Language Acquisition

Fall 2010/Winter 2011

Model Evaluation & Word Segmentation

(December 16, 2010)

Afra Alishahi, Heiner Drenhaus

Computational Linguistics and Phonetics
Saarland University

Evaluation of Computational Models

- Cognitive models cannot be solely evaluated based on their accuracy in performing a task
  - The behavior of the model must be compared against observed human behavior
  - The errors made by humans must be replicated and explained
- Evaluation of cognitive models depends highly on experimental studies of language

Language Acquisition Models: Evaluation

- What humans know about language can only be estimated/evaluated through how they use it
    - Language processing and understanding
    - Language production
- Analysis of child production data yields valuable clues
    - Developmental patterns such as error and recovery
- Comprehension experiments reveal biases and preferences
    - knowledge sources that children exploit, and their biases towards linguistic cues

Language Production Data

- CHILDES database (MacWhinney, 1995)
  - An ever-growing collection of the recorded interactions (text, audio, video) between children and their parents

```
@Languages:     en
@Participants:  CHI Adam Target_Child, URS Ursula_Bellugi Investigator, MOT Mother, ...
@ID: en|brown|CHI|3;1.26|male|normal|middle_class|Target_Child||
@ID: en|brown|PAU|||||Brother||
@ID: en|brown|MOT|||||Mother||
..
@Date:          30-AUG-1963
@Time Duration: 10:30-11:30
*CHI:           one busses .
%mor:           det:num|one n|buss-PL .
%xgra:          1|2|QUANT 2|0|ROOT 3|2|PUNCT
*URS:           one .
%mor:           det:num|one .
%xgra:          1|0|ROOT 2|1|PUNCT
*CHI:           two busses .
%mor:           det:num|two n|buss-PL .
%xgra:          1|2|QUANT 2|0|ROOT 3|2|PUNCT
*CHI:           three busses .
%mor:           det:num|three n|buss-PL .
%xgra:          1|2|QUANT 2|0|ROOT 3|2|PUNCT
```

Experimental Methods

- Online methodologies
  - Reading time studies: measure relative processing difficulties
  - Eye-tracking studies: Monitor gaze as people hear a spoken utterance; anticipatory eye-movements reflect interpretation
  - Visual world paradigm: monitor subjects' eye movements to visual stimuli as they listen to an unfolding utterance
- Offline methodologies
  - Preferential looking studies: monitor infants' preferences of certain scene depictions based on linguistic stimuli
  - Act-out scenarios: describe an event and ask the child to act it out using a set of toys and objects
  - Elicitation tasks: persuade the child to describe an event or action

Reading Times

- Reading the whole sentence

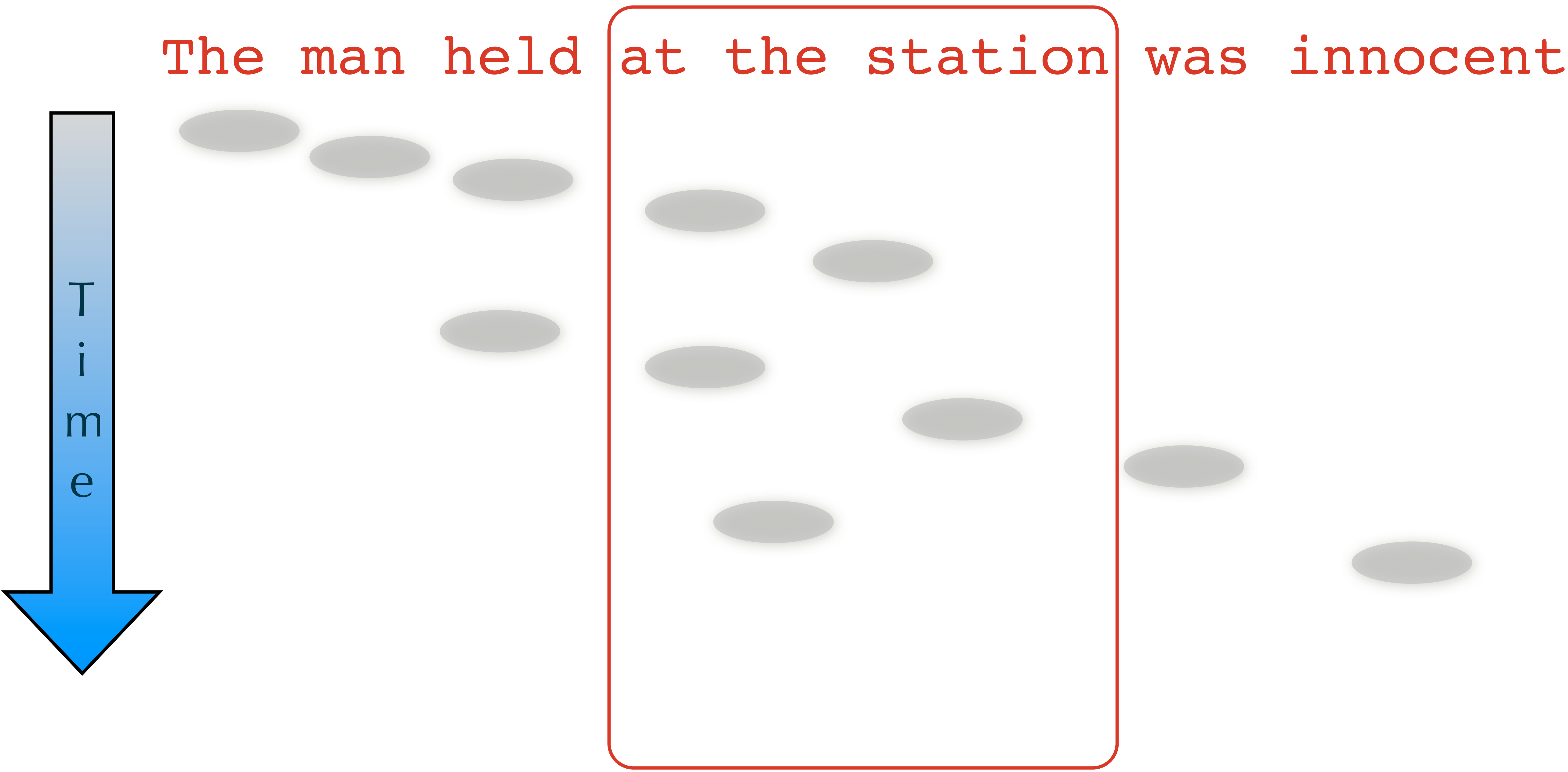
The man held at the station was innocent

- Self-paced reading, central presentation

- Self-paced reading, moving window

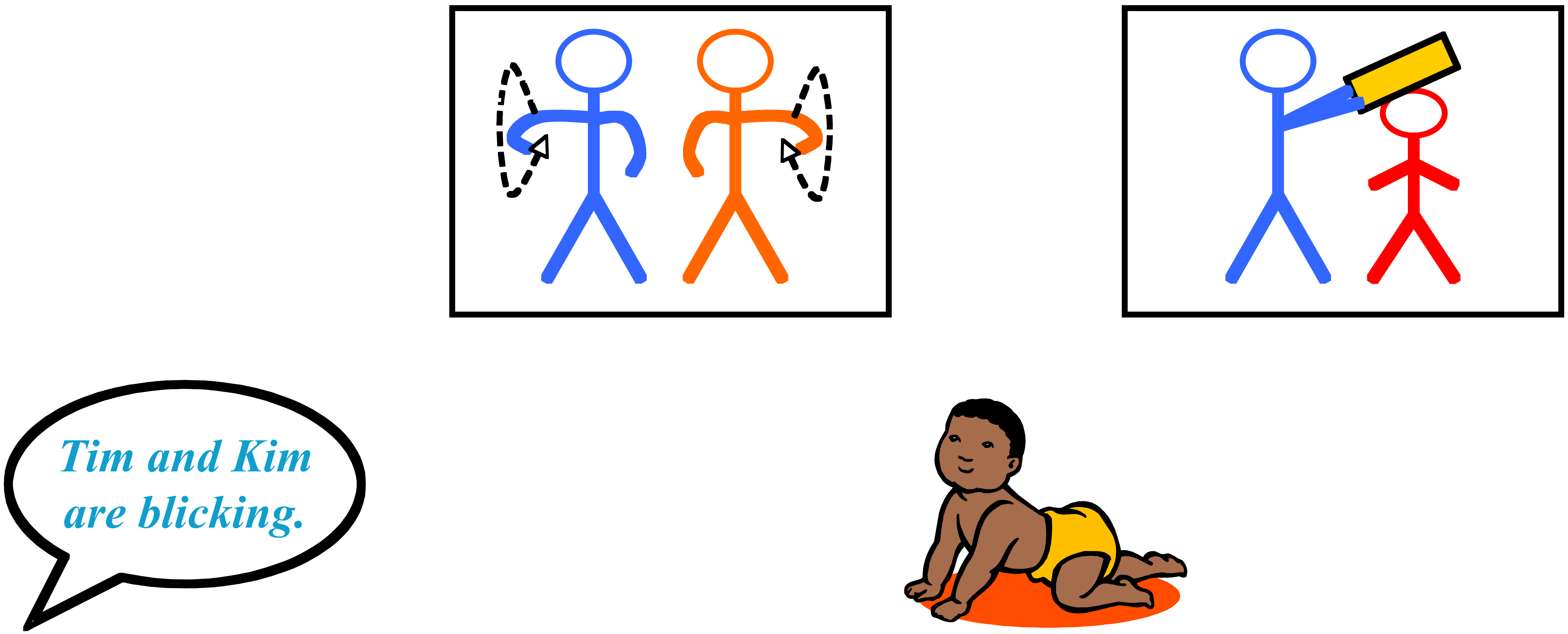
The man held at the station was innocent

Eye-tracking



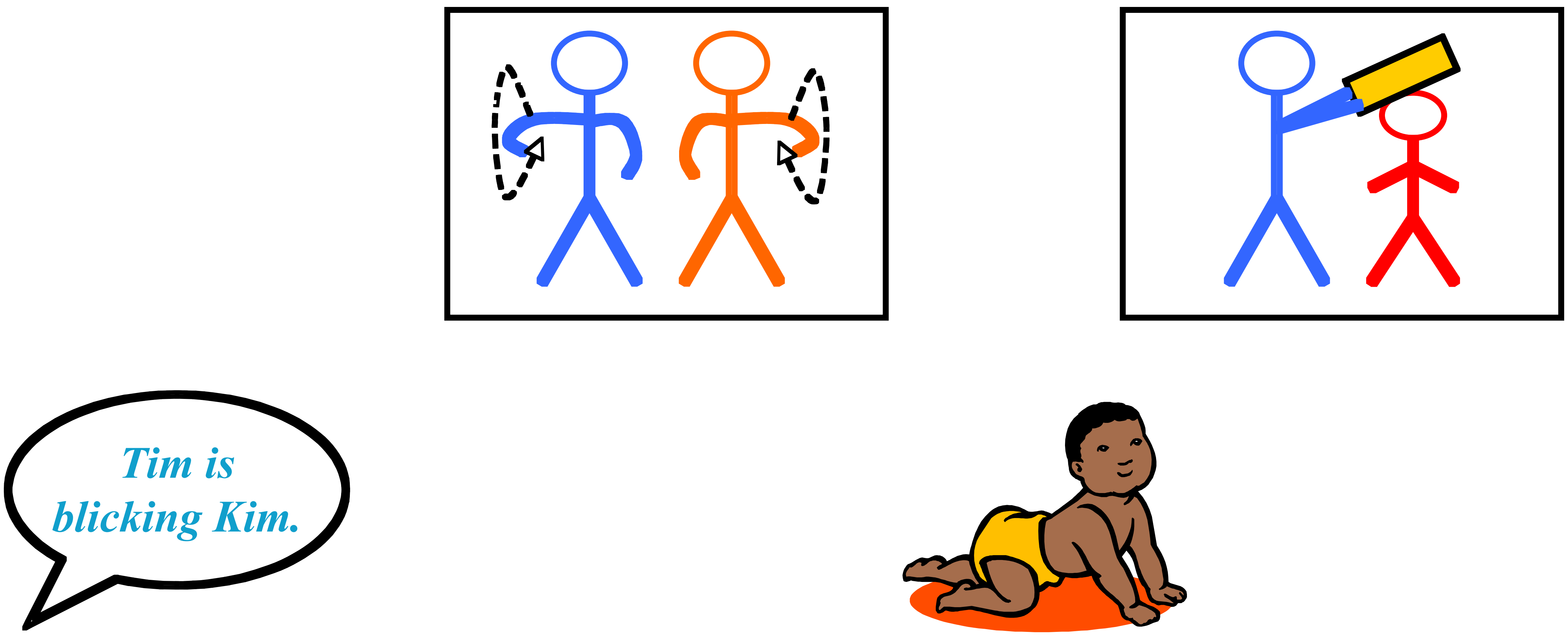
Preferential-looking Studies

- Monitor infants' preference of visual stimuli based on linguistic stimuli



Preferential-looking Studies

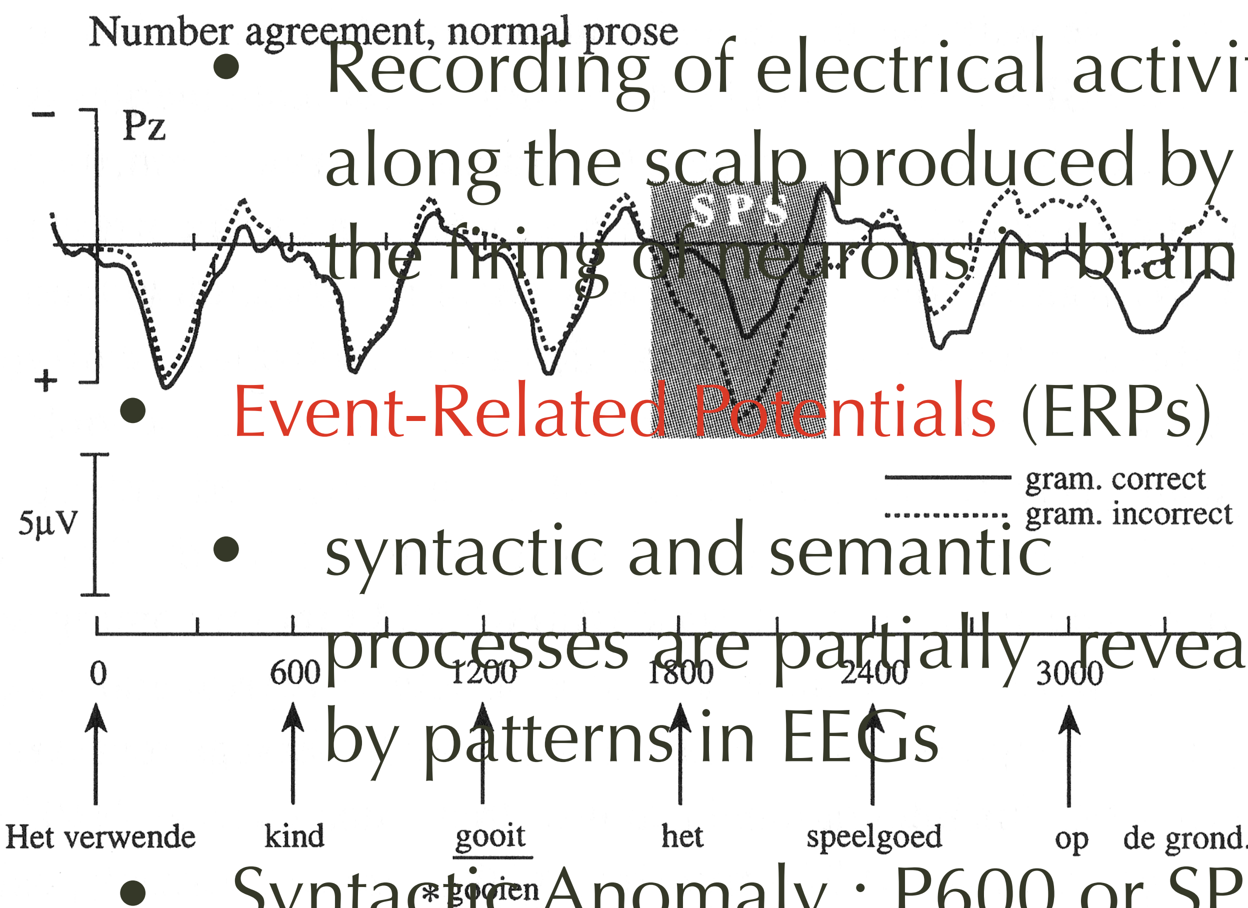
- Monitor infants' preference of visual stimuli based on linguistic stimuli

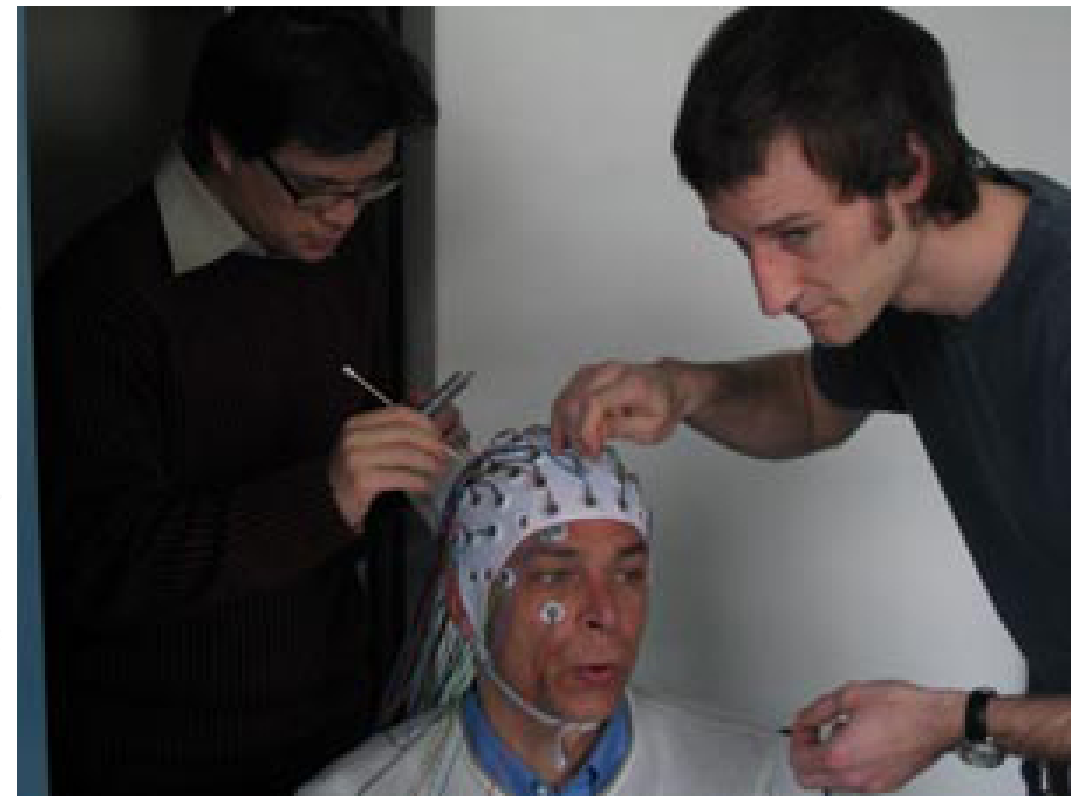


Neuroscientific Methods

Syntactic and semantic processes are partially revealed by activation patterns in brain

- Electroencephalography (EEG)
  - Recording of electrical activity along the scalp produced by the firing of neurons in brain
- Event-Related Potentials (ERPs)
  - syntactic and semantic processes are partially revealed by patterns in EEGs
- Syntactic Anomaly : P600 or SPS
- Semantic Anomaly: N400

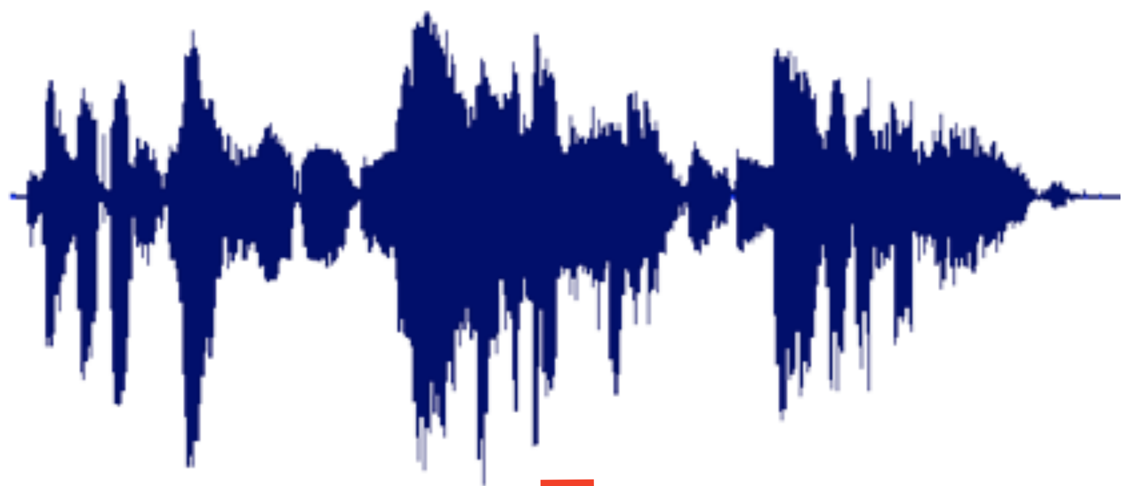




"The spoilt child throw(s) the toy on the ground"

Word Segmentation

Identifying Word Boundaries



ābigməngkēizētɪŋārɛdapəl

ā•big•məngkē•iz•ētɪŋ•ā•rɛd•apəl

Identifying Word Boundaries

امروزبایدبریمدکترواکسنبزنیمتاخوببشیم

Identifying Word Boundaries

امروز.باید.بریم.دکتر.واکسن.بزنیم.تا.خوب.بشیم.

- There are no consistent cues to word boundary in the speech signal that children receive

Supervised Word Segmentation

- Resources
  - Pre-defined lexicon
  - Manually segmented data
- Techniques
  - Match the longest possible substrings to lexicon entries
  - Use heuristics to resolve ambiguities
  - Use training data to evaluate the probabilities of different possible segmentations and choose the most probable one
- These models are useful in practice, but irrelevant to infant word segmentation

How do Infants Begin to Segment?

- Isolated words
  - About 9% of utterances directed at English-learning infants
  - Isolated words might be used to bootstrap word segmentation
- Utterance boundaries
  - Unlike word boundaries, utterances are usually marked by pause
  - Beginning and end of an utterance can guide word segmentation
- Phonological cues
  - phonotactics, allophonic variation, prosodic cues, etc
- Statistical regularities in syllable sequences found in speech

Phonological Cues

- Phonotactic constraints
  - restrictions on permissible sequences of sounds in language
  - English: no /zw/ or /vl/ at the beginning of a word (unlike Dutch)
- Prosodic characteristics
  - sound patterns of language, e.g. stress or intonation
  - strong/weak stress patterns are dominant in English
- Allophonic cues
  - auditory variants of the same phoneme in different positions
  - e.g., *nitrates* vs. *night rates*

Infants' Sensitivity

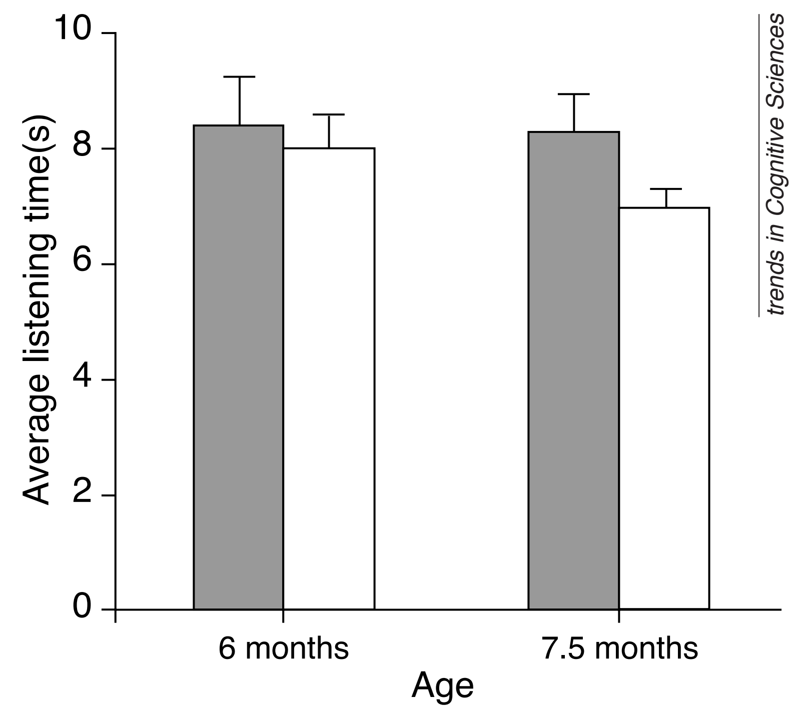


Fig. 1. Infants' segmentation of fluent English speech.

- Six-month olds are less sensitive to phonological properties of words than 7.5-month olds (Jusczyk & Aslin, 1995)
- Sensitivity to Allophonic cues develops more slowly in English learners

Distributional Cues

- Statistical regularities in the sequences of syllables found in speech can indicate word boundaries
    - Methods based on these regularities are language-independent
    - Infants as young as 7 months are sensitive to these cues

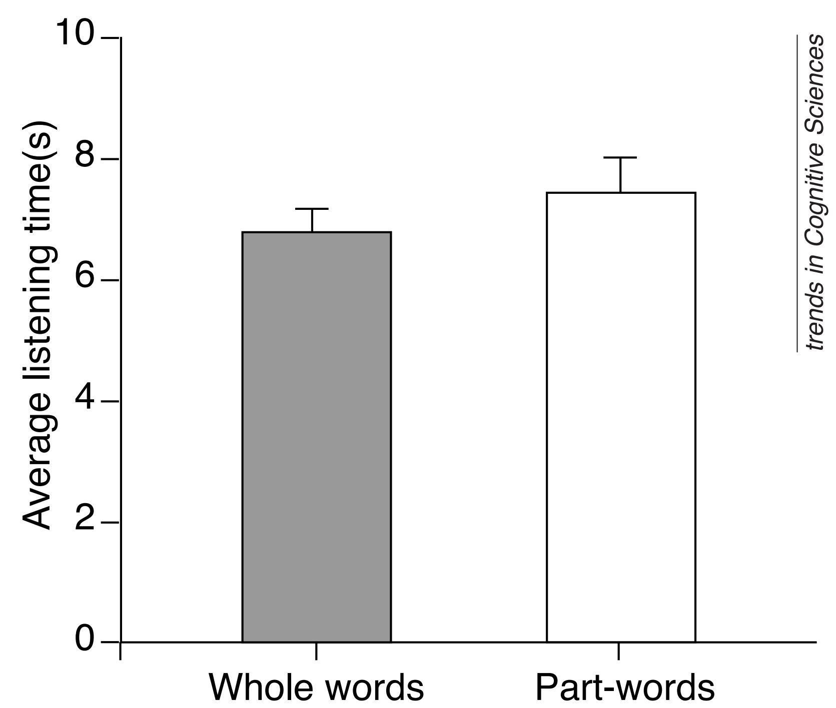


Fig. 3. Segmentation from statistical regularities in speech.

Transitional Properties

- Experimental findings suggest that children use transitional probabilities between words and syllables

big ripe apple	←	bigrīpapəl	←	bigrīpapəl
bi gripe apple				

- word level: P(*apple*|*ripe*) > P(*apple*|*gripe*)
- syllable level: P(rīp|big) > P(grīp|bi)

Unsupervised Word Segmentation

- Transitions between linguistic units within words are more predictable than transitions across word boundaries
- Other statistics measuring the degree of association between adjacent units or groups of units
  - Mutual information, n-gram frequencies, boundary entropy, etc
- General strategy:
  - calculate the chosen statistics at each possible boundary point
  - insert a boundary at every local minimum

Case Study: Harris (1955)

- Input: utterance as a phoneme sequence
- Algorithm:
  - Measure number of successors of each subsequence of the utterance
    - How many different phoneme types follow a subsequence?
  - Segment utterance at points where the number of successors reaches a peak

Case Study: Harris (1955)

- Test utterance: `/hiyzklevər/`

Phoneme subsequences	# of successors
/h/	**9**

Case Study: Harris (1955)

- Test utterance: `/hiyzklevər/`

Phoneme subsequences	# of successors
/h/	**9**
/hi/	**14**

Case Study: Harris (1955)

- Test utterance: `/hiyzklevər/`

Phoneme subsequences	# of successors
/h/	9
/hi/	14
/hiy/	29
/hiyz/	29
/hiyzk/	11
/hiyzkl/	7
/hiyzkle/	8
/hiyzklev/	1
/hiyzklevə/	1
/hiyzklevər/	28

Case Study: Harris (1955)

- Test utterance: /hiyzklevər/

Phoneme subsequences	# of successors
/h/	9
/hi/	14
/hiy/	29
/hiyz/	29
/hiyzk/	11
/hiyzkl/	7
/hiyzkle/	8
/hiyzklev/	1
/hiyzklevə/	1
/hiyzklevər/	28

Case Study: Harris (1955)

- Test utterance: /hiy.z.klevər./

Phoneme subsequences	# of successors
/h/	9
/hi/	14
/hiy/	29
/hiyz/	29
/hiyzk/	11
/hiyzkl/	7
/hiyzkle/	8
/hiyzklev/	1
/hiyzklevə/	1
/hiyzklevər/	28

Case Study: Brent (1999)

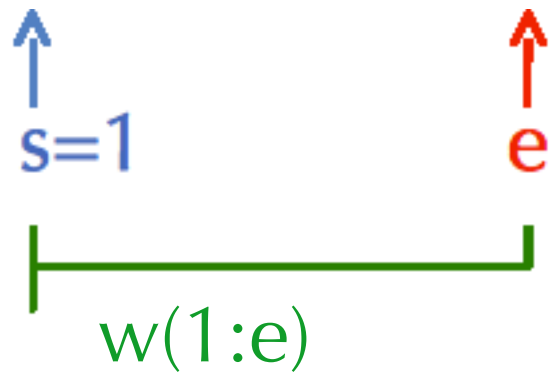
- Input: unsegmented corpus of phoneme sequences
- Approach:
  - Segment input incrementally, one utterance at a time
  - Assume words in an utterance are generated independently
    - word unigram
  - Assume phonemes in a word are generated independently
    - no phonotactics

Case Study: Brent (1999)

- At each step (t):
  - C(t-1): part of corpus segmented so far
  - U(t): current utterance
- Algorithm:
  - Hypothesize words in U(t) by considering a word-end **e** at each position
  - For each **e**, find best start **s** as the one with highest score
  - Starting from end of utterance as **e**, insert a boundary at its best start **s**

Case Study: Brent (1999)

- U(t): /yoōwänttoōsēðēboŏk/



Case Study: Brent (1999)

- U(t): `/yoōwänttoōsēðēboŏk/`

s=2 e

w(2:e)

Case Study: Brent (1999)

- U(t): /yoōwänttoōsēðēboŏk/

s=3 e

w(3:e)

Case Study: Brent (1999)

- U(t): /yoōwänttoōsēðēboŏk./

best start

e

Case Study: Brent (1999)

- U(t): /yoōwänttoōsēðē.boŏk./

best start

e

Case Study: Brent (1999)

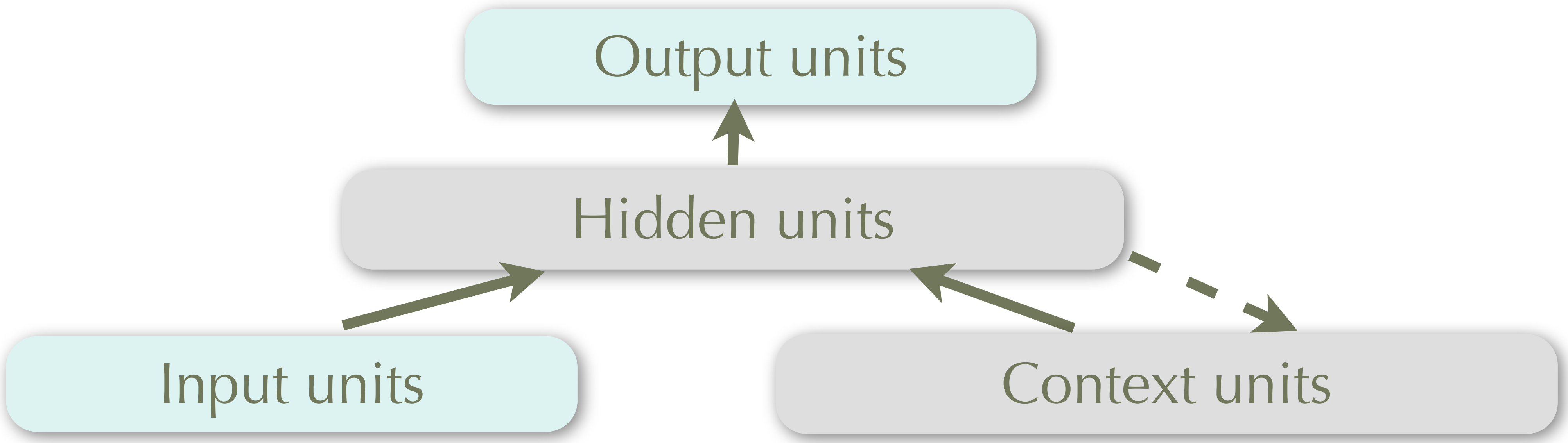
- U(t): `/yoōwänttoōsē.ðē.boŏk./`

Connectionist Models

- Neural networks have been used to segment representations of speech using distributional cues
- Input:
  - artificial corpora
  - phonological transcriptions of natural speech
- Common architecture: Simple Recurrent Network (SRN)
- Recurrence allows predictions based on context
- But it is difficult to determine exactly what part of context is useful for prediction

Case Study: Elman (1990)

Network is trained to predict
the next letter as output



input: word representation

A copy of the hidden units is kept as context

Case Study: Elman (1990)

- Input: an artificial sequence of letters

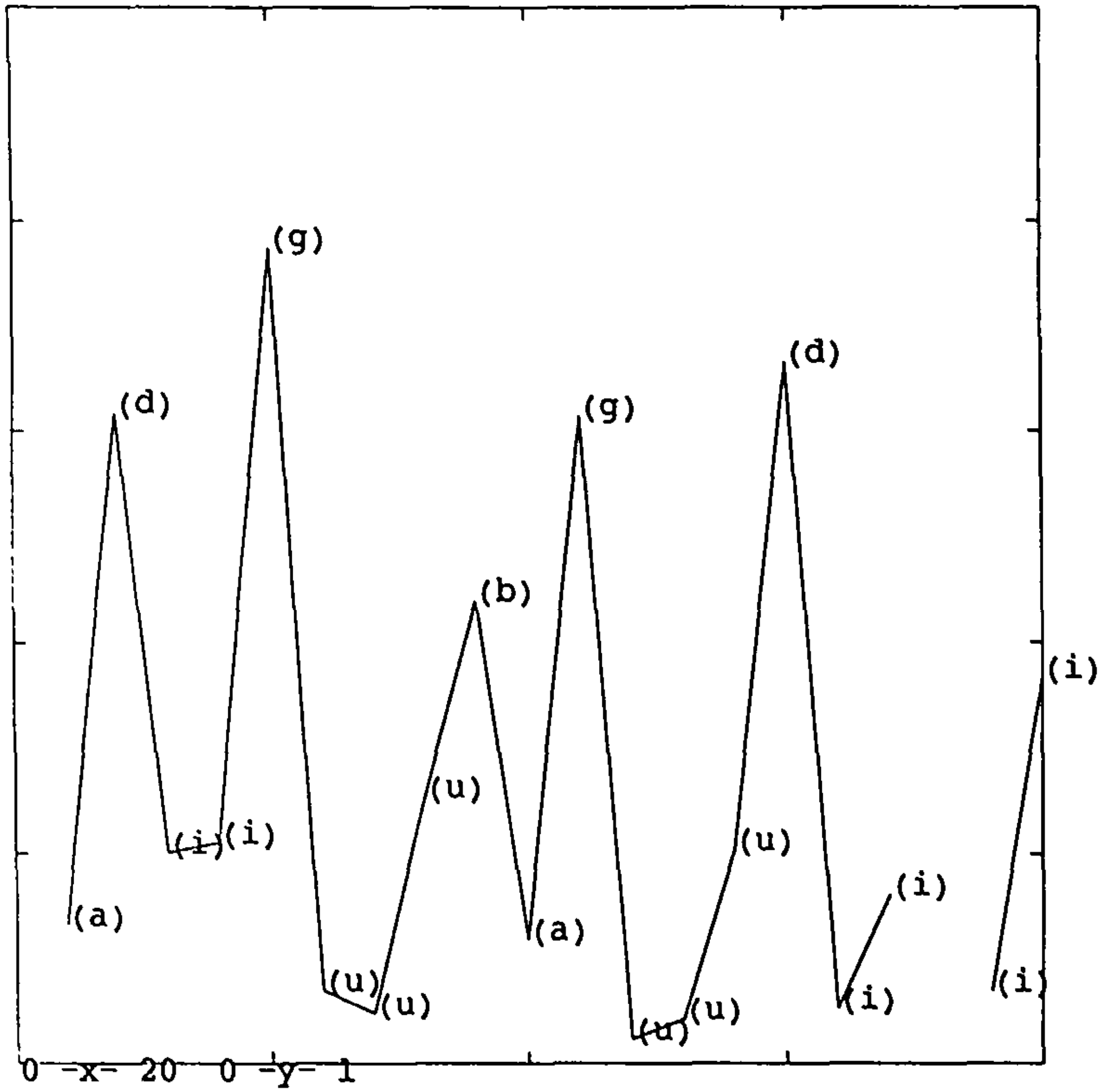
b -> ba
d -> dii
g -> guuu

- Representation of letters: vectors of phonological features

Vector Definitions of Alphabet

		Consonant	Vowel	Interrupted	High	Back	Voiced	
b	[	1	0	1	0	0	1	]
d	[	1	0	1	1	0	1	]
g	[	1	0	1	0	1	1	]
a	[	0	1	0	0	1	1	]
i	[	0	1	0	1	0	1	]
u	[	0	1	0	1	1	1	]

Case Study: Elman (1990)

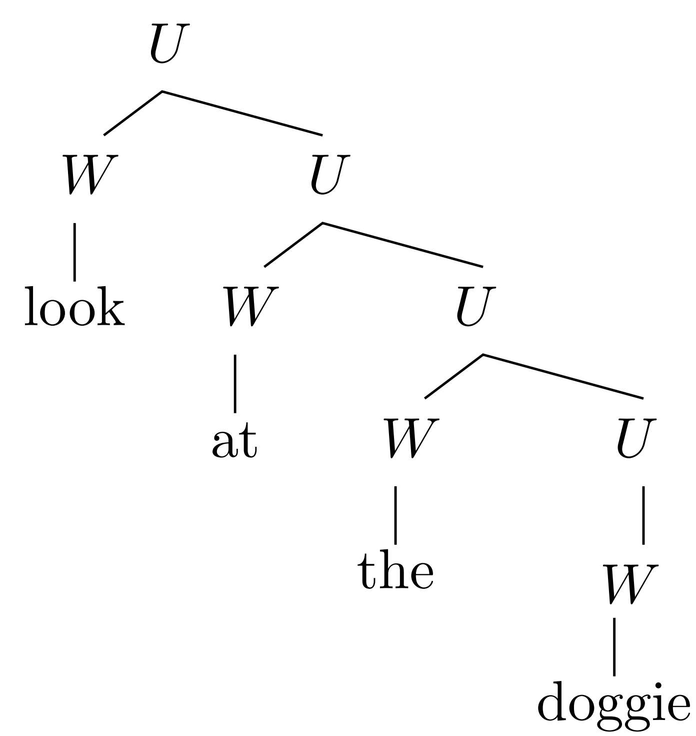


Association-based Models: Limitations

- Input representations in different models are usually not comparable
- Utterance boundaries are essential to learning, but infants can segment without utterance boundaries
- The assumption that words are generated independently of each other is limiting, and affecting the performance
  - Natural language displays many complex dependencies
- These models use unprincipled methods of constraining the number of parameters (words)
  - A better way is by using a Bayesian prior

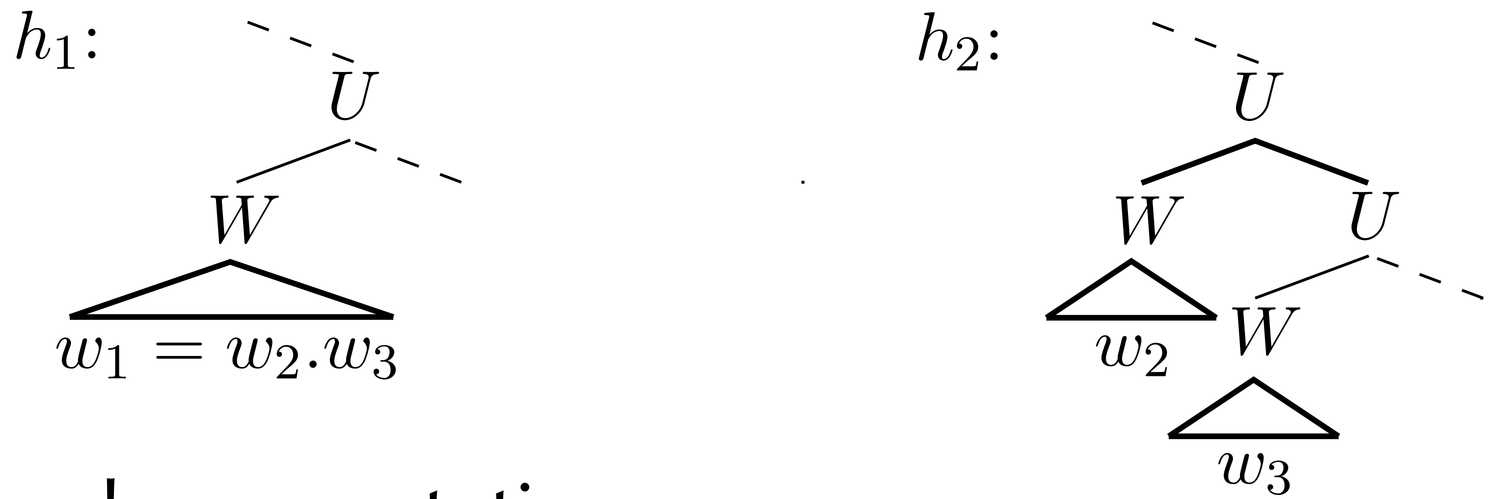
Bayesian Models

- The input phoneme sequence is "generated" by a "grammar", which has a particular distribution
- the parameters of the distribution can be estimated from the generated data, that is, the observed utterances
- A hypothesized utterance:

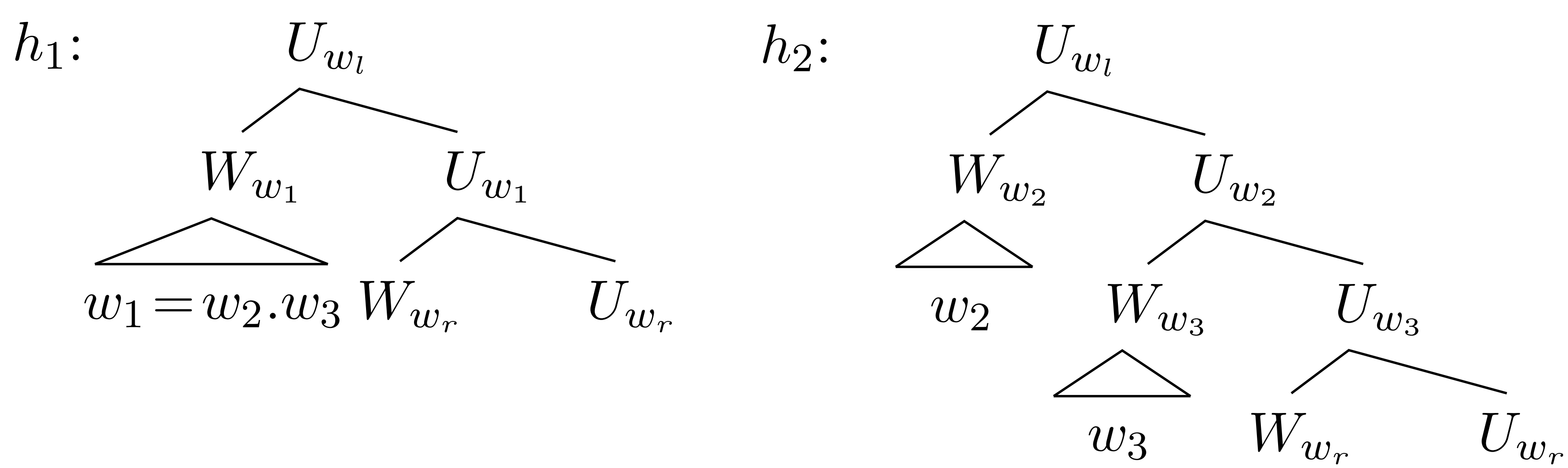


Case Study: Goldwater (2007)

- Unigram word segmentation:



- Bigram word segmentation:



Hierarchical Bayesian Models

- Findings:
  - Models incorporating a unigram assumption tend to under-segment data
  - Incorporating sequential dependencies into a model of word segmentation can greatly reduce this problem
- High transitional probabilities can occur in language
  - either because there is no word boundary
  - or because there is a boundary between two words that frequently co-occur

Open Questions

- Computational level: which information is important?
  - It seems that children use a variety of cues for segmentation
    - Phonemic cues, statistical regularities, utterance boundaries
  - But they can segment in the absence of any of these cues
- Algorithmic level: what is the most plausible strategy?
  - How are these cues combined?
  - Association-based models have poor performance
  - Bayesian models do not explain human errors