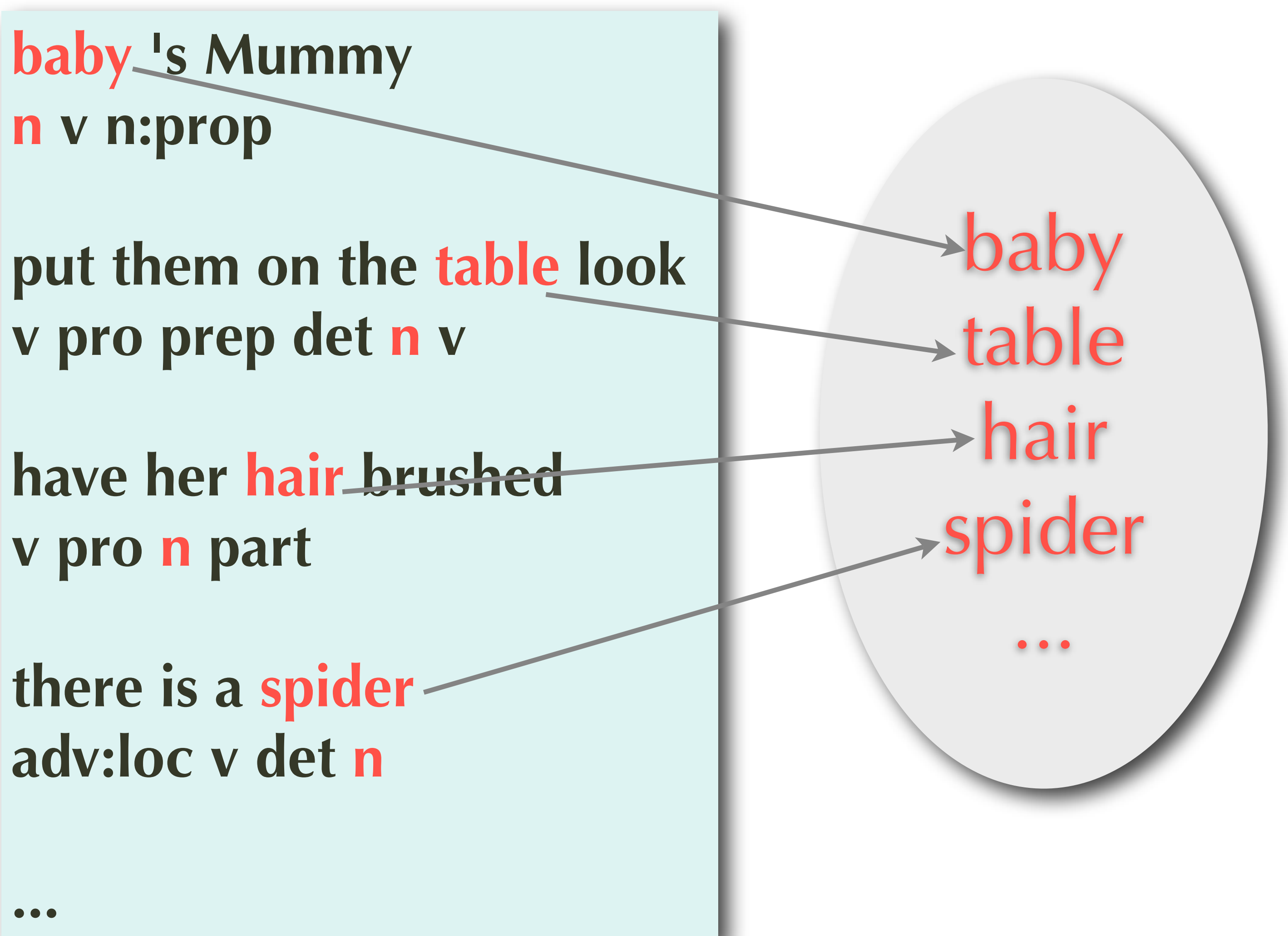
have her hair brushed
v pro n part

there is a spider
adv:loc v det n

...

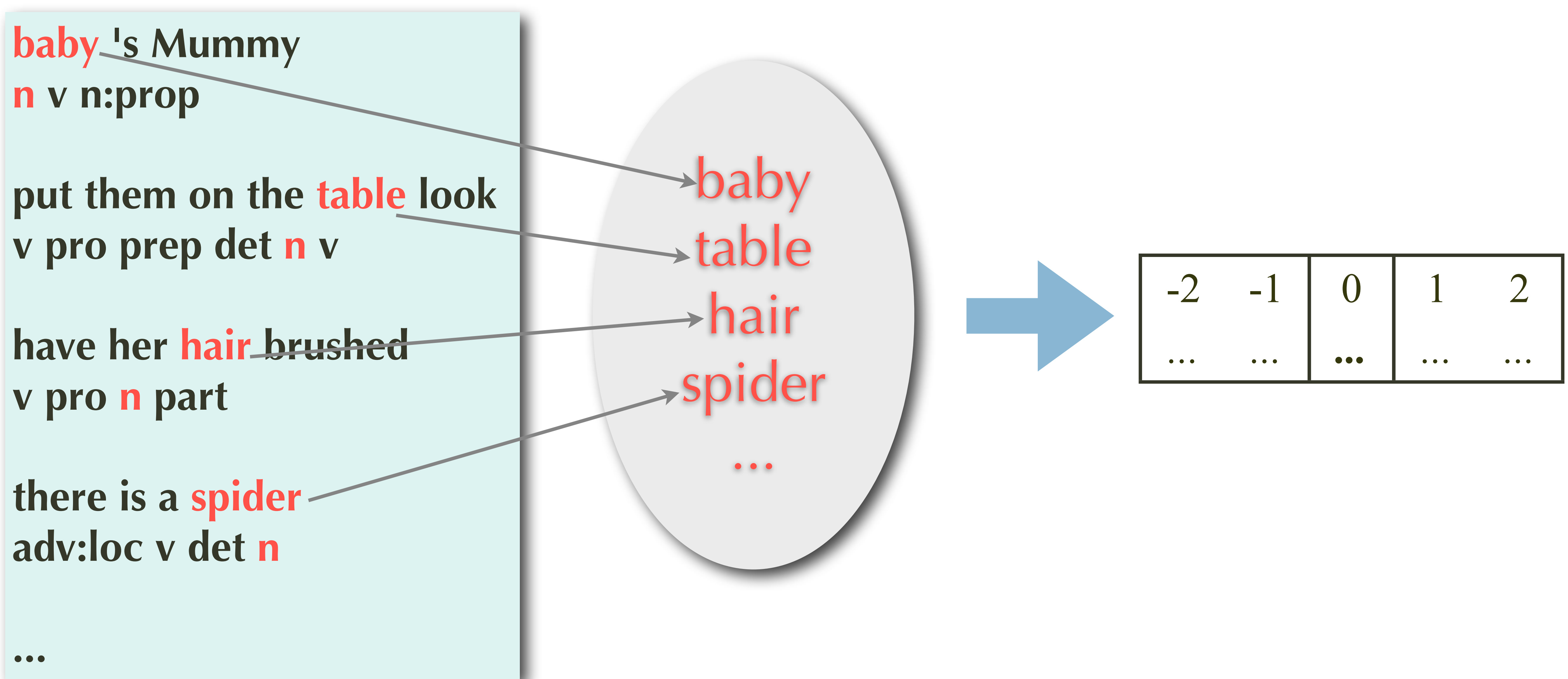
Labelled Data

Word Prediction - POS Categories



Labelled Data **Noun Category**

Word Prediction - POS Categories



Labelled Data

Noun Category

Feature Representation

Word Prediction - Results

Category Type	Mean Reciprocal Rank
POS	0.073
Induced	**0.198**
Word type	0.009

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)

Word Semantic Properties

- Semantic features of each word are extracted from WordNet:

cake
→baked goods
→food
→solid
→substance, matter

WordNet hypernyms for cake **Semantic vector for cake**

- Semantic feature vector for each category is the mean of the semantic vectors of its members
  - Note: semantic features are not used in categorization

Word Semantic Properties

- Semantic features of each word are extracted from WordNet:

cake
→baked goods
→food
→solid
→substance, matter

cake
baked goods
food
solid
substance

WordNet hypernyms for cake **Semantic vector for cake**

- Semantic feature vector for each category is the mean of the semantic vectors of its members
  - Note: semantic features are not used in categorization

Inferring Semantic Properties - Methodology

Test item:

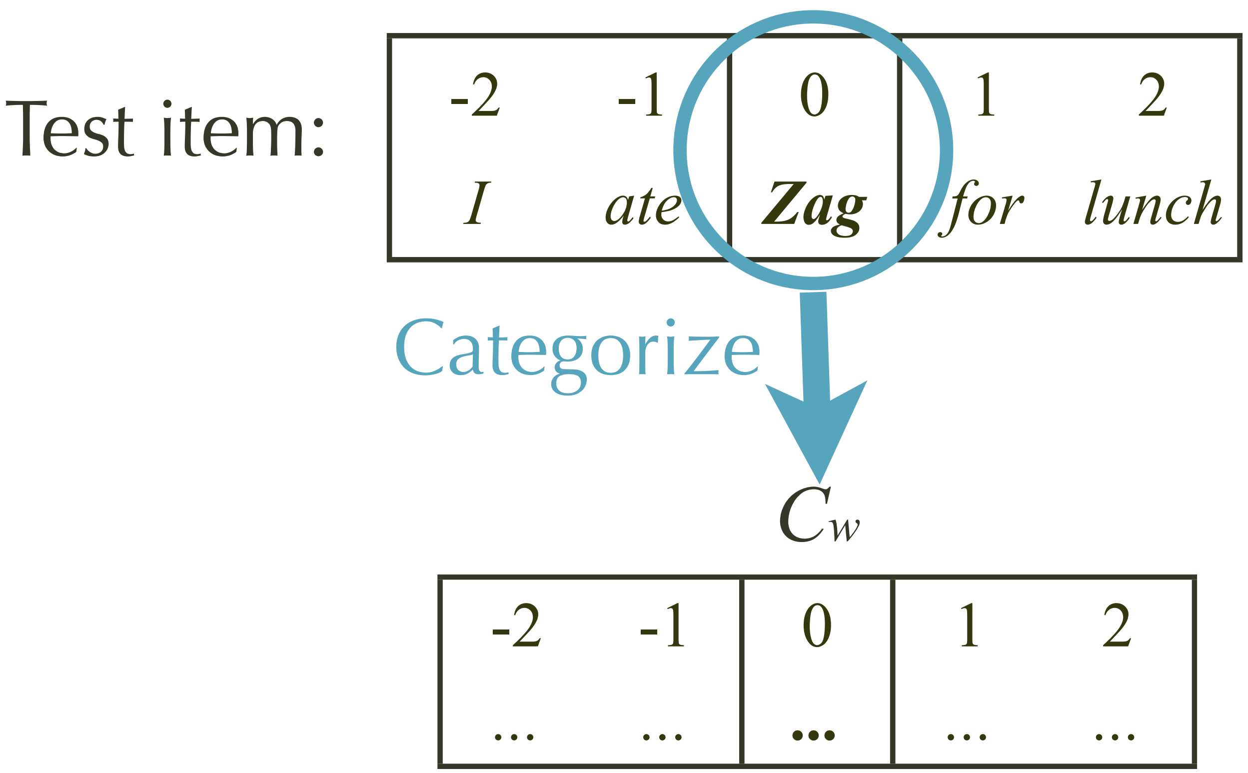
-2	-1	0	1	2
I	*ate*	***Zag***	*for*	*lunch*

Inferring Semantic Properties - Methodology

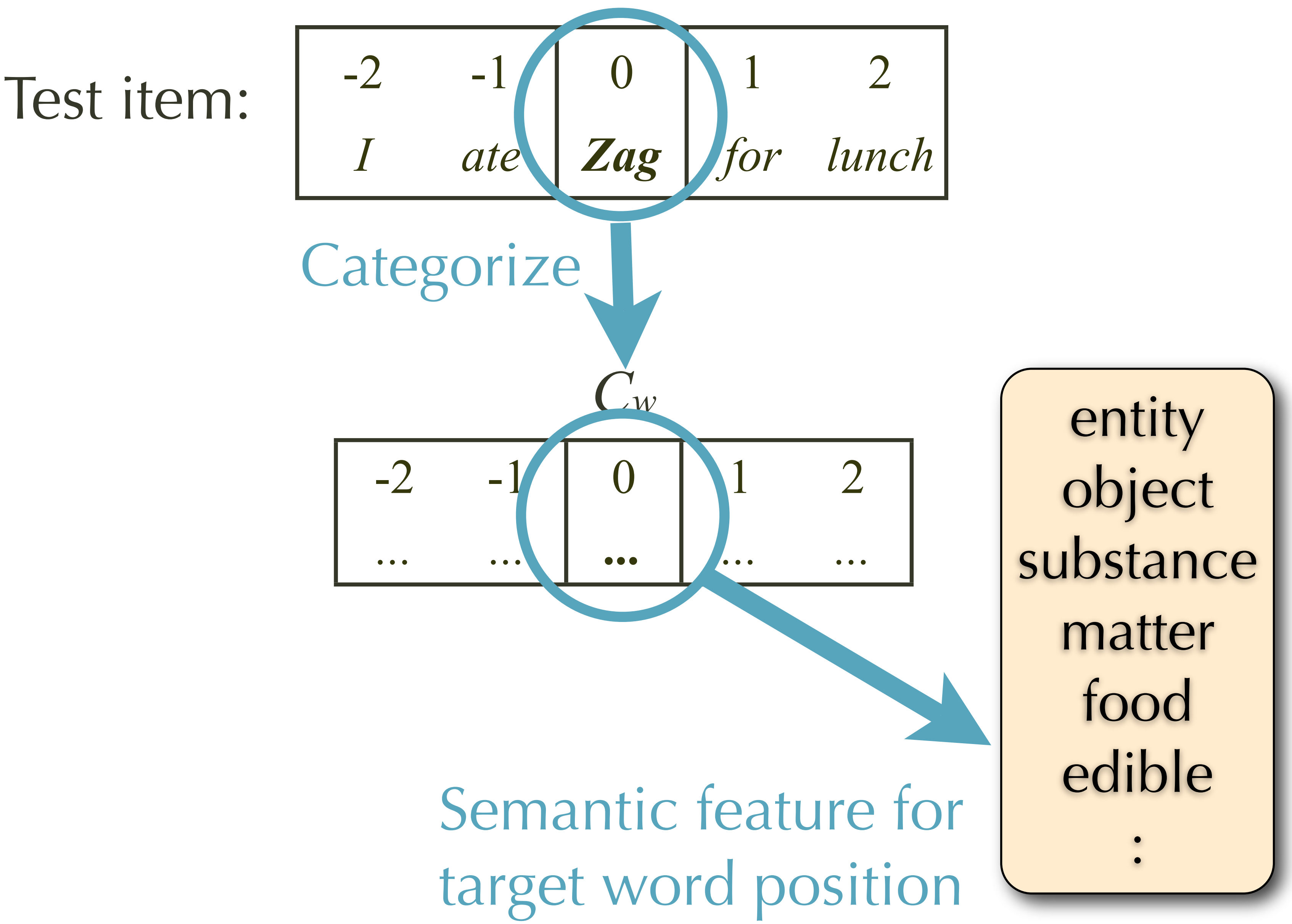
Test item:

-2	-1	0	1	2
I	*ate*	***Zag***	*for*	*lunch*

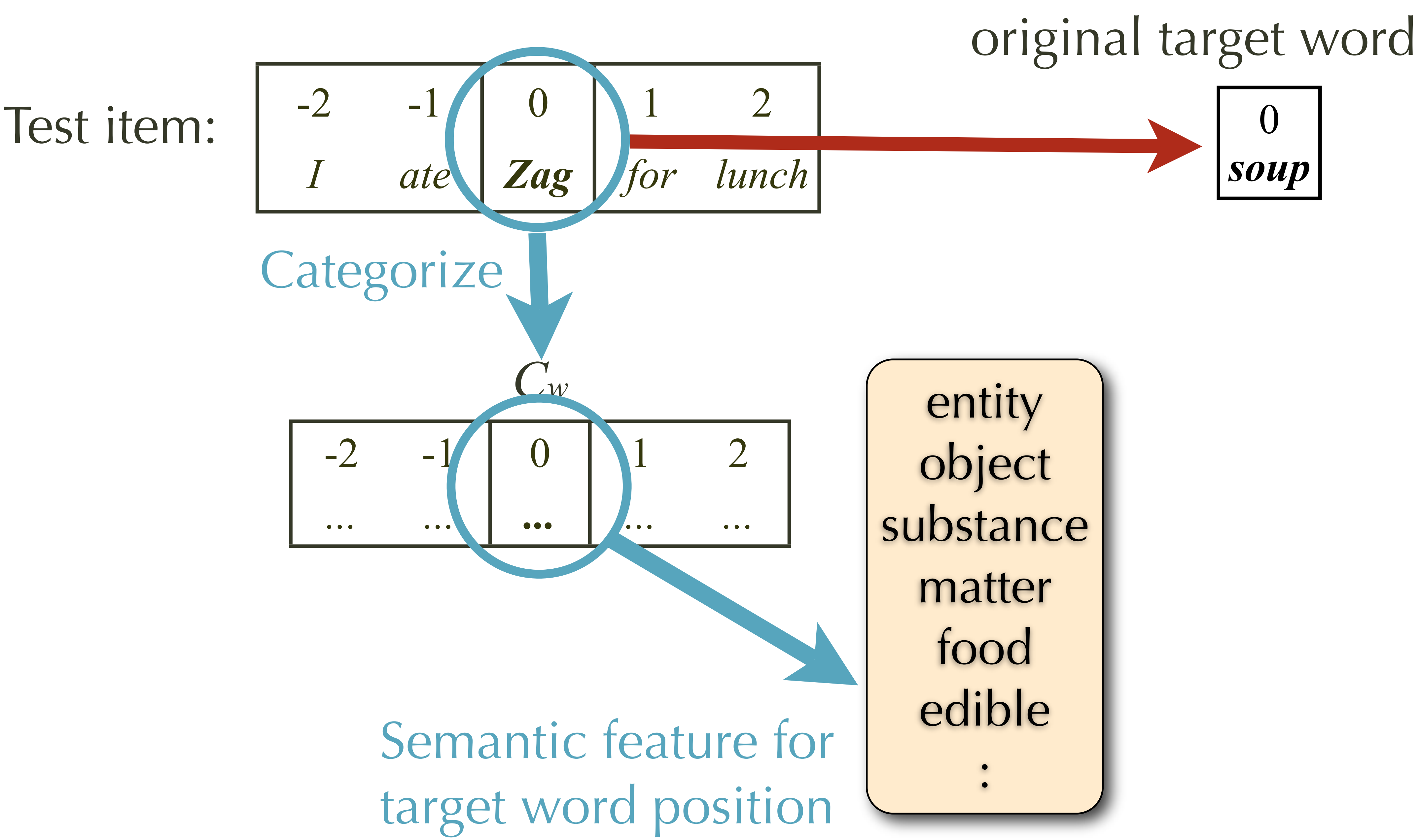
Inferring Semantic Properties - Methodology



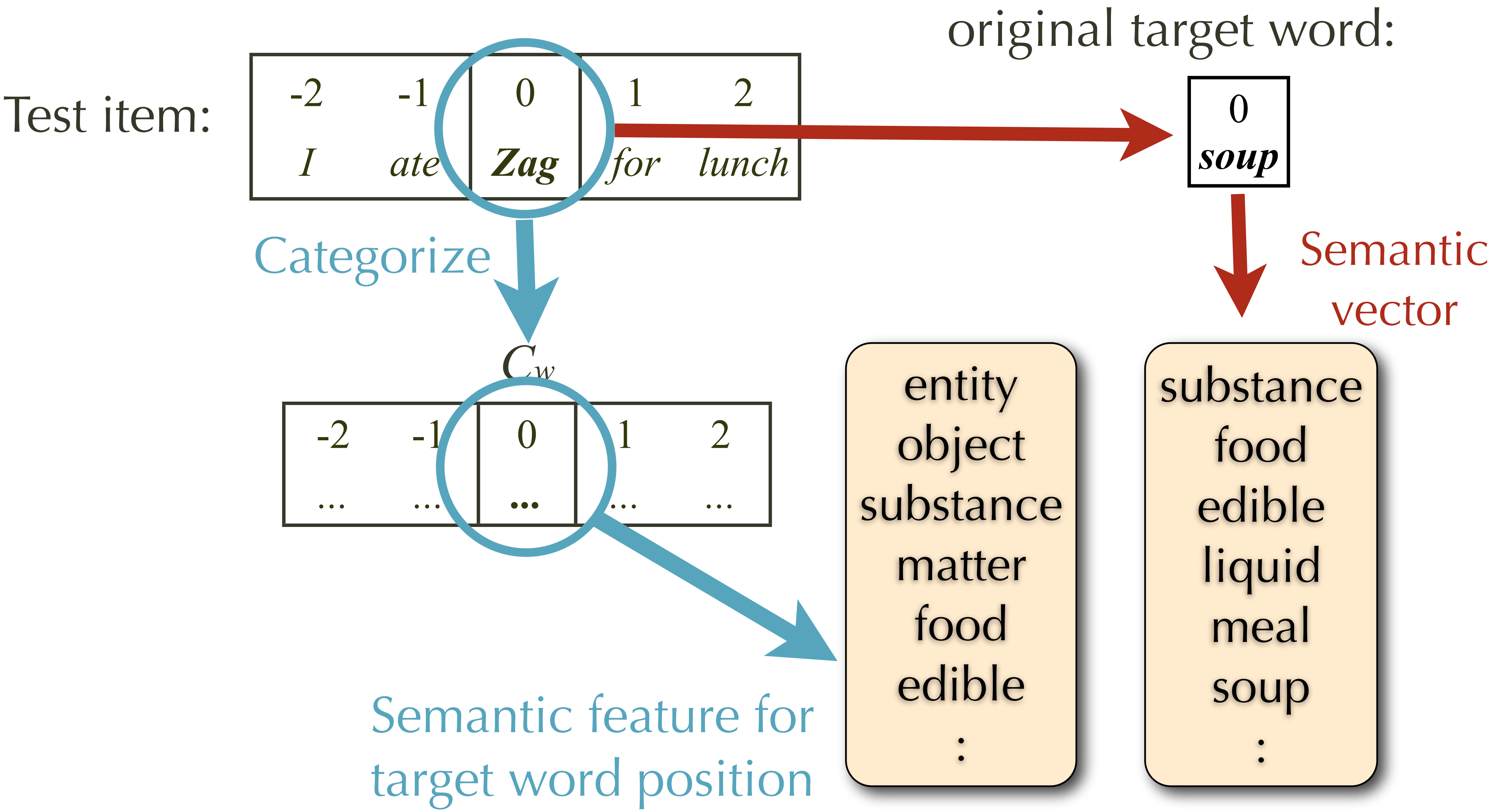
Inferring Semantic Properties - Methodology



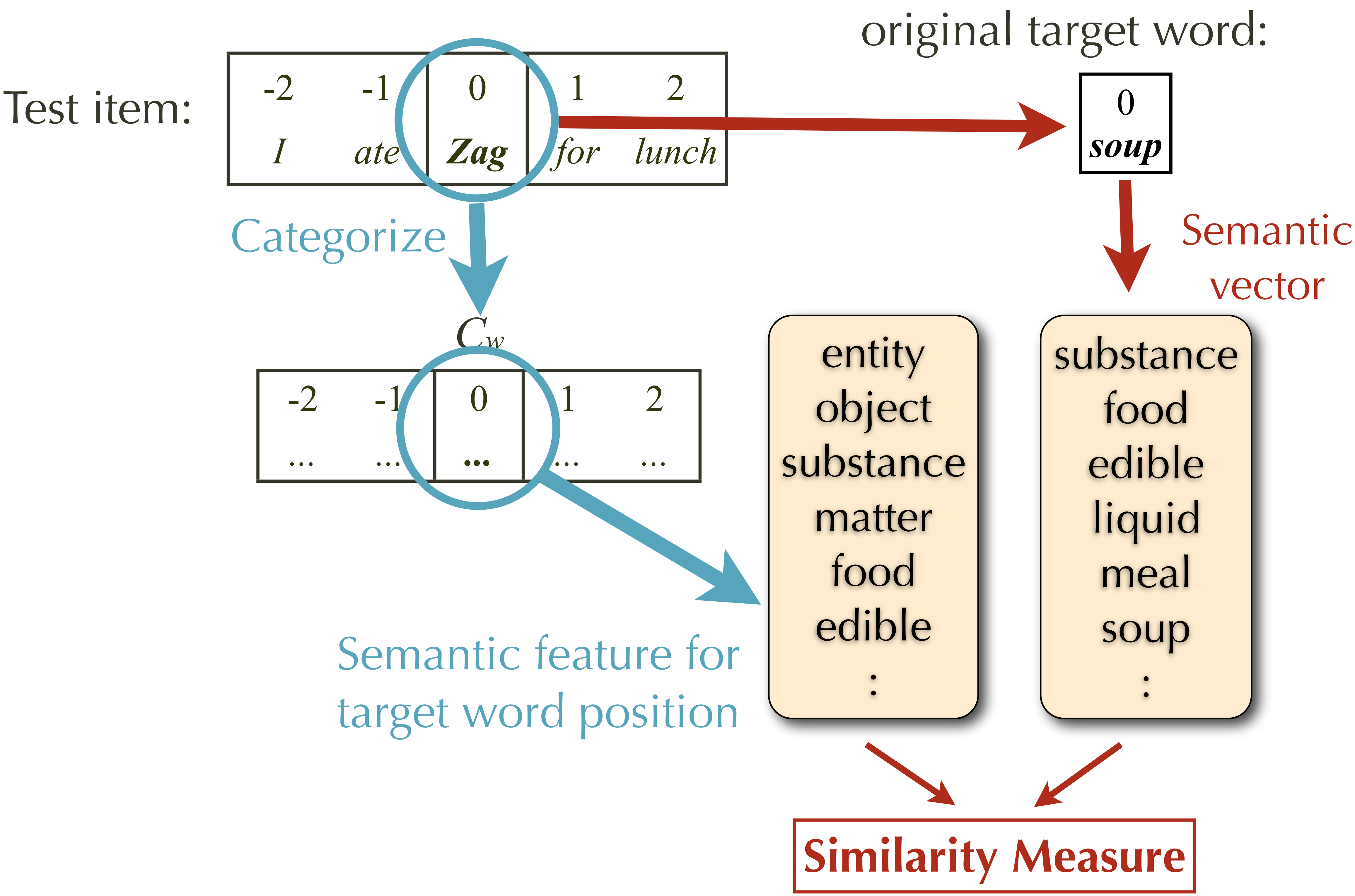
Inferring Semantic Properties - Methodology



Inferring Semantic Properties - Methodology



Inferring Semantic Properties - Methodology



Inferring Semantic Properties - Results

Category Type	Average Dot Product
POS	0.035
Induced	**0.048**

Discussion

- We propose an incremental model of lexical category acquisition based distributional properties of words
    - Model learns intuitive categories from child-directed speech
    - Categories are successfully used in word prediction and the inference of semantic properties of words from context
- 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

Future Directions

- Improving the model
    - Alternative representations of the local context
        - Applying a Gaussian filter on context window
    - Bootstrapping
        - Using categories of the previous words as feature
    - Alternative representations of categories and similarity measures
- Evaluating categories via more applications
    - Lexical decision
    - Grammaticality judgment