Using VWP to study speech, spoken word recognition and pitch accents

Michael K. Tanenhaus

Outline

- Basic issues
  - The problem of spoken word recognition
    - Accessing lexical knowledge in a continuous speech stream
  - Eye movement issues
    - **Linking hypothesis**
    - **Closed set concerns**
- Some empirical results
  - Frequency
    - **Real words/artificial lexicons**
  - Fine-grained sub-phonetic detail
    - Consonants
      - VOT (gradiency and perceptual learning)
    - Vowels
      - Vowel duration, prosodic domains, **lexical neighborhoods**
      - **Co-articulation**
      - Voice information
      - Pitch accents
  - Distributed perceptual representations
    - Perceptual features (behavioral and brain imaging)
  - Context

Spoken word recognition: the input

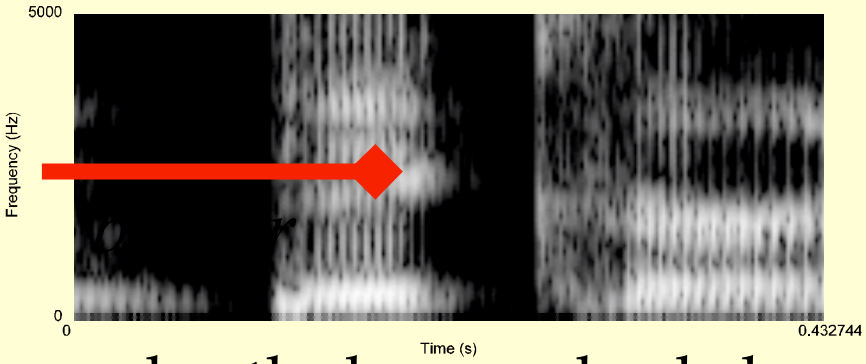
Spoken words unfold as a series of transient acoustic events extending over a few hundred ms, without reliable cues to word boundaries.

Imagine reading this page through a two-letter aperture, the text scrolling past without spaces separating the words, at a variable rate one could not control, with the visual features for each letter arriving asynchronously.

Consequences:

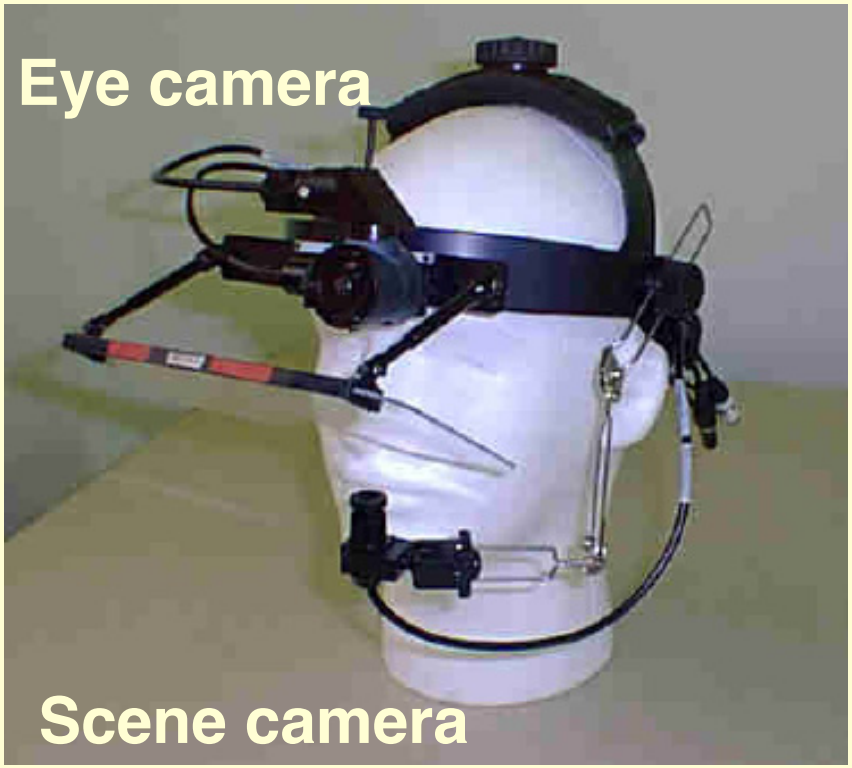
Asynchronous cues (VOT/vowel duration)

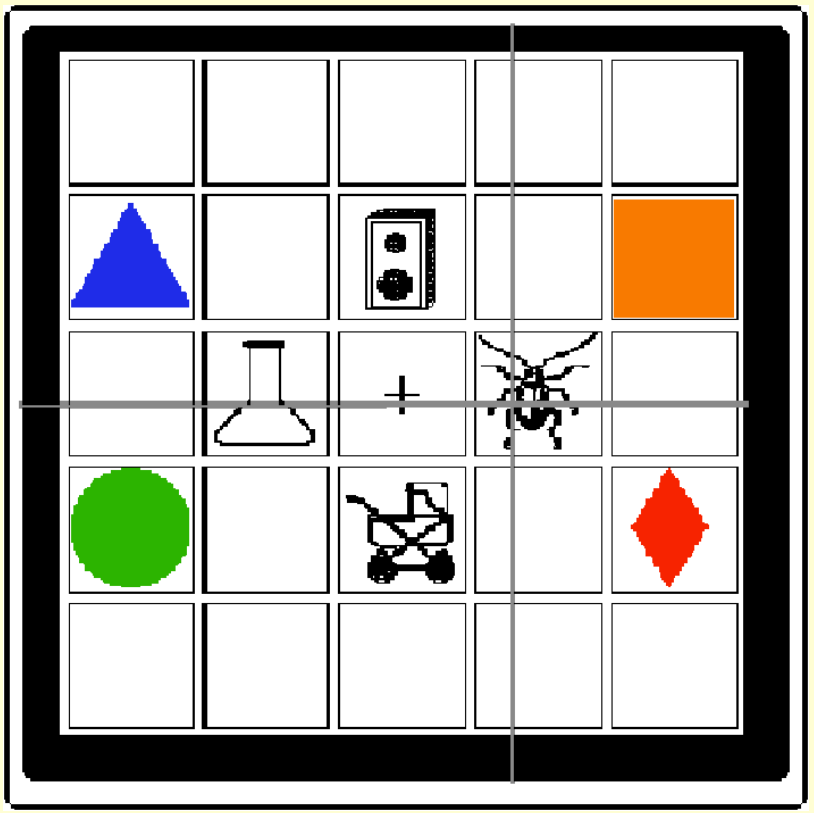
Lexical representations acoustically similar to the unfolding input must serve as a temporary memory of the input and as a set of candidate hypotheses.



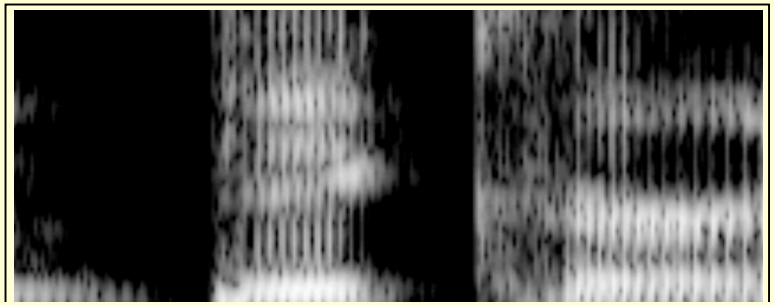
beetle, beacon, beak, beep…

Allopenna, Magnuson & Tanenhaus (1998)

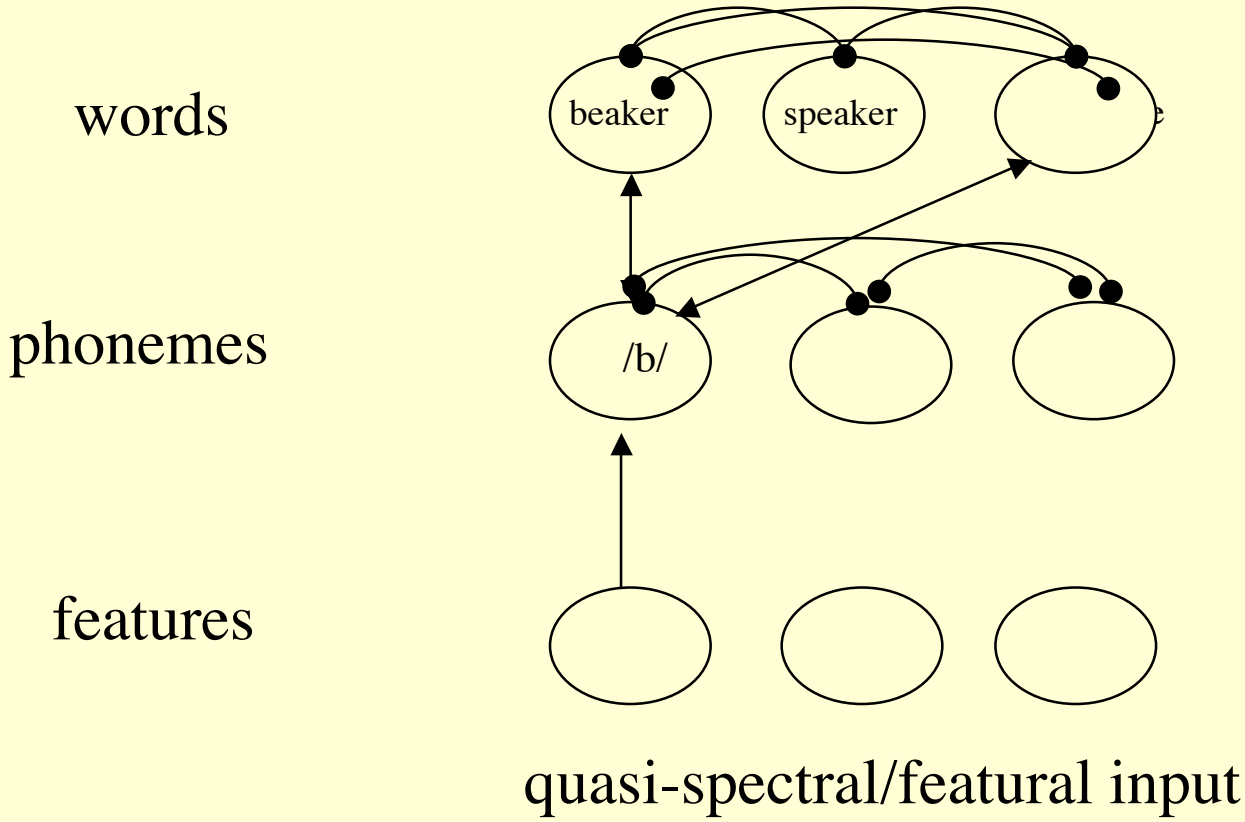




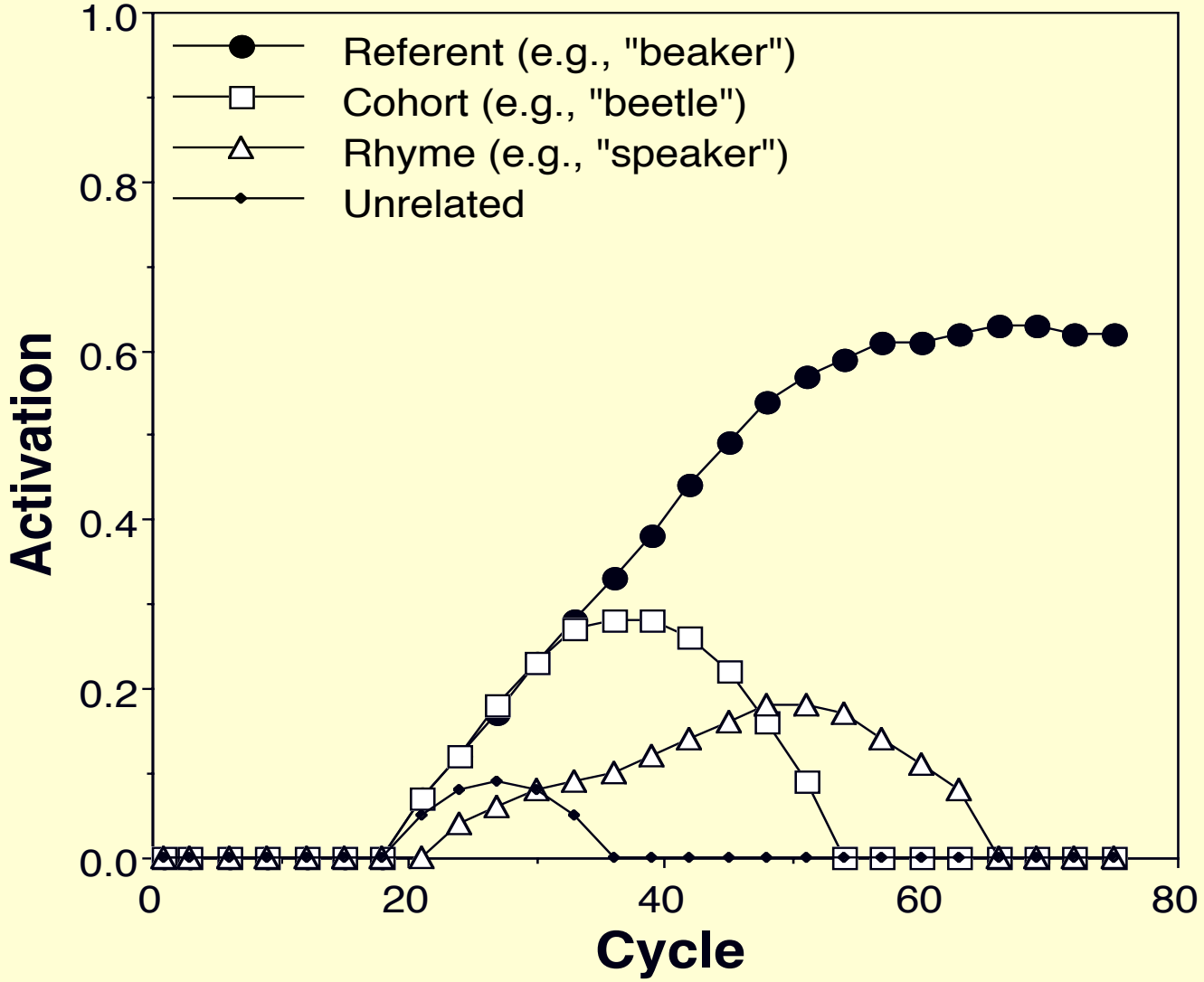
Pick up the beaker



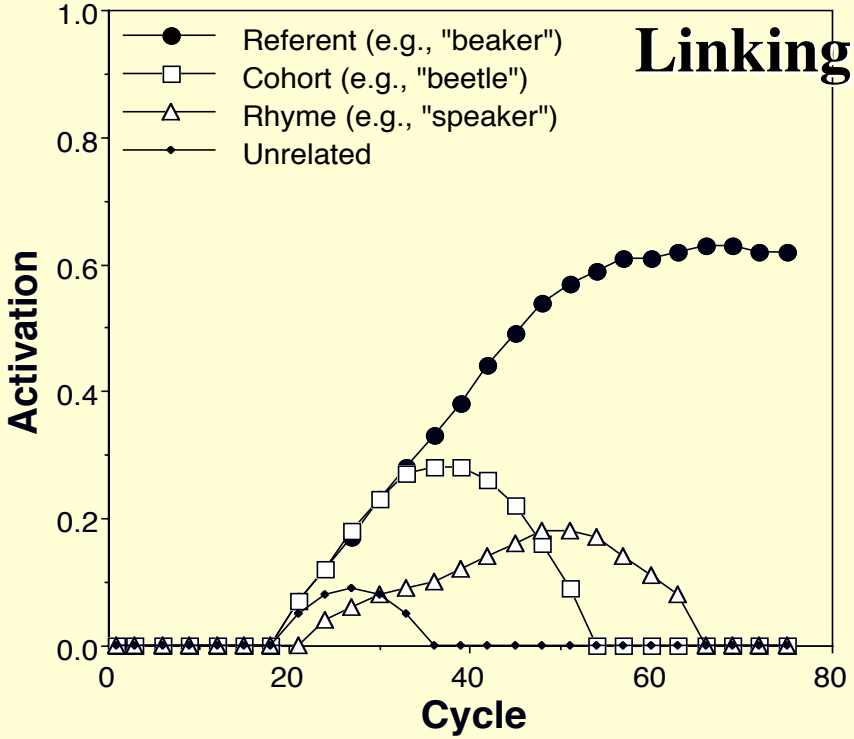
TRACE architecture



TRACE activation functions for target, cohort, rhyme and unrelated competitors



Linking hypothesis



- Activation converted to probabilities using the Luce (1959) choice rule

$$S_i = e^{ka_i}$$

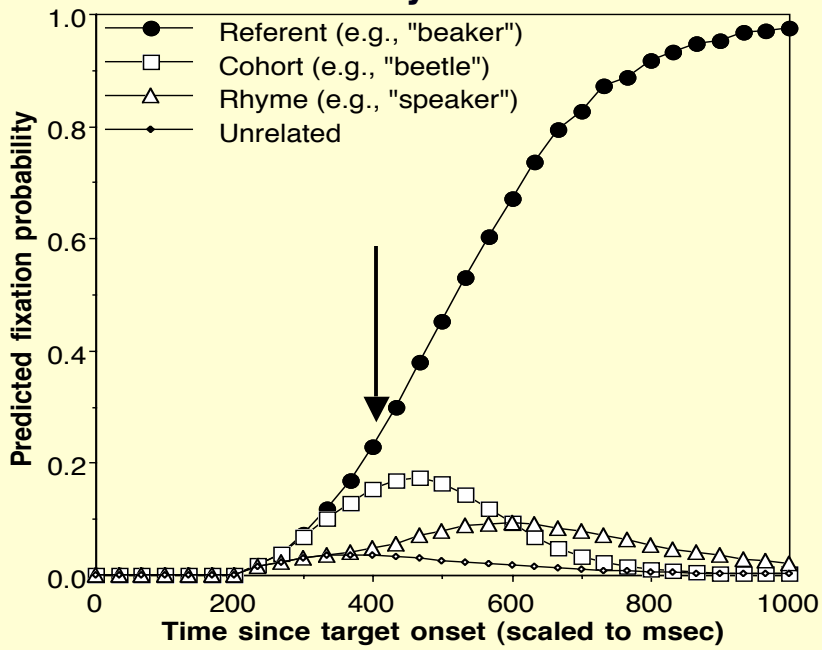
S: response strength for each item
a: activation
k: free parameter, determines amount of separation between activation levels (set to 7)

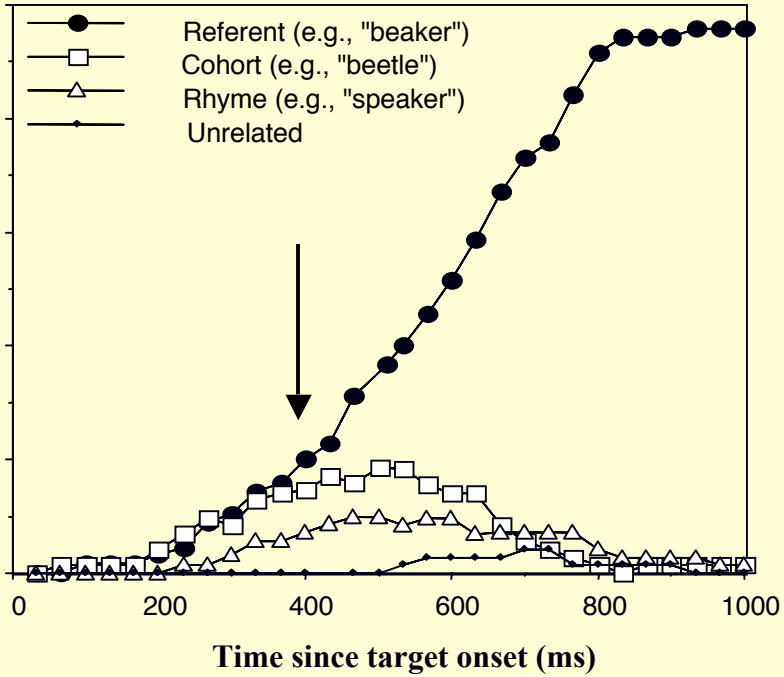
$$L_i = S_i / \sum S_i$$

Choice rule assumes each alternative is equally probable given no information; when initial instruction is *look at the cross* or *look at picture X*, we scale the response probabilities to be proportional to the amount of activation at each time step:

$$d_i = \text{max\_act}_i / \text{max\_act\_overall}$$

$$R_i = d_i L_i$$





What representations are involved in linking word to picture/object?

Some candidates:

sound to sound (picture name)

shape to shape (picture shape--words activate perceptual/motor representations

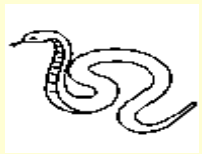
conceptual to conceptual

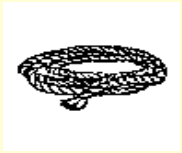
Pre-naming vs shape to shape

Are participants naming the pictures?

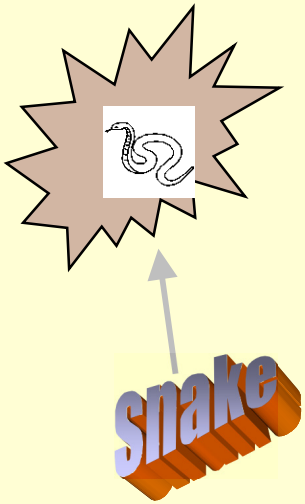
Assumption: Lexical representations include perceptuał conceptual information that is activated from earliest moments of lexical access

Prediction: Early fixations should show "visual" competitor effects

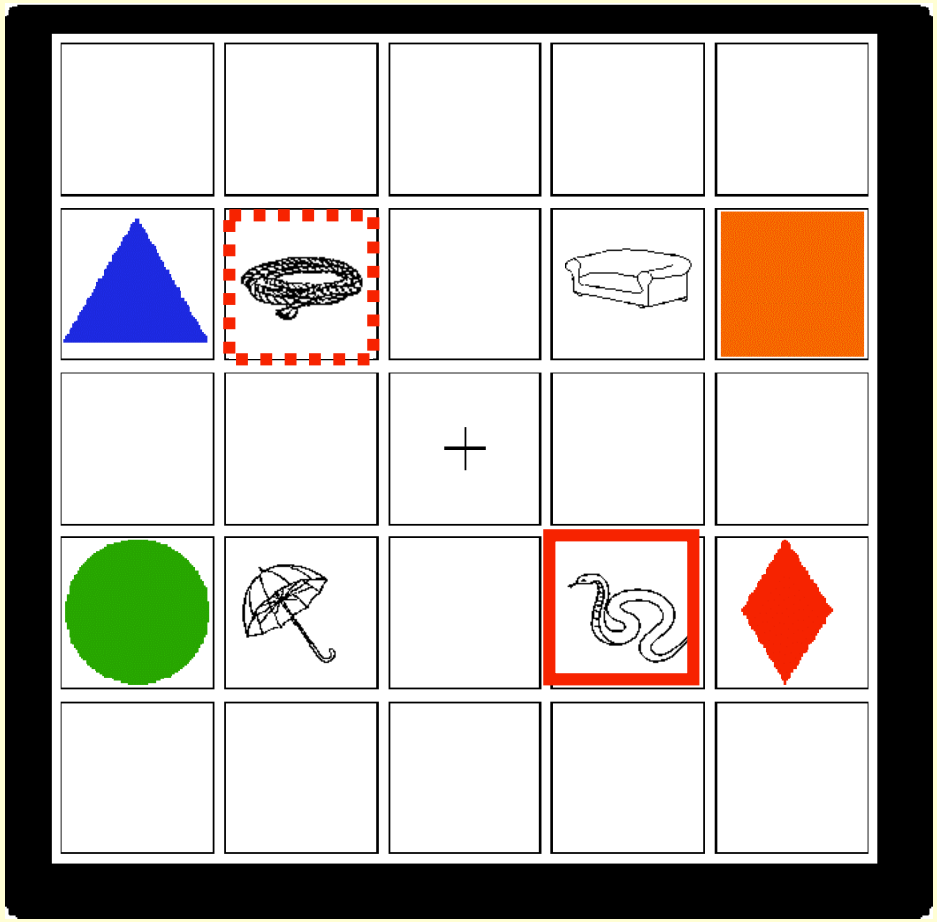




Would not predict this if people were naming the pictures



Dahan and Tanenhaus, 2005, *PBR*

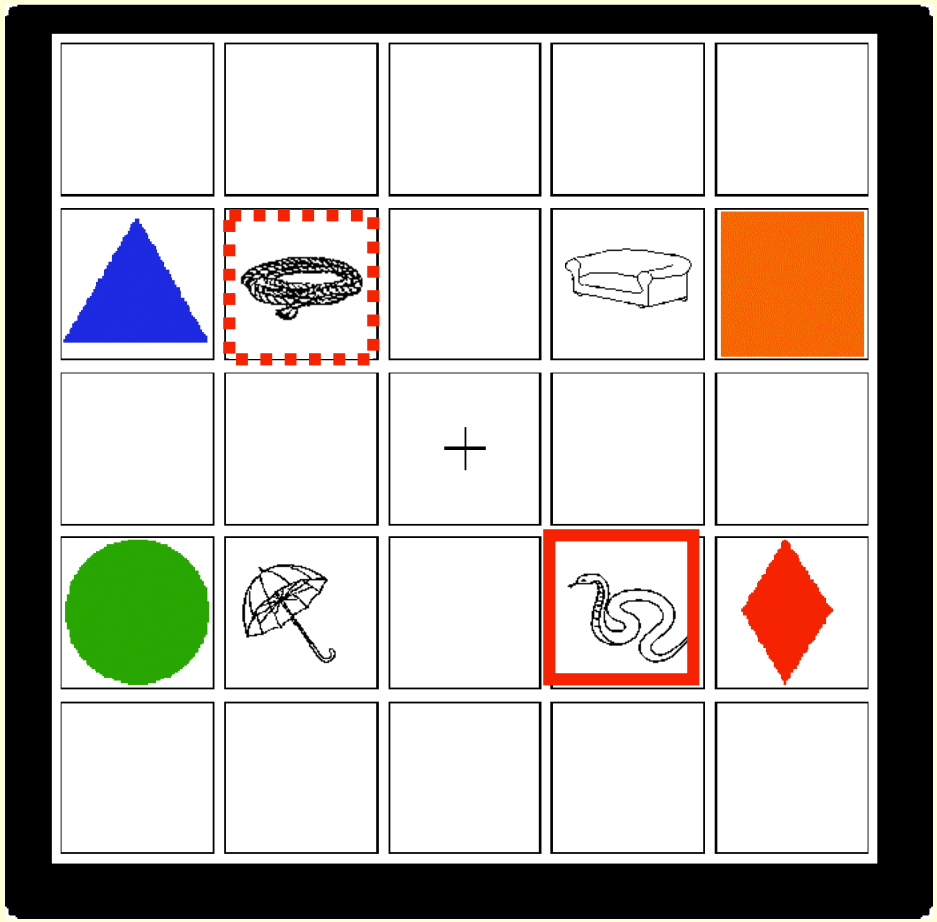


How materials were constructed:

(1) non-prototypical* picture of target

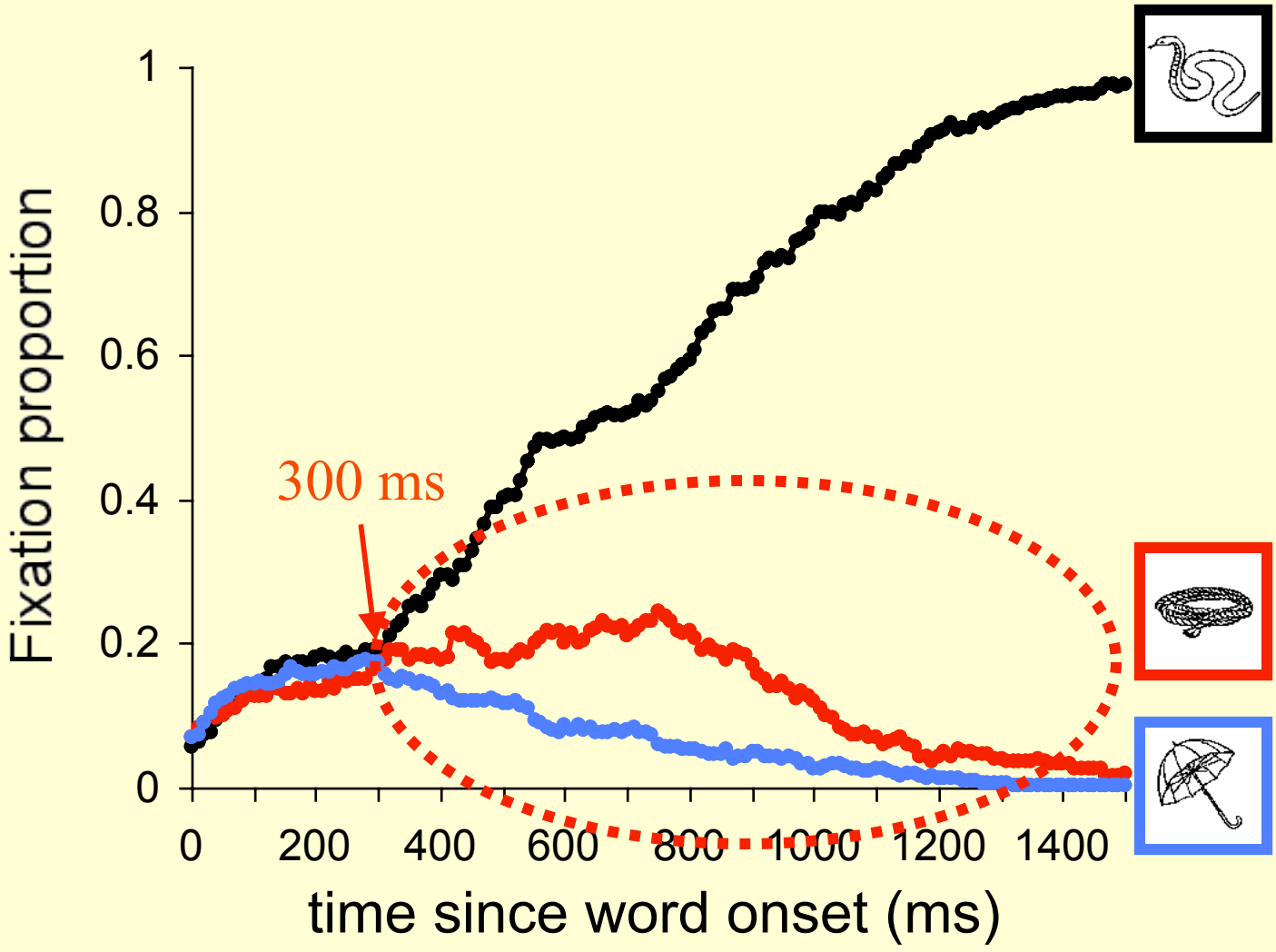
(2) visual competitor similar to prototype*

* DD-Prototypes



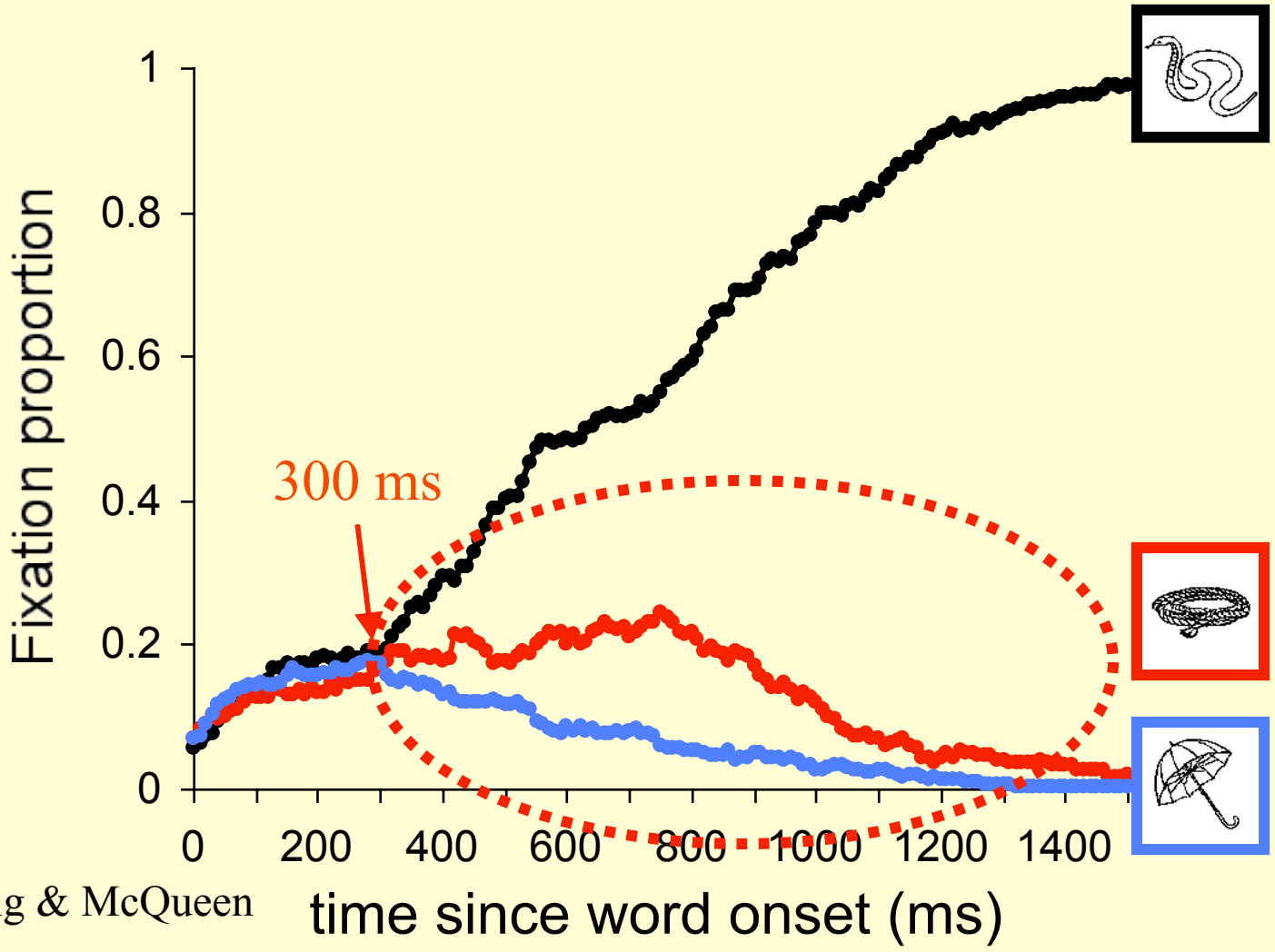
- Participants heard the name of the target picture in isolation (e.g., “slang”, *snake*)
- 300 or 1000 ms of pre-exposure

Results (300 ms)



Look contingent analysis:

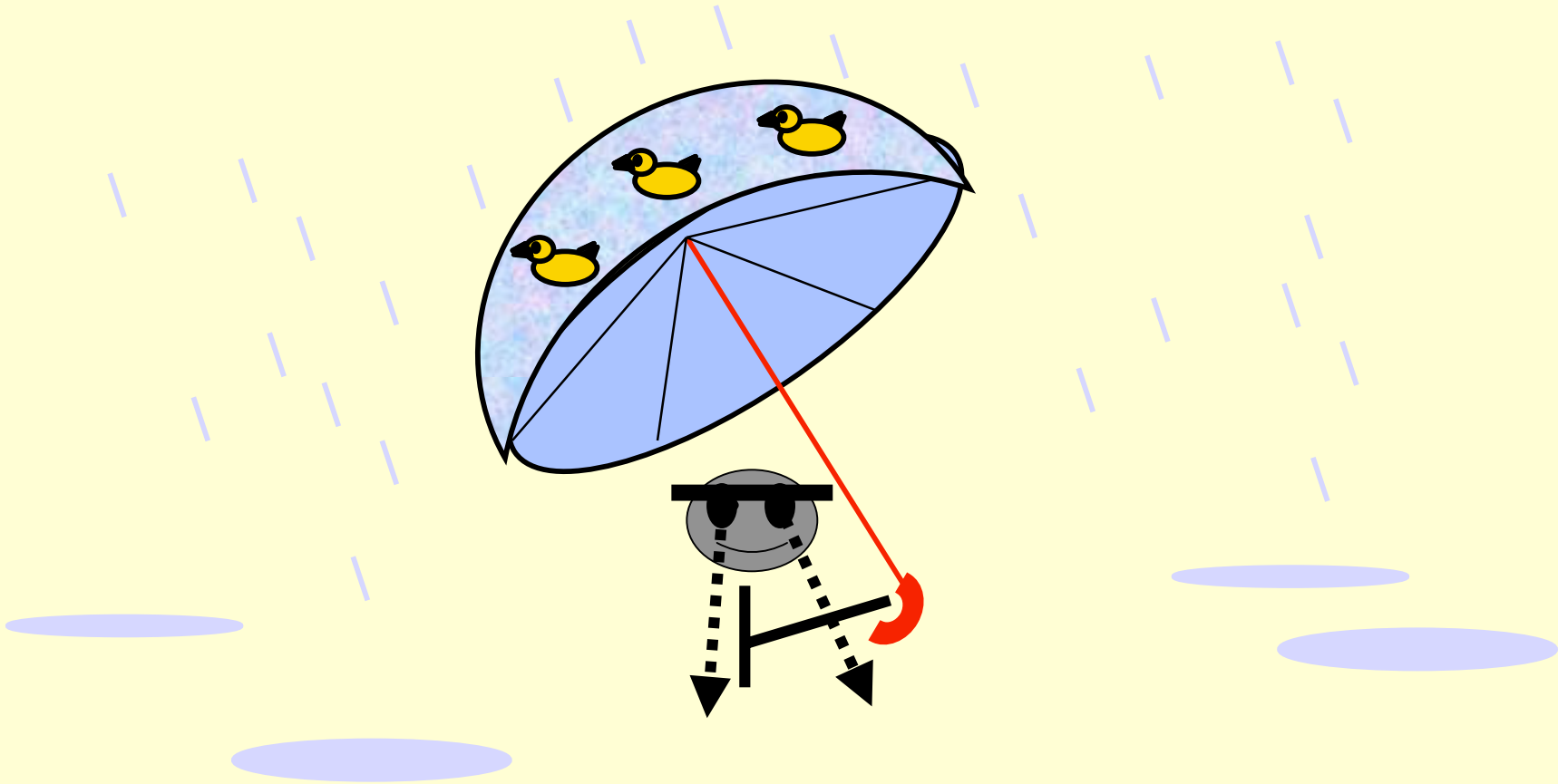
Same results with 1000 ms preview for trials where people had looked at the shape competitor ***before*** the onset of the word

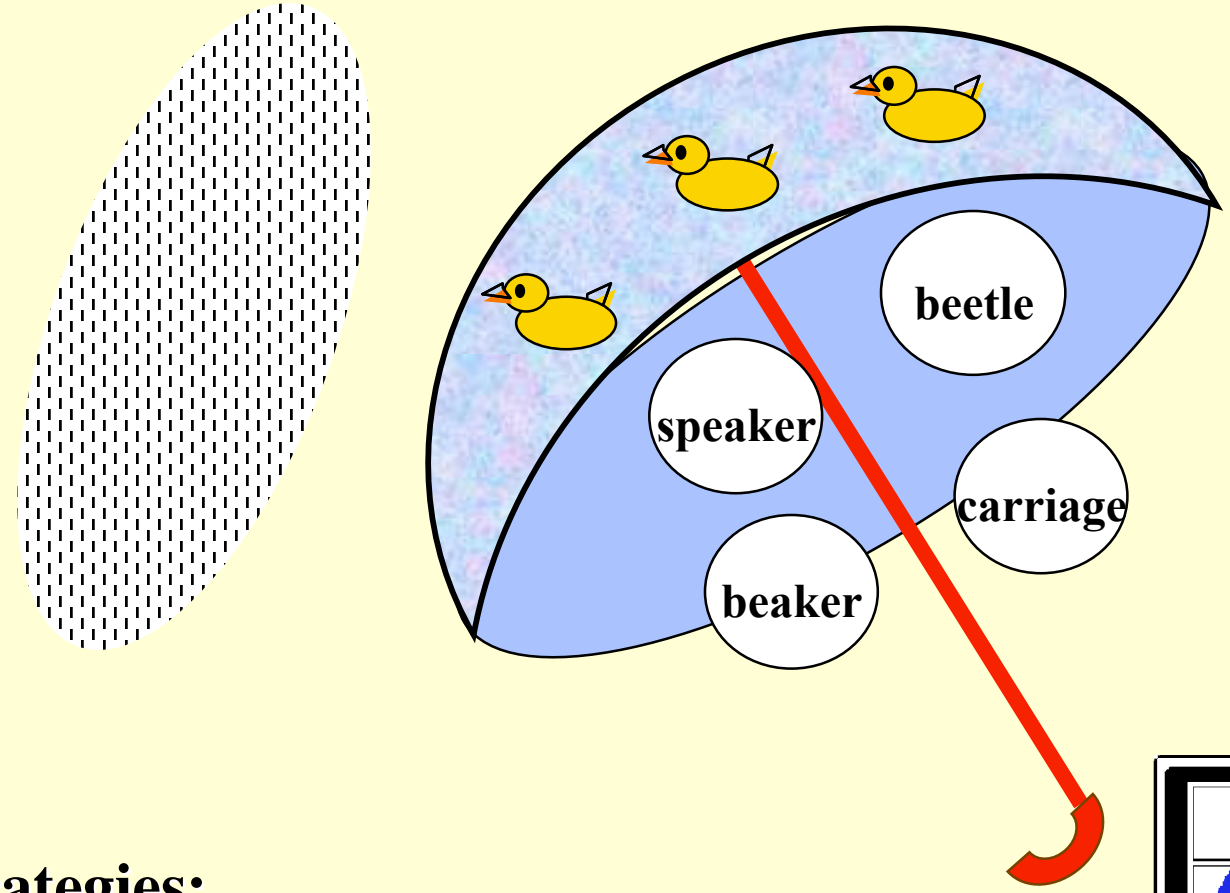


But cf. Huettig & McQueen (in press)

The closed set problem

Mike's blinders



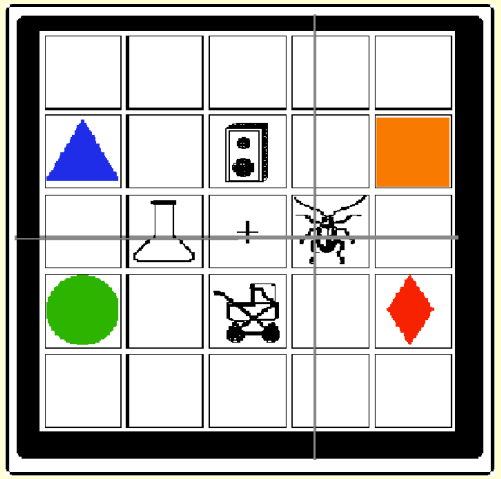


Strategies:

naming

(inconsistent with D&T, 05 PBR)

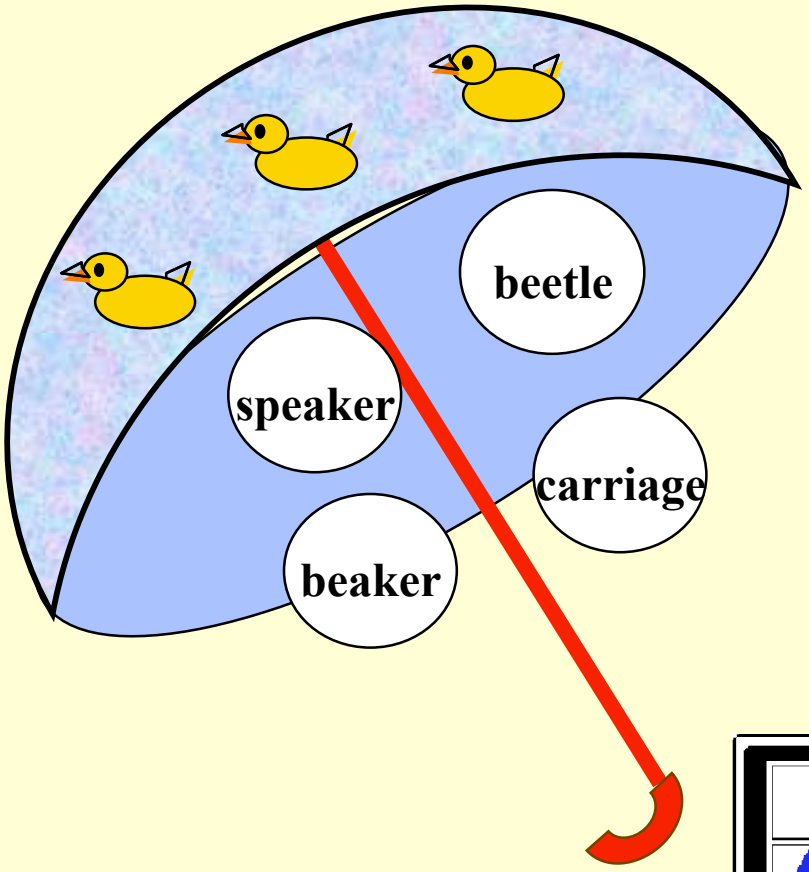
Reduced sensitivity outside of set



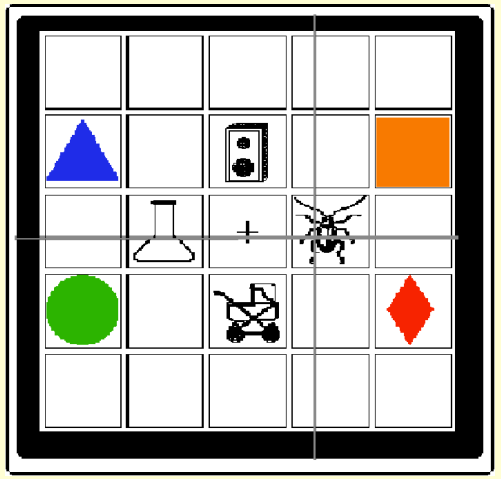
bead

beagle

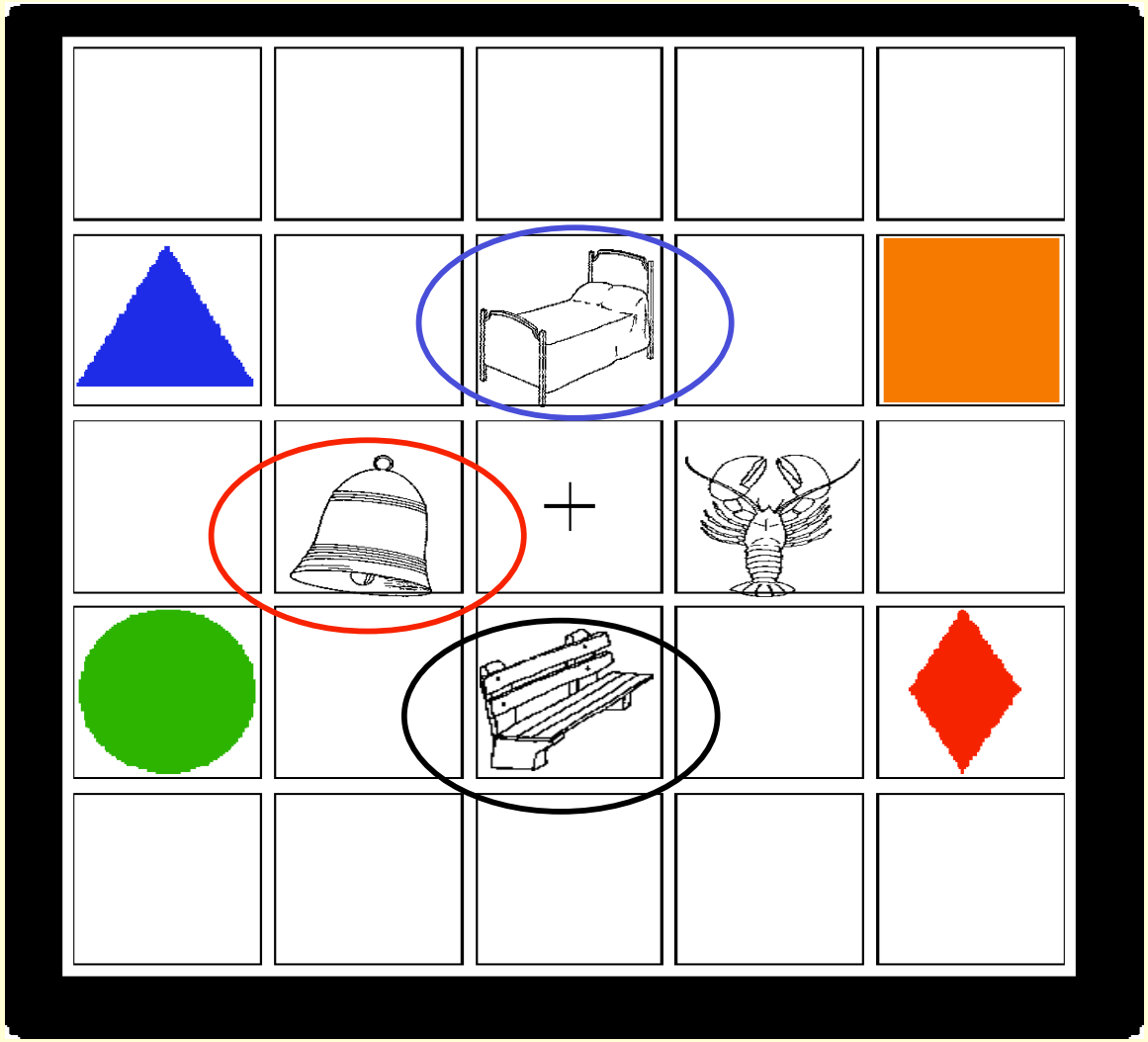
beast



Frequency?

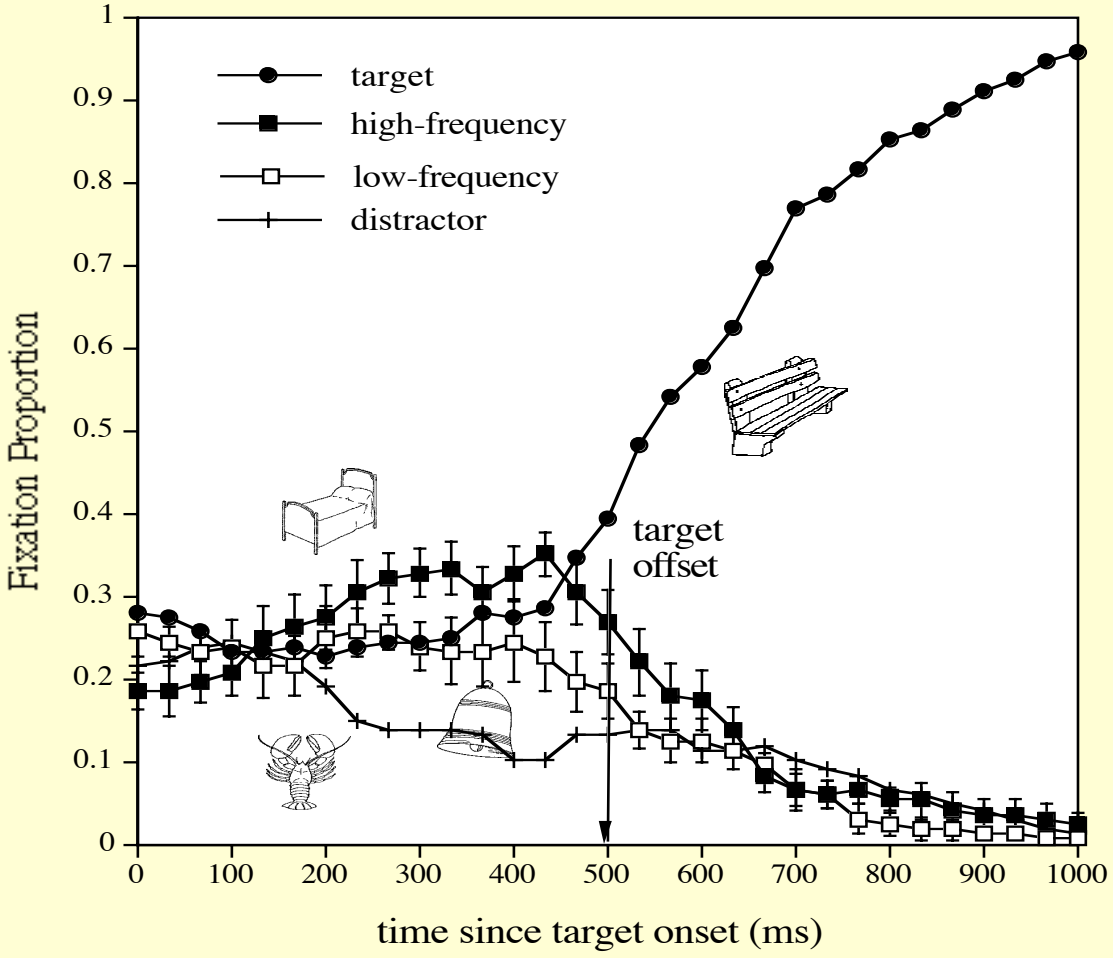


Frequency

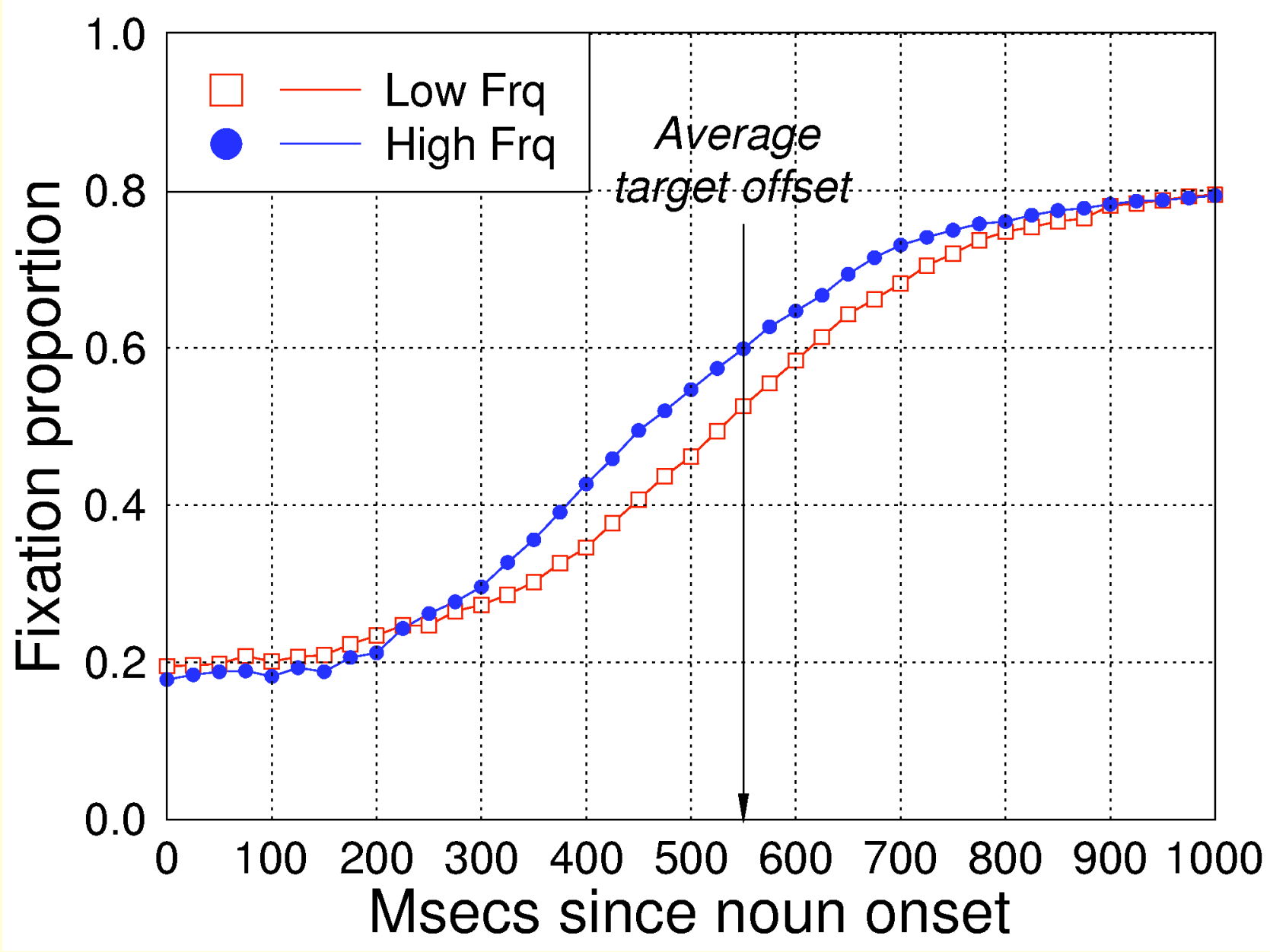


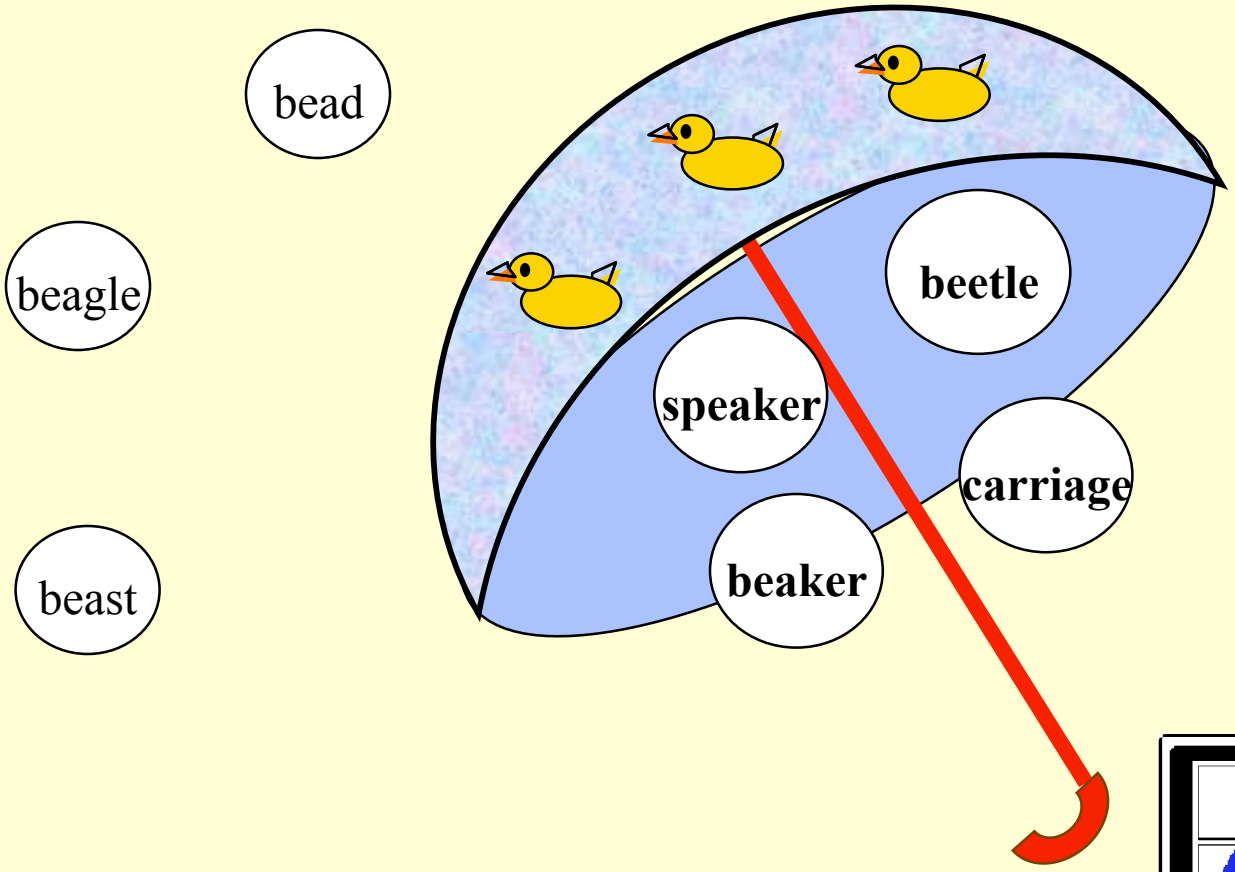
Dahan, Magnuson & Tanenhaus, *Cognitive Psychology*, 2001

High and low frequency cohorts

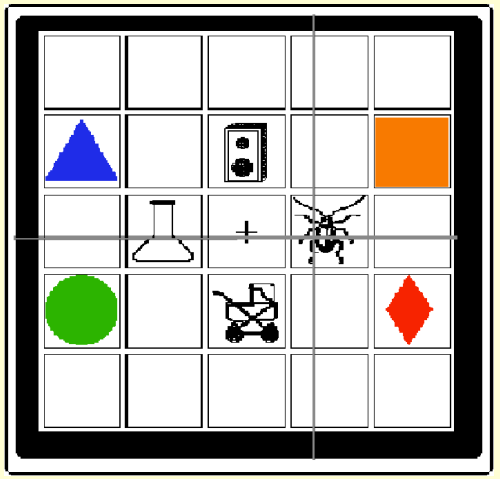


Frequency (Magnuson et al., 2007, Cognitive Science)





What about non-displayed, not-mentioned candidates?



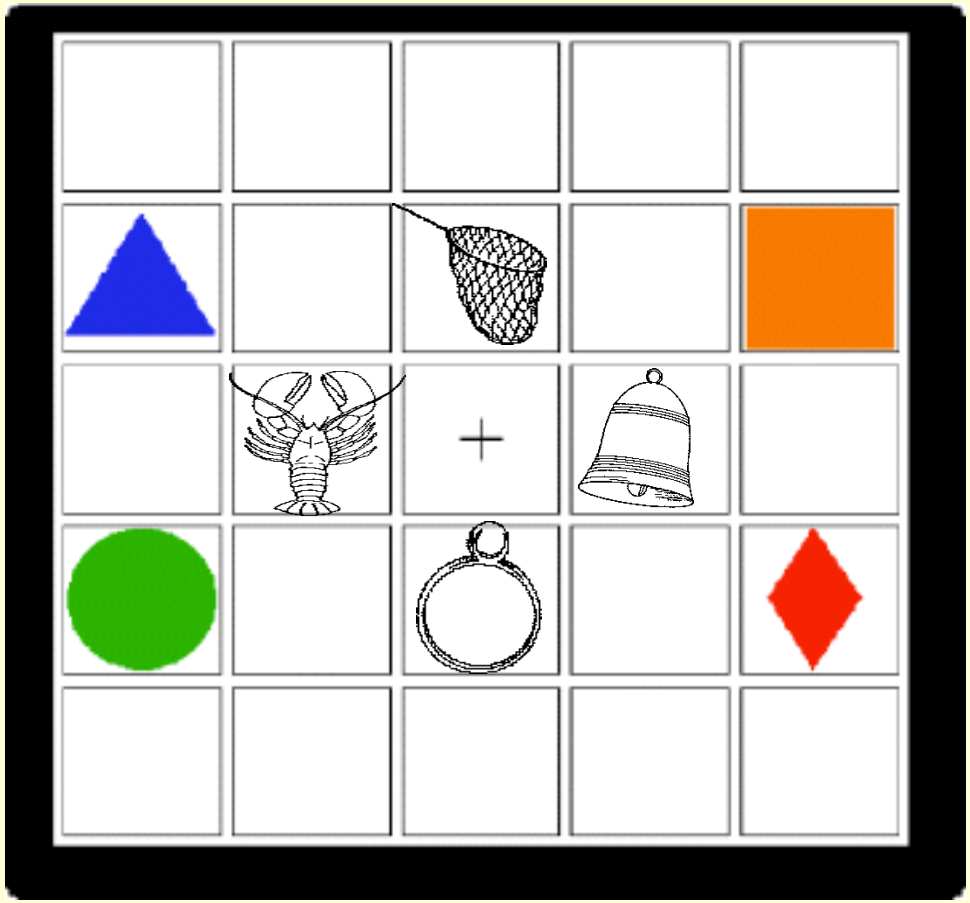
X-spliced stimuli with misleading coarticulatory information

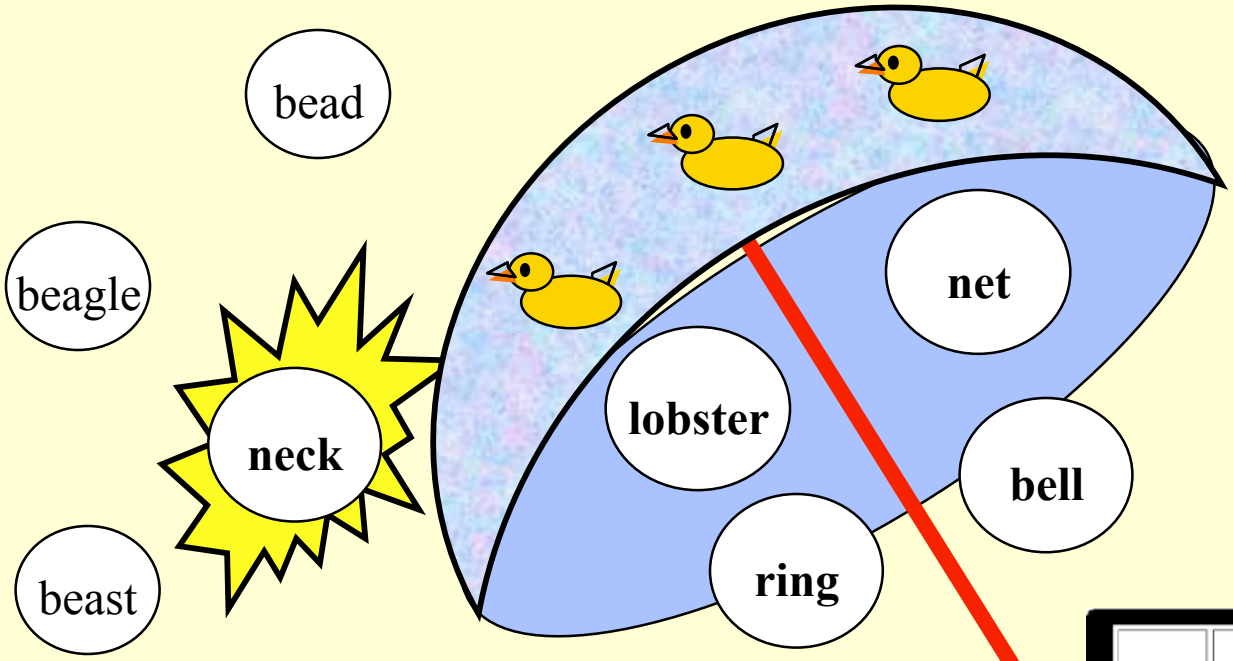
- $NE_{(t)}T$ (word from itself)
  - (net cross-spliced using the onset and vowel from another token of net)
- $NE_{(k)}T$ (word from another word)
  - (net cross-spliced using the onset and vowel from neck)
- $NE_{(p)}T$ (word from a nonword)
  - (net cross-spliced using the onset and vowel from nep)

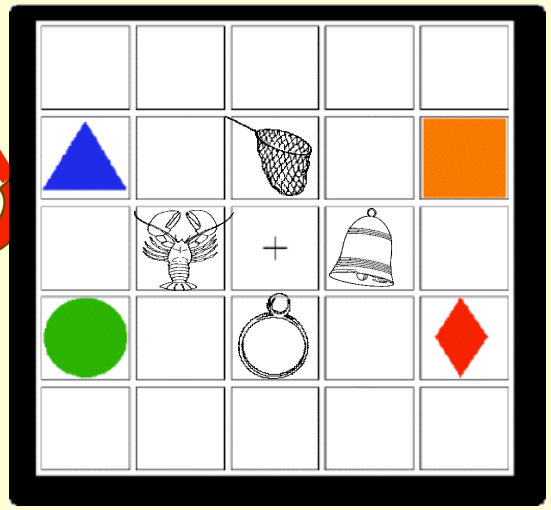
Dahan, Magnuson, Tanenhaus & Hogan, LCP, 2001

Lexical competitor not mentioned or displayed

Click on the { ne$_{(k)}$t / ne$_{(p)}$t / ne$_{(t)}$t }





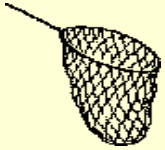


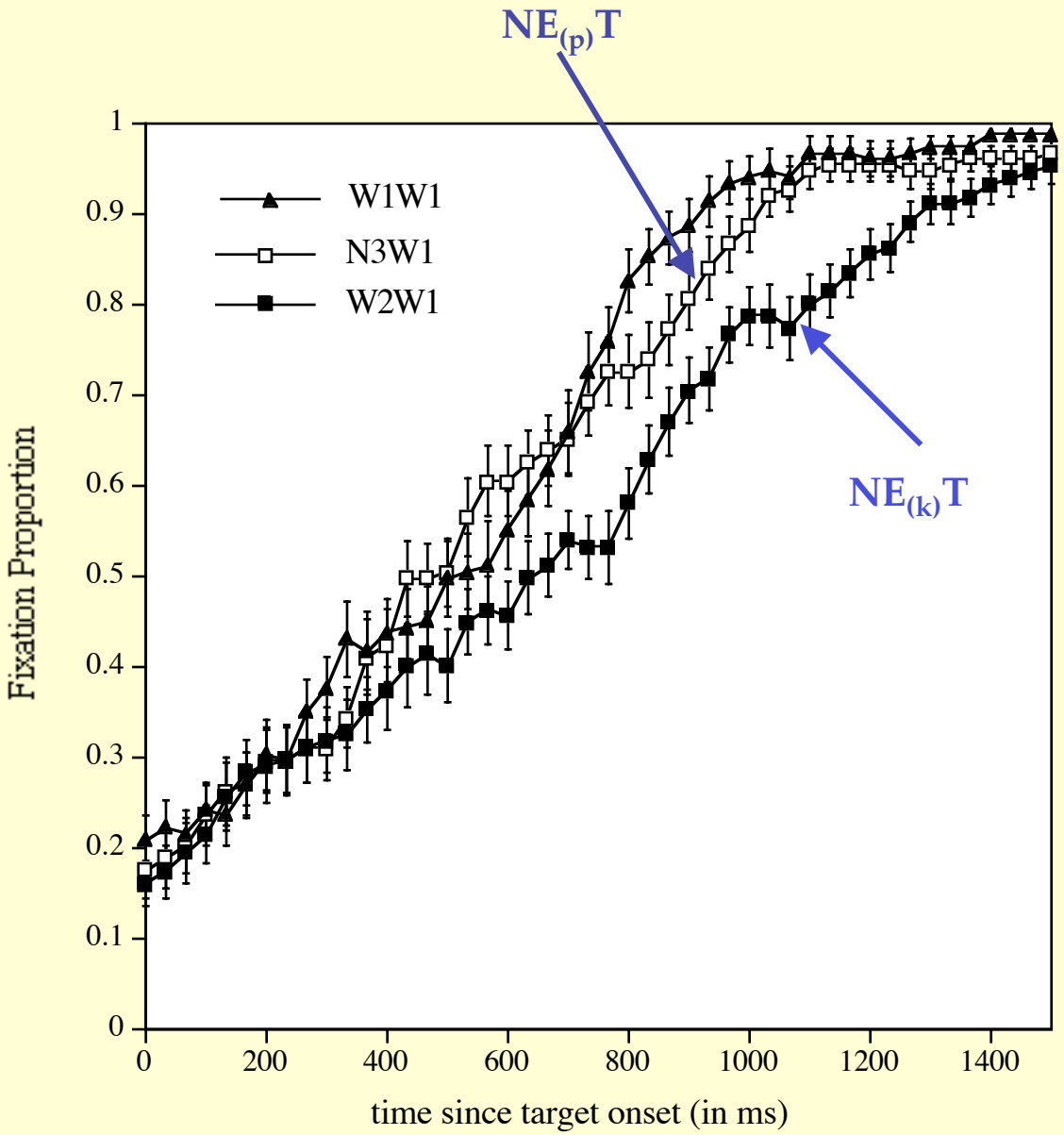
Prediction: Delayed target looks to the net with

$NE_{(k)}T$ compared to $NE_{(p)}T$

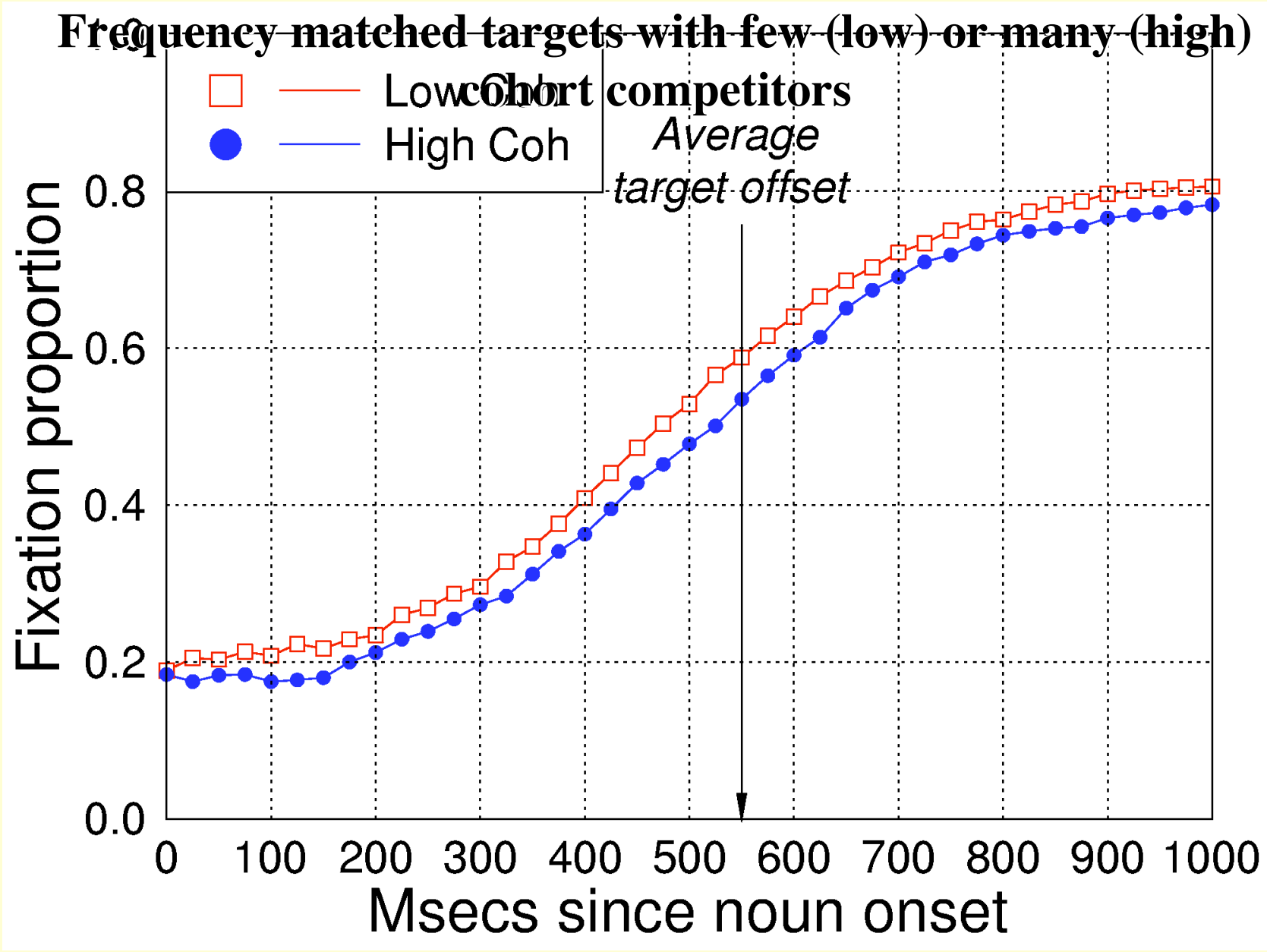
Figure 1

We see an effect of lexical status indicating that activation is affected by the full lexicon

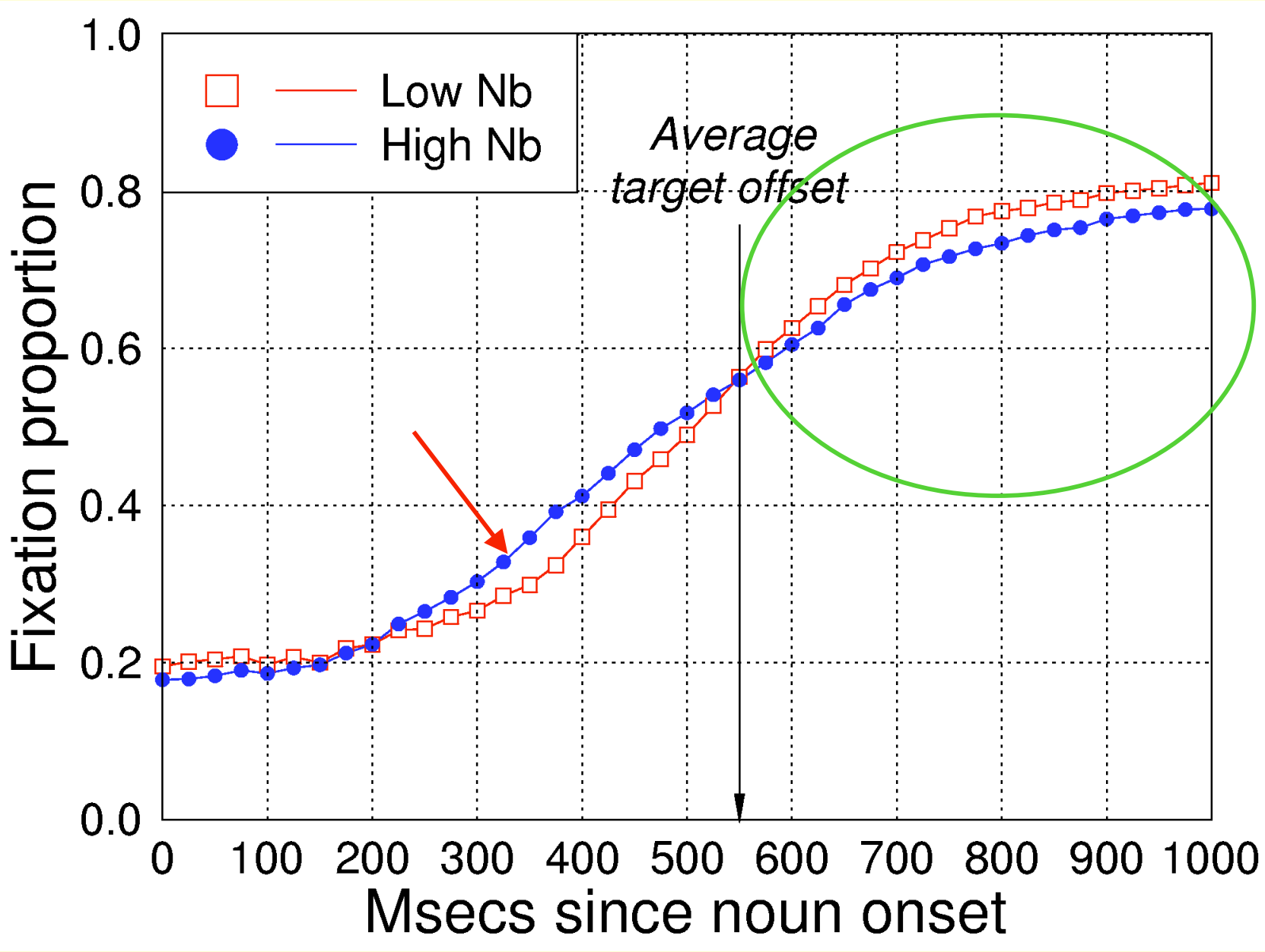




Cohort density (Magnuson et al., 2007)



Neighborhood density



Conclusions

We can study effects of entire lexicon, not just displayed words. (contra closed set claim)

Linking hypothesis issues need more work

Seeing neighborhood effects requires using targets and unrelated competitors*

***We often include one picture with name that has the same initial phoneme or just a one feature difference (not clear whether this is necessary).**

Artificial lexicon paradigm

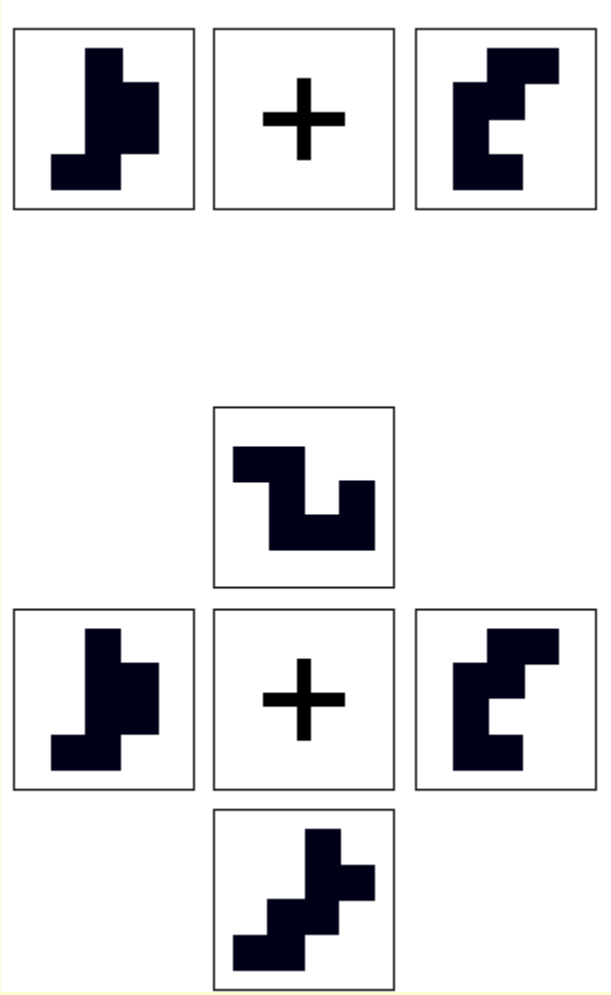
- Why use an artificial lexicon/language?
  - Control
    - Distributional properties of lexicon
    - Referential world/semantics
  - Learning and processing
- Are artlex words processing like real words?
- Is the art lexicon(lang) encapsulated?
  - is there a “bleeding” problem?

Artificial lexicon paradigm

Magnuson, Tanenhaus, Aslin & Dahan (2003) JEP:Gen

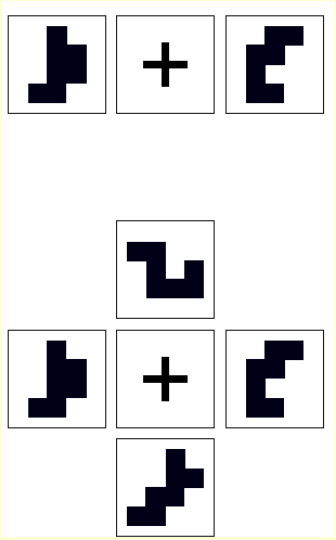
- **15 participants learned a 16-word lexicon**
- **Words refer to shapes →**
  - **Random mapping for each subject**
- **Four sets like:**

  pibo pibu dibo dibu
- **Allows high- and low-frequency (HF vs. LF) items with HF or LF neighbors**

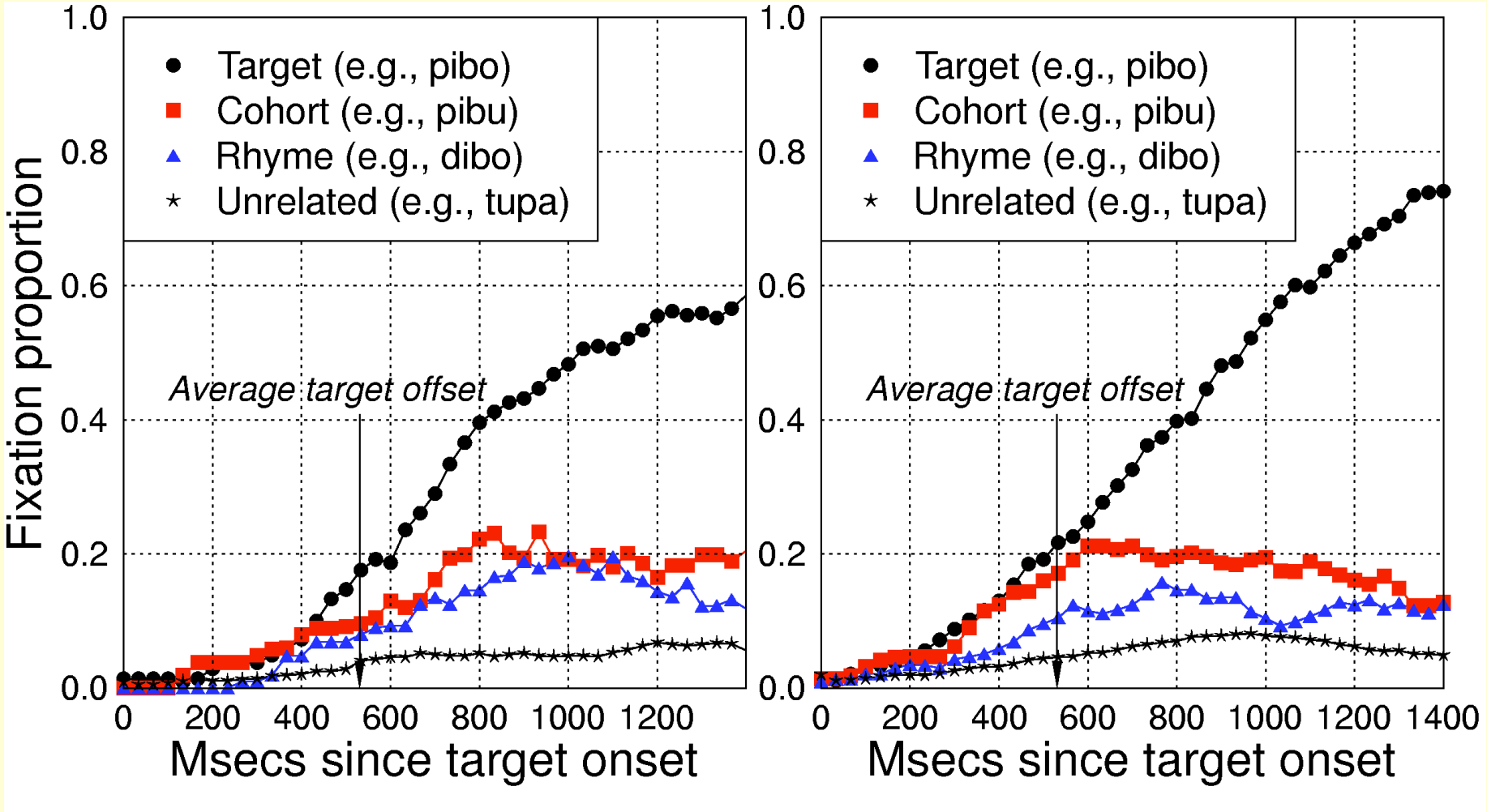


Procedure

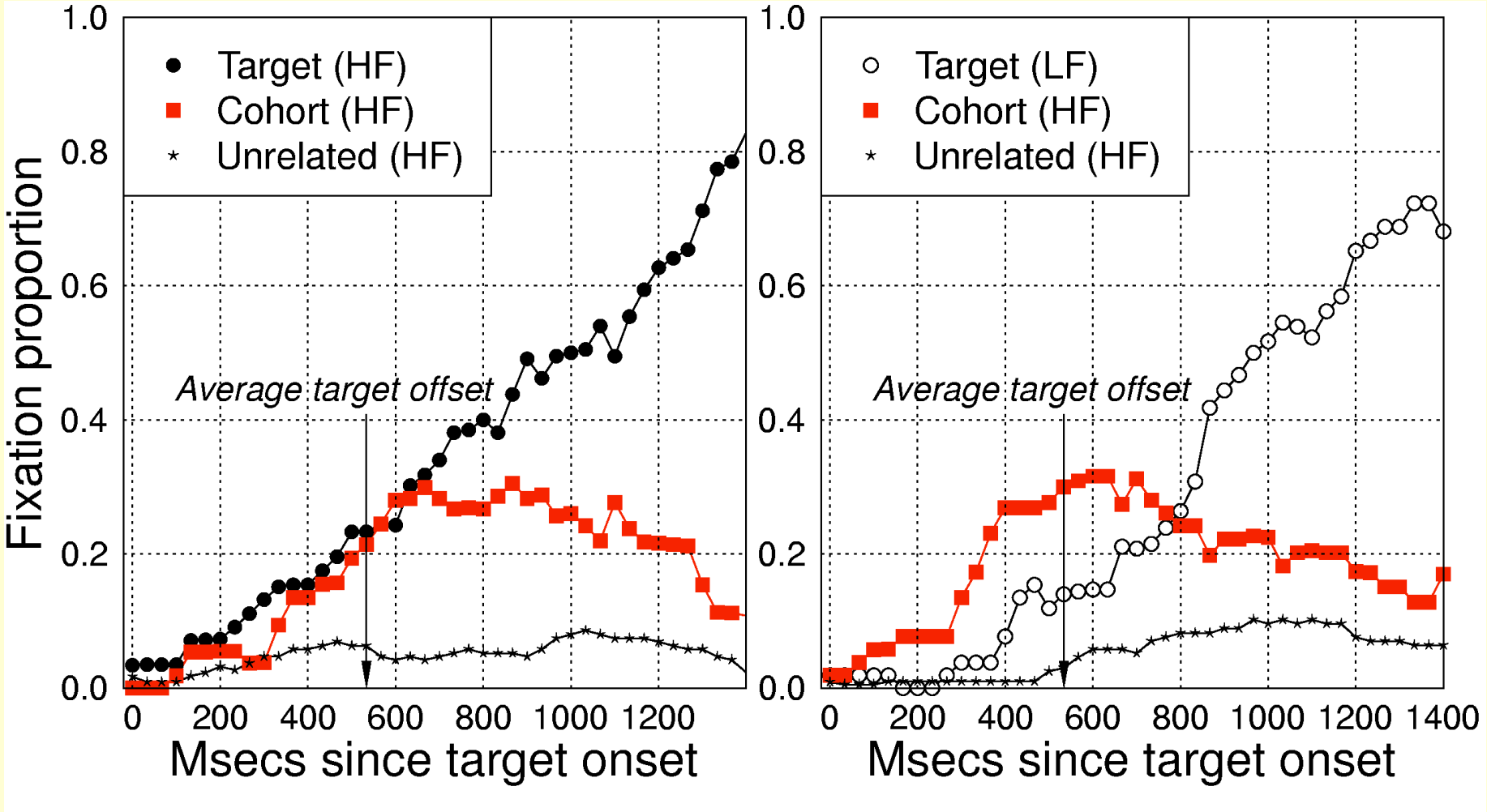
- Two days of training with a test at the end of each day
- HF items: 7 times / training block
- LF items: 1 time / training block
- Day 1:
  - 3 blocks of 2AFC, 4 blocks of 4AFC
  - Test (4AFC, eye tracking, no feedback)
- Day 2:
  - 7 blocks of 4AFC
  - Test



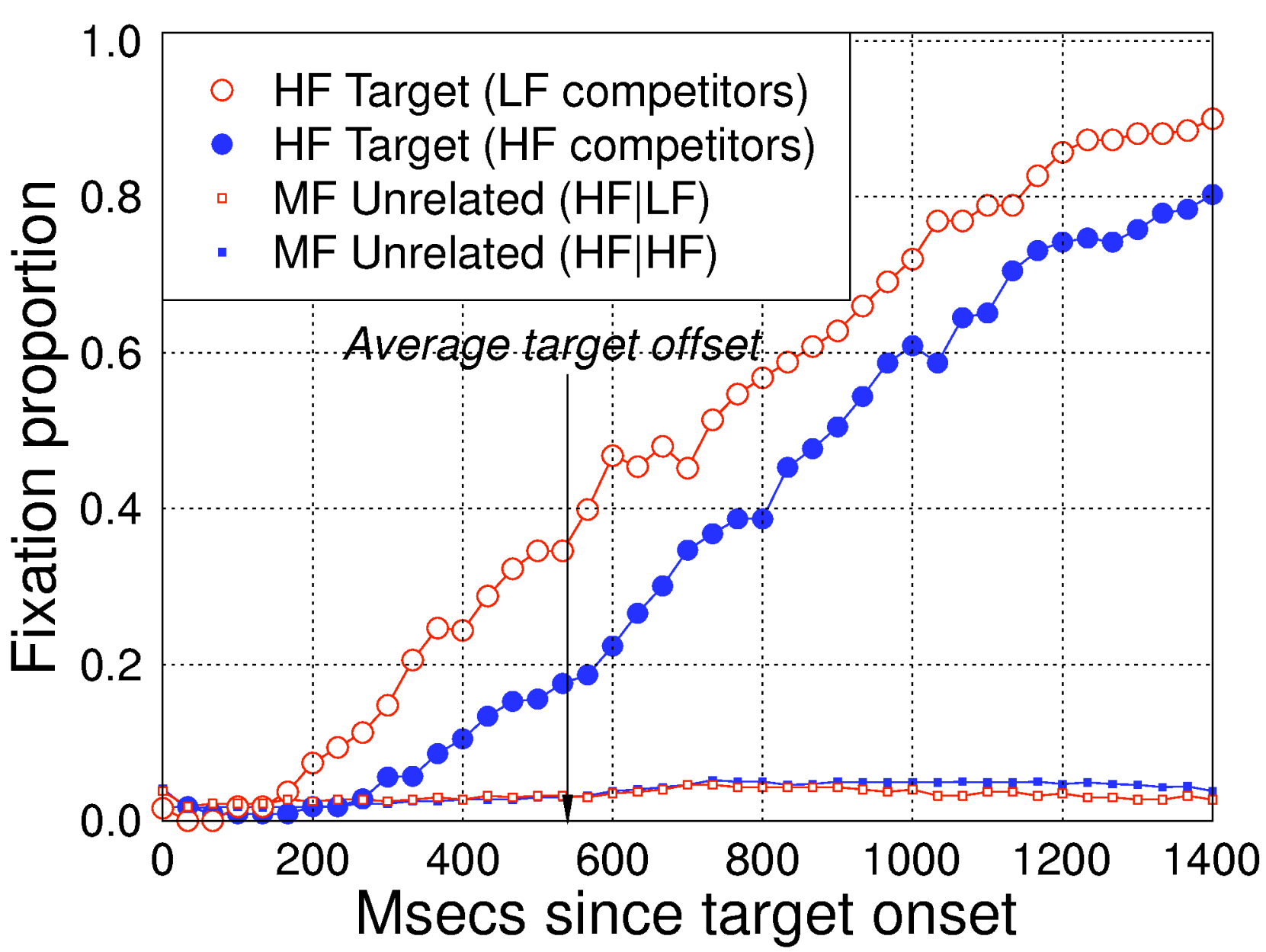
Replicated cohort and rhyme effects



Effects modulated by target and competitor frequency: high frequency targets



Absent neighbors compete



Summary

- After short training:
  - Replicated cohort and rhyme effects
  - Replicated frequency effects
    (Dahan, Magnuson & Tanenhaus, 2001)
  - Measured time course of interaction of item and neighbor frequency
- Effects not limited to displayed items

What about bleeding?

Magnuson et al. (03, JEP:General)

Created HF and LF novel words in:

(1) Dense and Sparse English neighborhoods

(2) Dense and Sparse ARTLEX neighborhoods

With the exception of small (ns) effects for low frequency words, only ARTLEX mattered; no bleeding density effects.

To first approximation, lexicon is encapsulated*

Useful tool (we'll see some examples)

Conclusions

Eye movements provide a sensitive measure of the time course of lexical activation that is sensitive to fine-grained phonetic/acoustic variation in the signal.

Linking hypothesis between fixations and lexical representations seems plausible.

Very early access to lexically-based perceptual/conceptual information.

Closed set does not induce strategies that bypass or mask effects of full lexicon

Combining with artificial lexicon allows for tight control

Fine-grained acoustic/phonetic detail

Cartoon of traditional view

Spoken word recognition is a problem of information reduction.

The signal is noisy and filled with irrelevant detail.

The listener must abstract away from meaningless variation to extract the relevant linguistic features

Categorical perception is the paradigmatic example:

good discrimination between categories
poor discrimination with a category
within category variability rapidly discarded

Reasons to expect use of sub-phonetic detail

Interactions between asynchronous acoustic cues

Voicing: VOT and vowel duration

Assimilation:

good run/m picks you up

Statistical learning

Adults and children are sensitive to distributional information

Perceptual learning/plasticity

adjustment to accents, cue-reweighting

Indexical effects

speaker specificity

Systematic variation in phonetic realization with prosodic domains

Articulatory strengthening of consonants, lengthening of vowels

cap/cap/captain example

Classic evidence for within-category sensitivity

Goodness judgment; RT differences

Lexical priming results (Blumstein & colleagues)

Evidence for use of fine-grained sub-phonetic variation in spoken word recognition

- VOT
  - Lexical access
    - (McMurray, Tanenhaus & Aslin, 2002; McMurray et al. in prep)
  - Recovery from phonetic/lexical garden-paths
    - (McMurray, Tanenhaus & Aslin, submitted)
  - Sensitivity to distributions in perceptual learning
    - (Clayards, Tanenhaus, Aslin & Jacobs, submitted)

Effects of within-category VOT on lexical access

McMurray, Tanenhaus & Aslin, 2002, *Cognition*

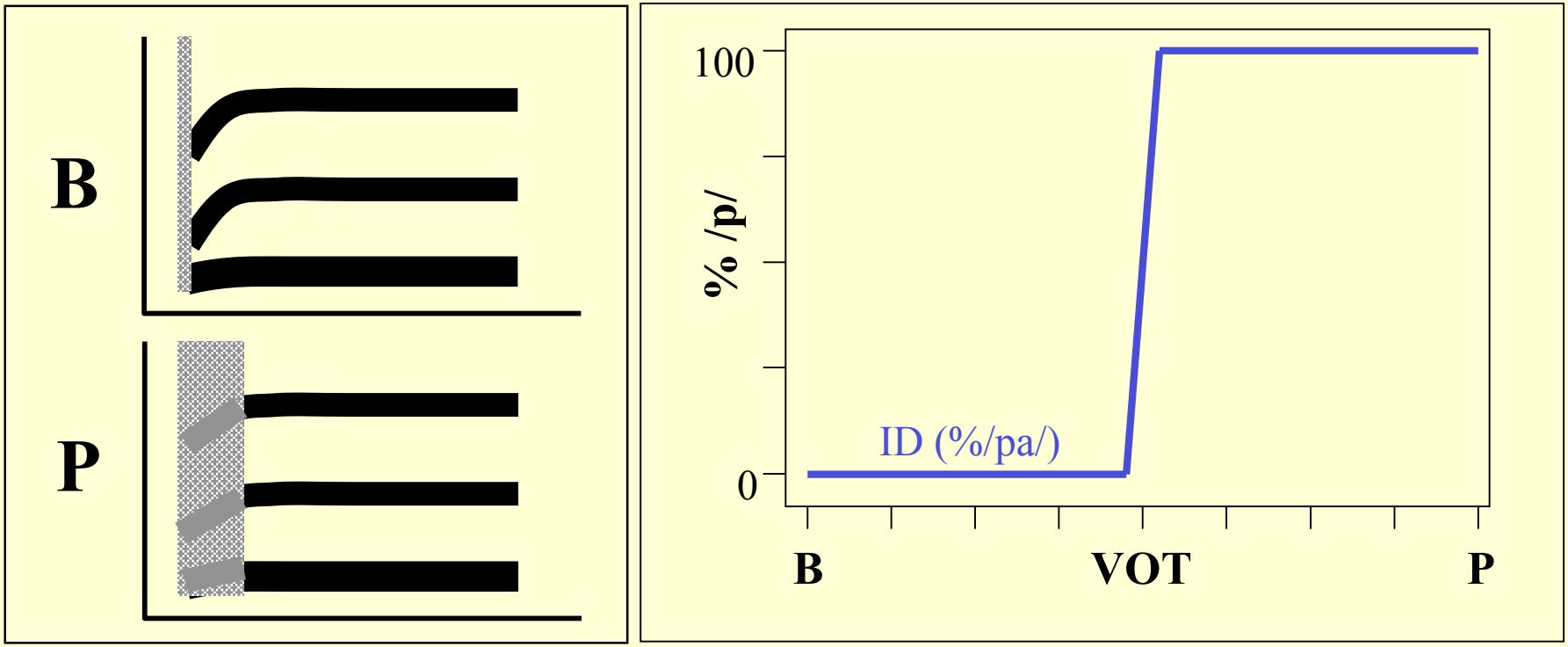
EyeLink 2 excellent all-purpose tracker:

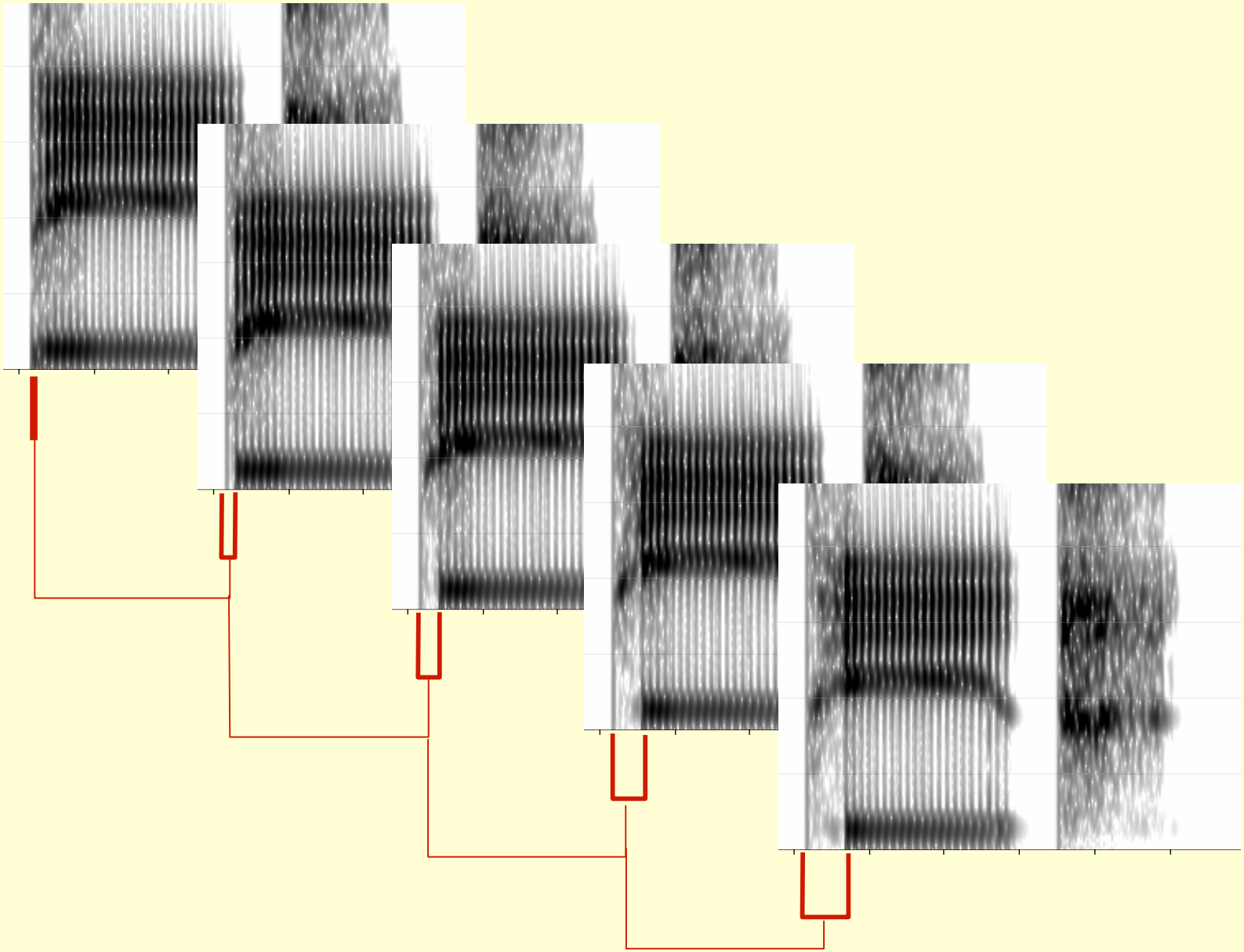
Reading*, visual world with screen, real objects**

***with chin rest**

****with scene camera attachment**

Categorical Perception





Use a speech continuum—more steps yields a better picture of acoustic mapping.

KlattWorks: generate synthetic continua from natural speech.

9-step VOT continua (0-40 ms)

6 pairs of words.

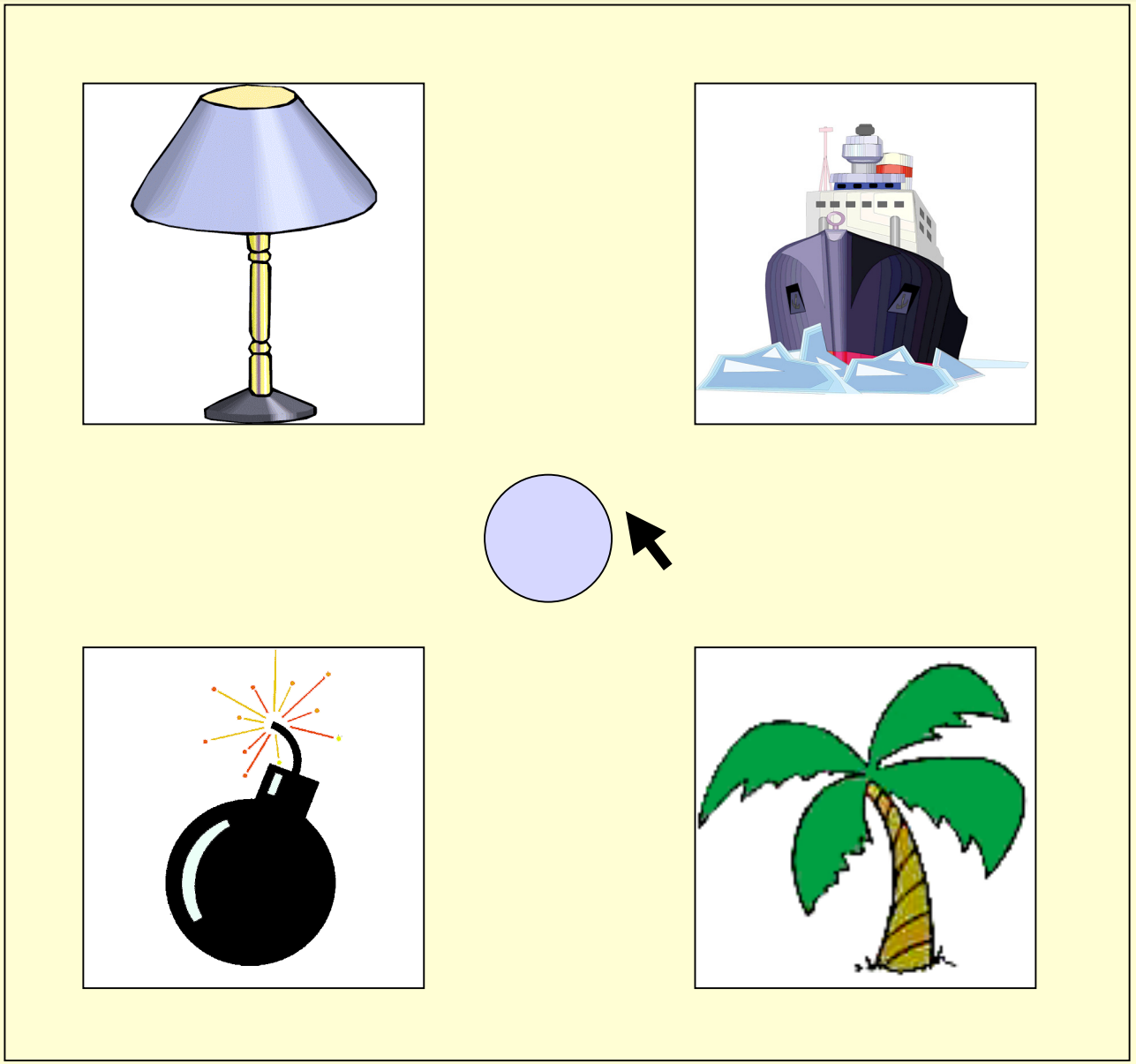
beach/peach	bale/pale	bear/pear
bump/pump	bomb/palm	butter/putter

6 fillers.

lamp	leg	lock	ladder	lip	leaf
shark	shell	shoe	ship	sheep	shirt

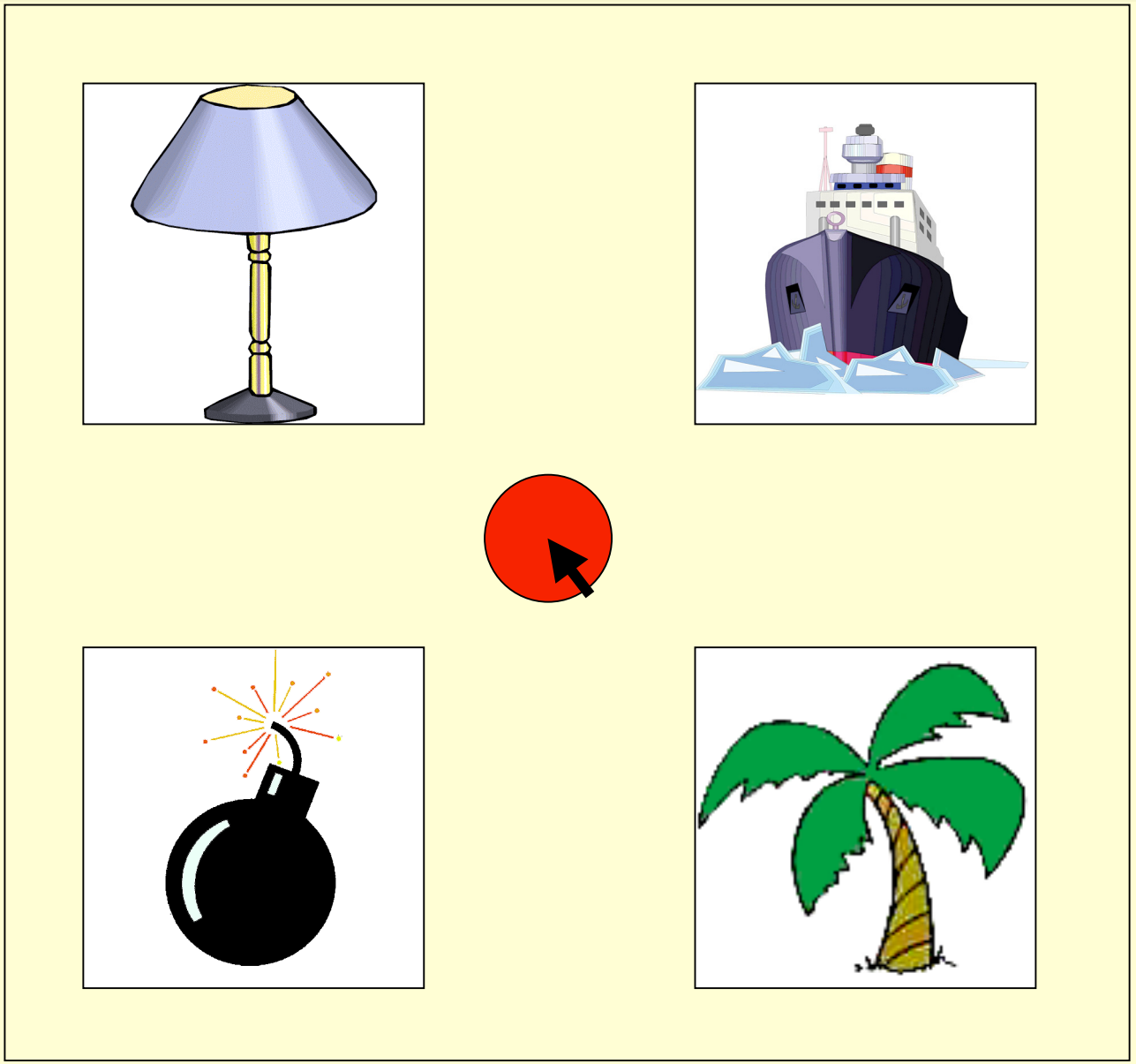
Lexical Identification

A moment to view the items

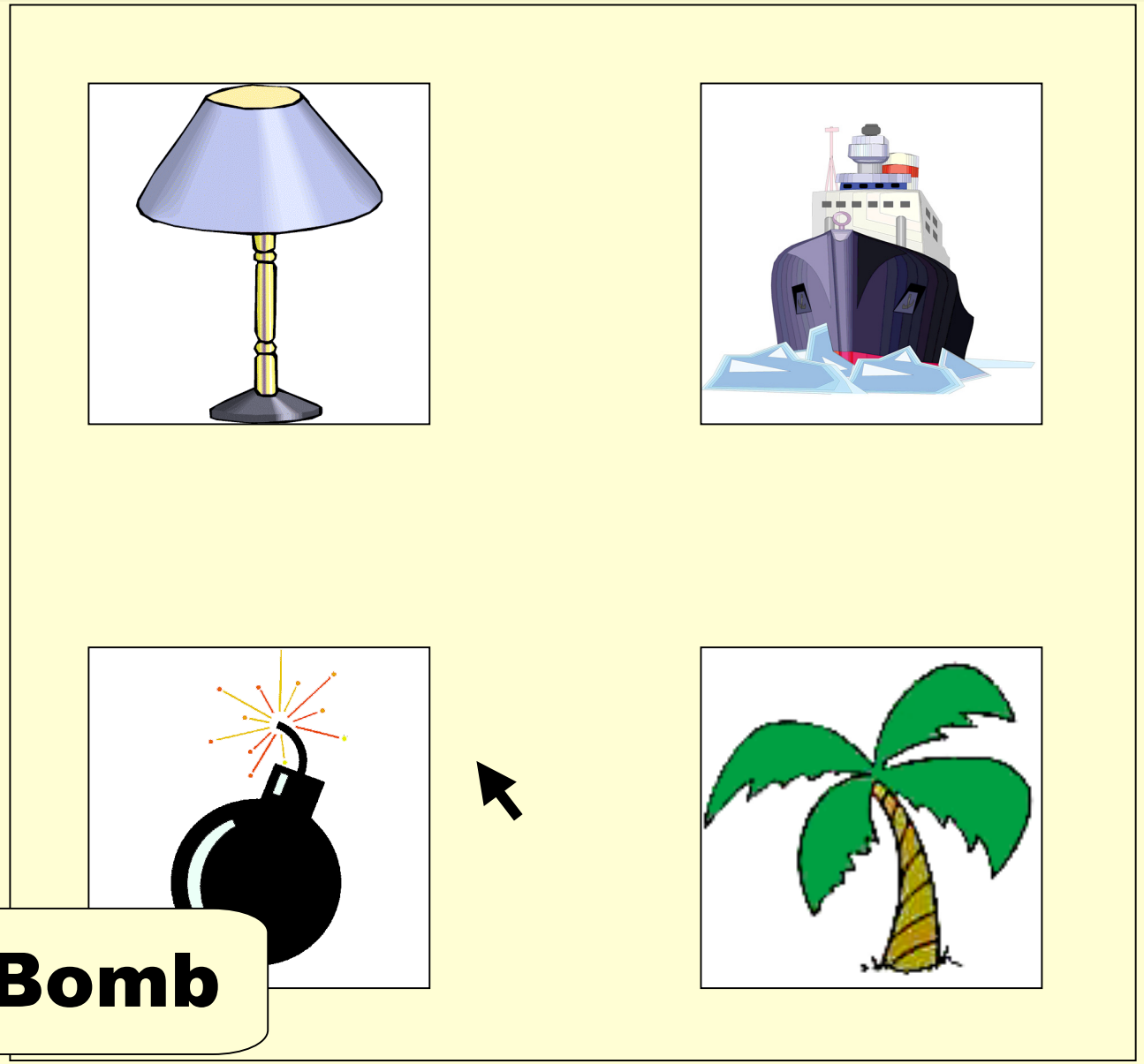


Lexical Identification

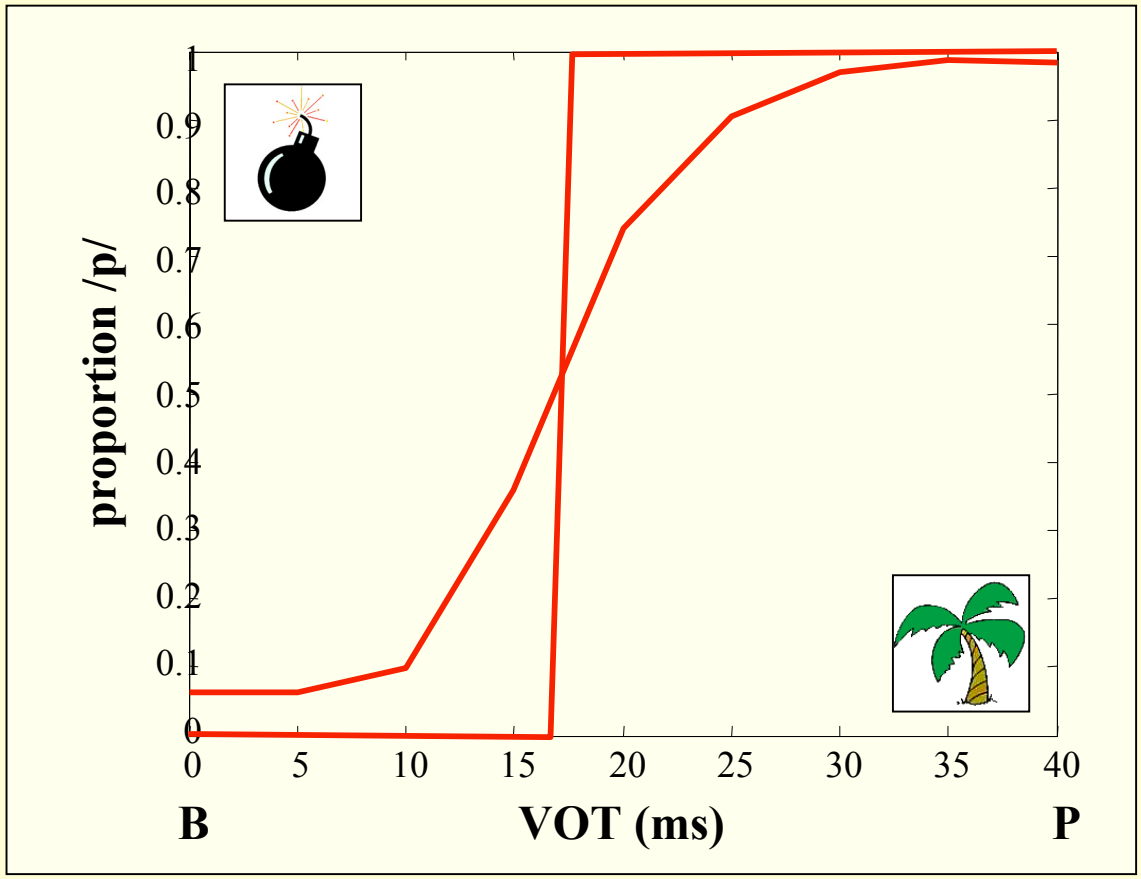
500 ms later



Lexical Identification



Identification Results

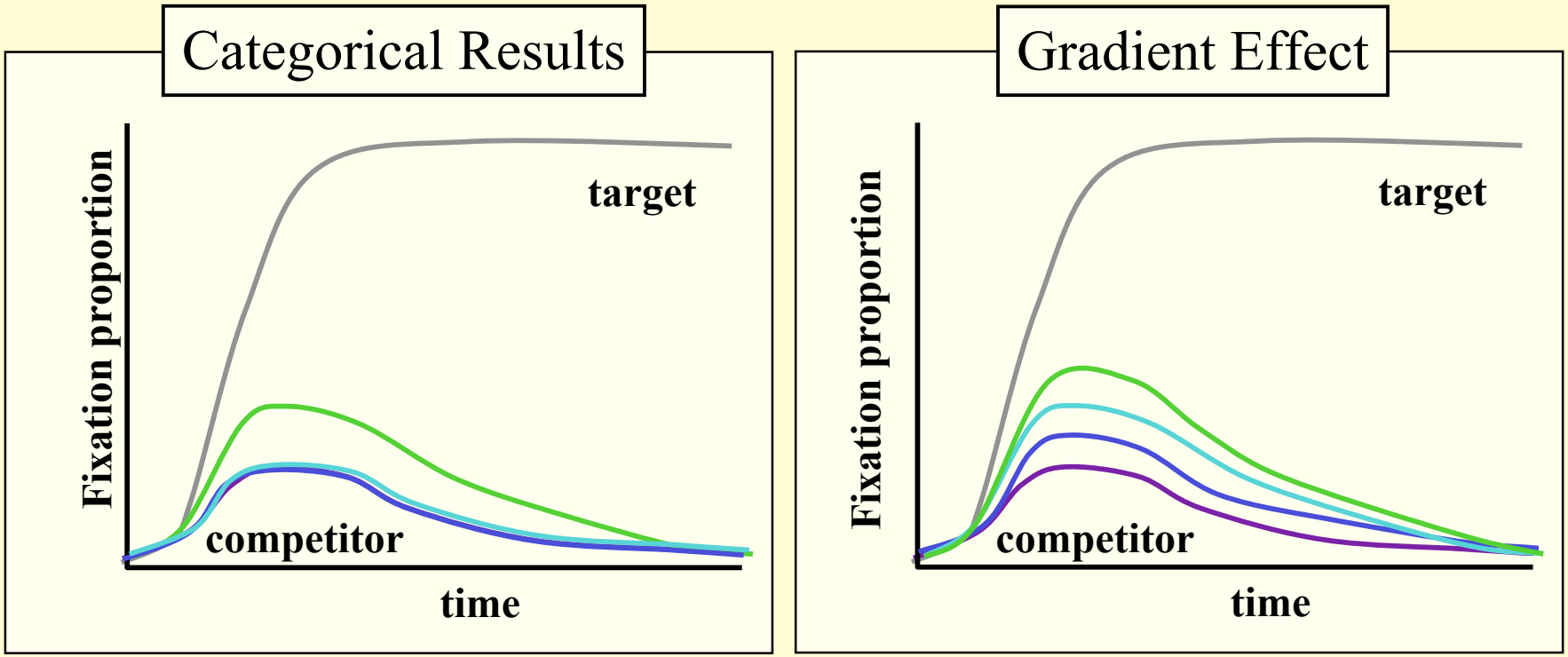


ID Function after filtering

Trials with low-frequency response excluded.

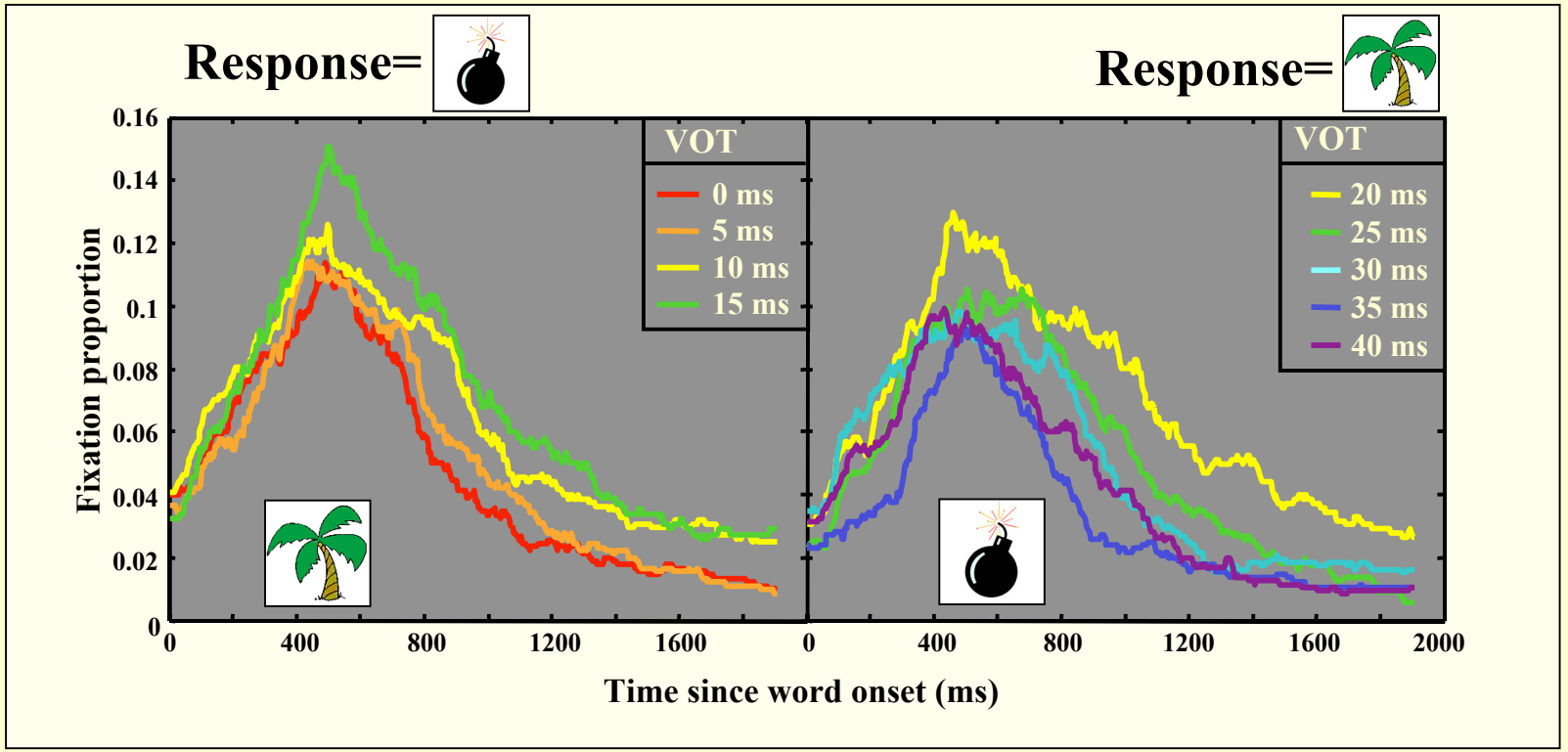
Yields a “perfect” categorization function.

Data Analysis (Response contingent analysis)

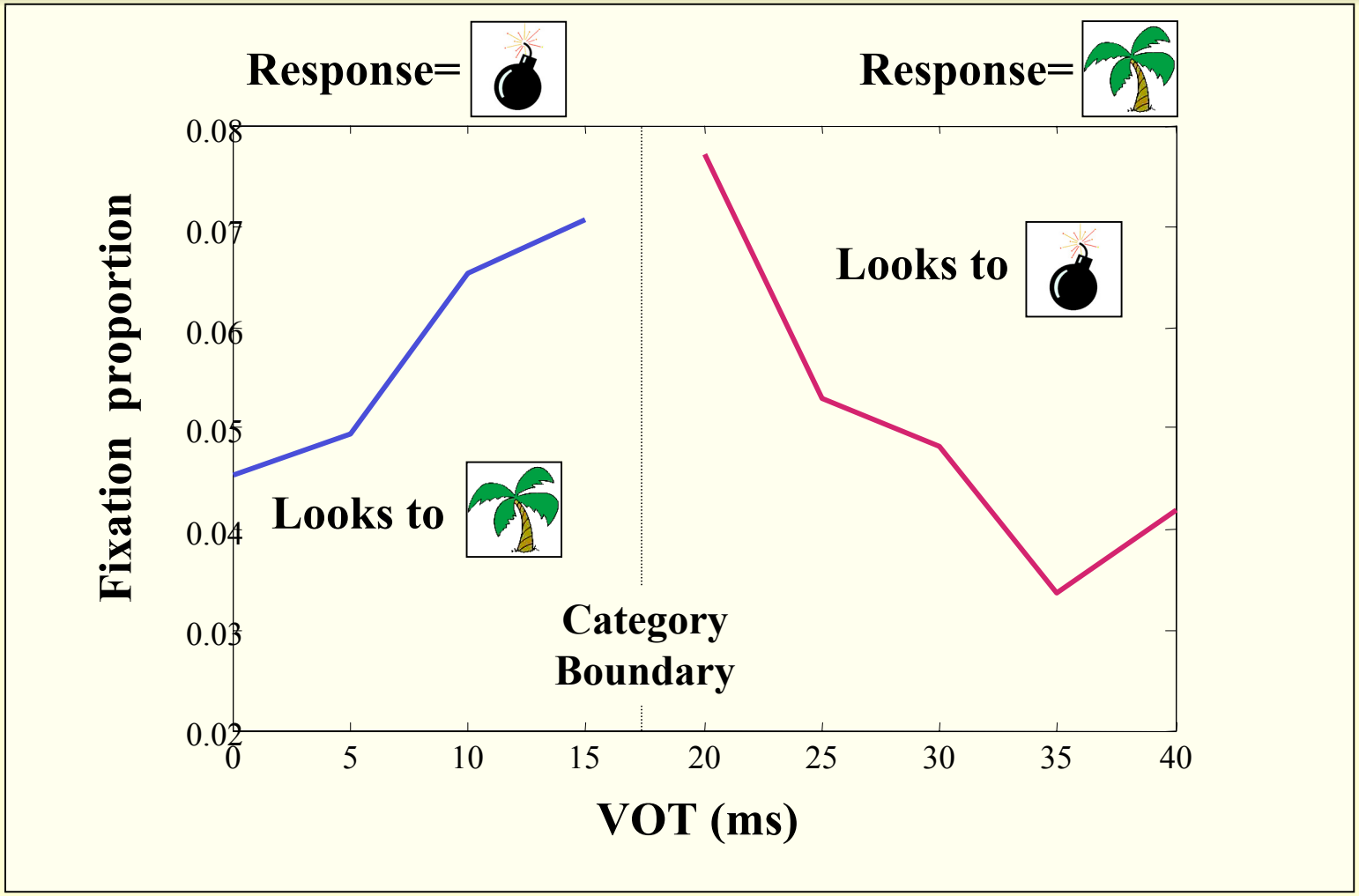


Eye Movement Results (response contingent)

Gradient effects of VOT?



Competitor looks as function of VOT



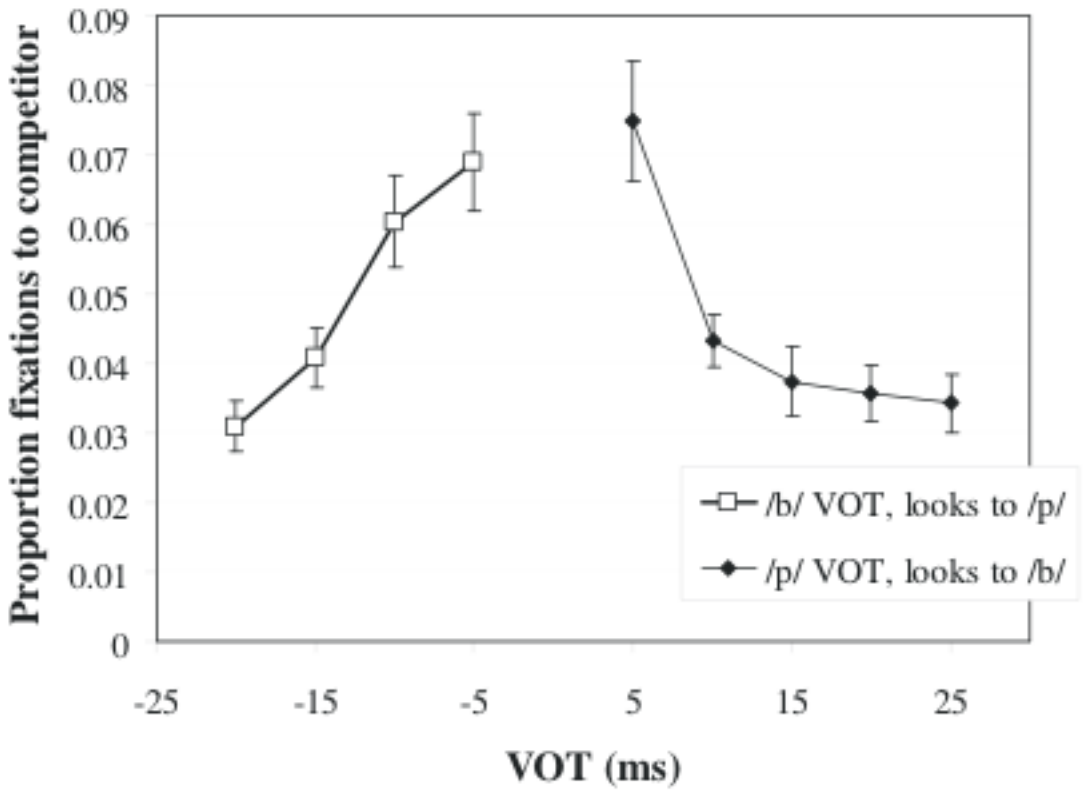


Figure 2: Proportion of fixations to the competitor as a function of VOT for voiced (left series) and voiceless (right series) tokens. Note that each curve reflects one 5-ms bin of rVOT (equivalent to grouping performed in the second ANOVA). Error bars reflect SEM.

Does sensitivity to detail persist over time?

10 Pairs of b/p items (*parakeet, barricade).*

- Differ in voicing of initial consonant
- Otherwise overlap for at least two syllables
- 0 – 45 ms VOT continua.
- Creates:
  - barakeet--> parakeet
  - barricade--> parricade continuum

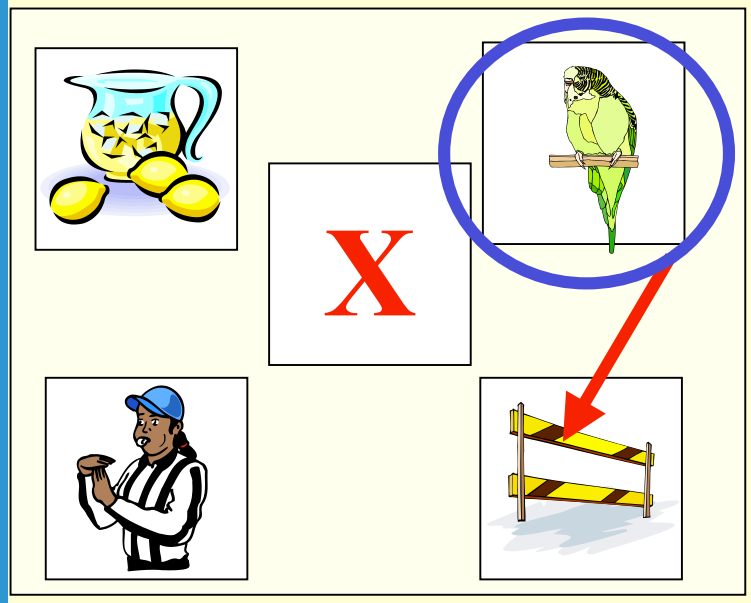
20 Filler items (*lemonade, restaurant, saxophone…)*

Option to click "X" (Mispronounced).

26 Subjects

1240 Trials over two days.

Look/response contingent



1. Looking at parakeet when the disambiguating information arrives (ade)

2. Next fixation is to barricade

3. Participant clicks on barricade

How long from POD to fixation shift?

Is it a function of VOT?

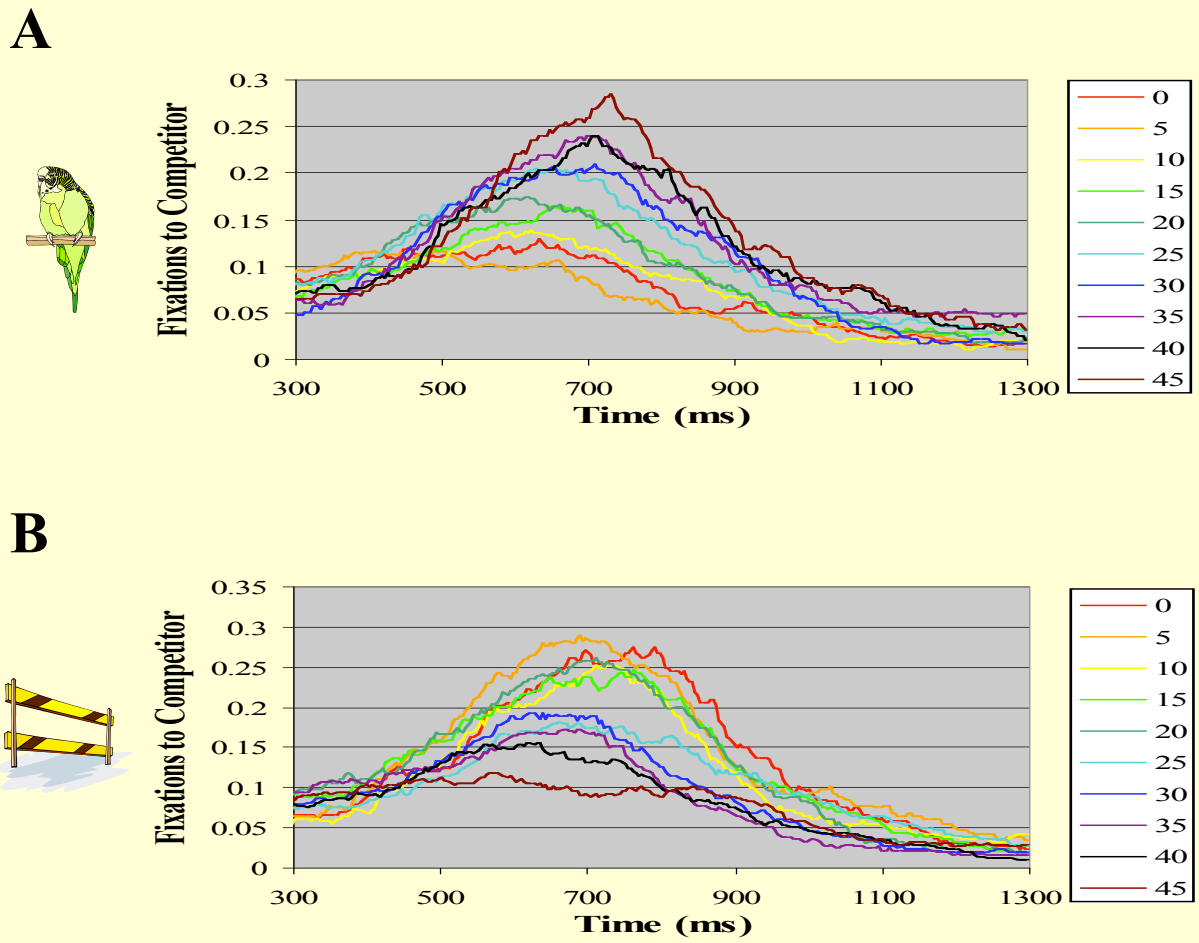


Figure 8: Proportion of fixations to the competitor as a function of time and VOT for voiced (Panel A) and voiceless (Panel B) targets.

Look-contingent analysis

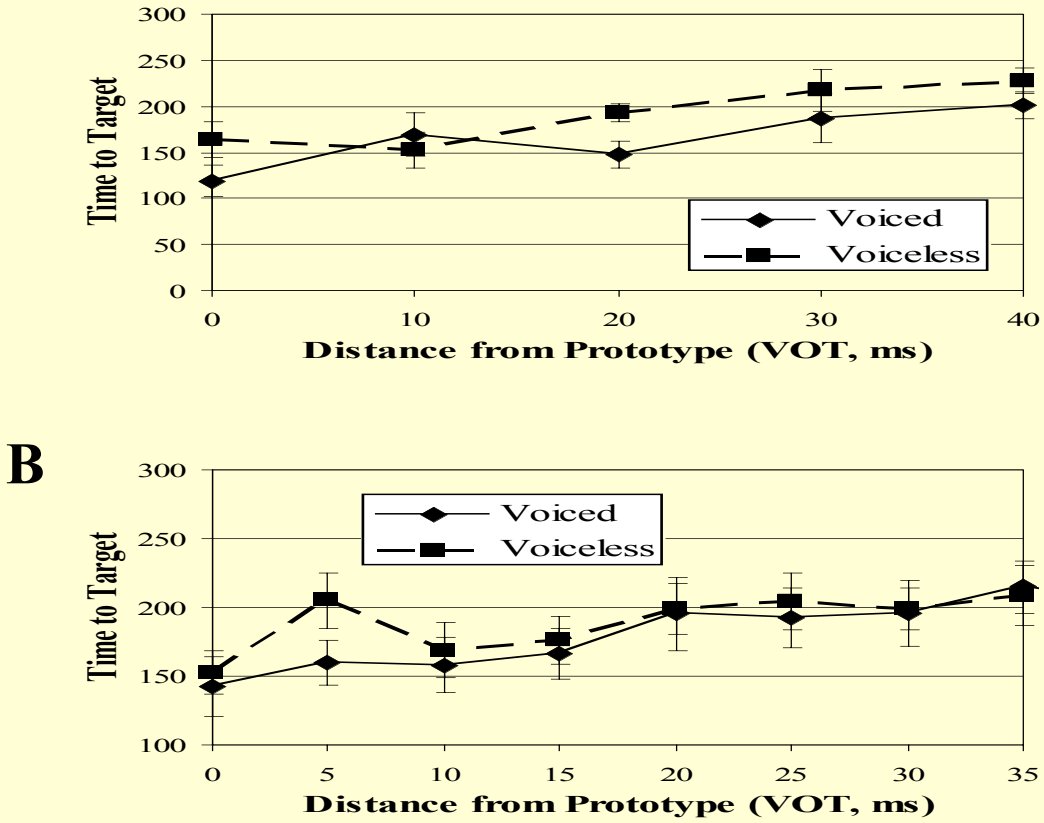
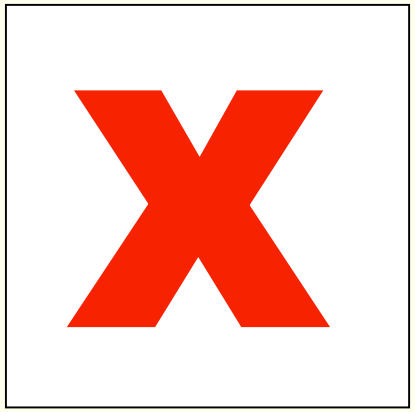
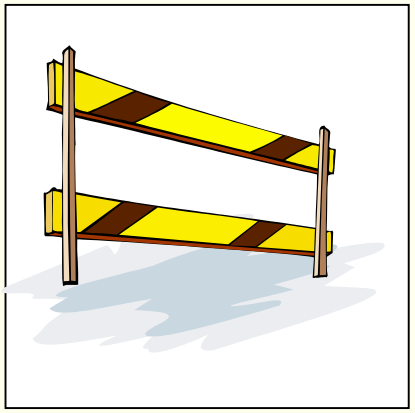


Figure 10: Time to fixate the target after the point of disambiguation as a function of distance from the prototype (in VOT) for voiced and voiceless targets. Panel A: time-to-target for trials in which subject was fixating the competitor just before the point of disambiguation. Panel B: time-to-target for trials in which the subject was fixating any non-target.

Same results with non-displayed competitor





Look-contingent analysis

A

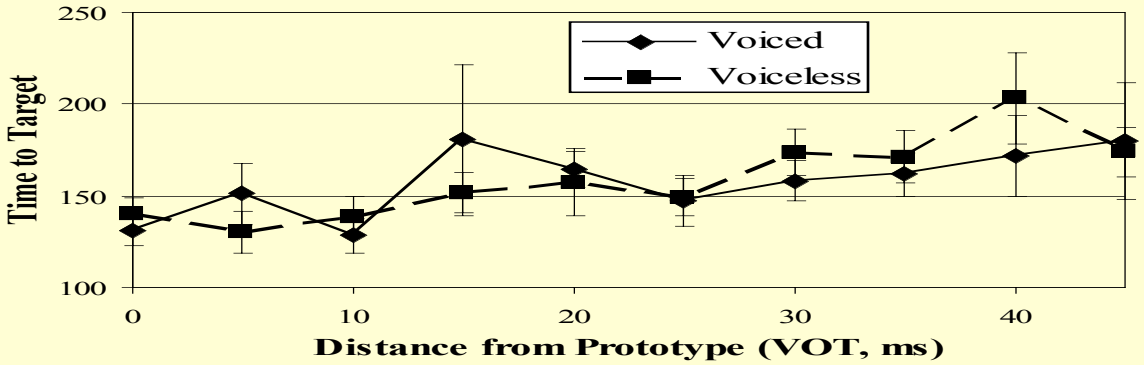


Figure 15: Time to fixate the target after the point of disambiguation as a function of distance from the prototype (in VOT) for voiced and voiceless targets. Time-to-target was computed for trials in which the subject was fixating any non-target object.

Gradient effect of within-category variation without minimal-pairs.

Gradient effect long-lasting: mean POD = 240 ms.

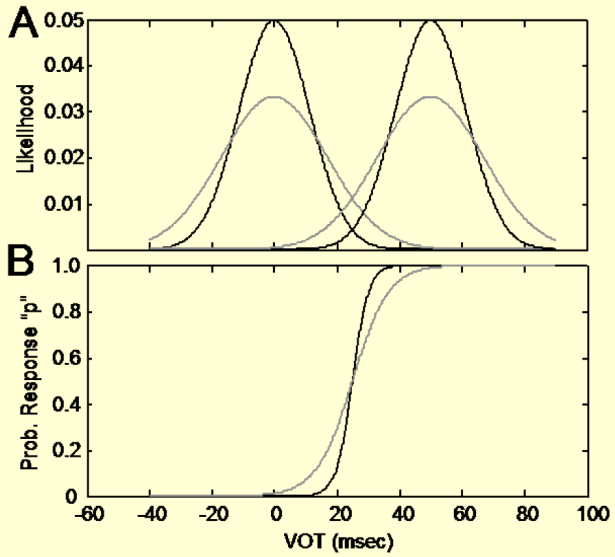
Regressive ambiguity resolution:

- Subphonetic detail still available when disambiguating information arrives.

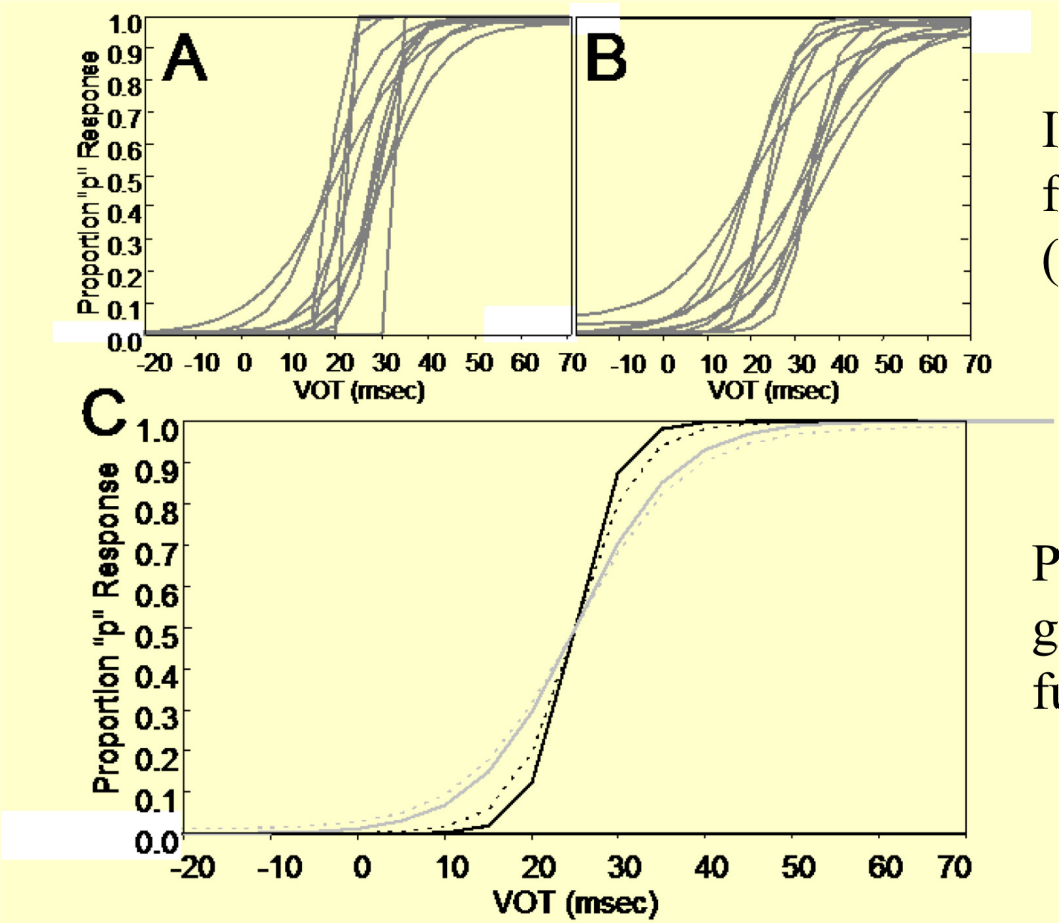
Perceptual Learning: sensitivity to distributions

If listener's are storing probabilistic detail like this, should be continuous sensitivity to distributions--continuous perceptual learning (Bayesian mechanism)

(Clayards, Tanenhaus, Aslin & Jacobs, submitted):

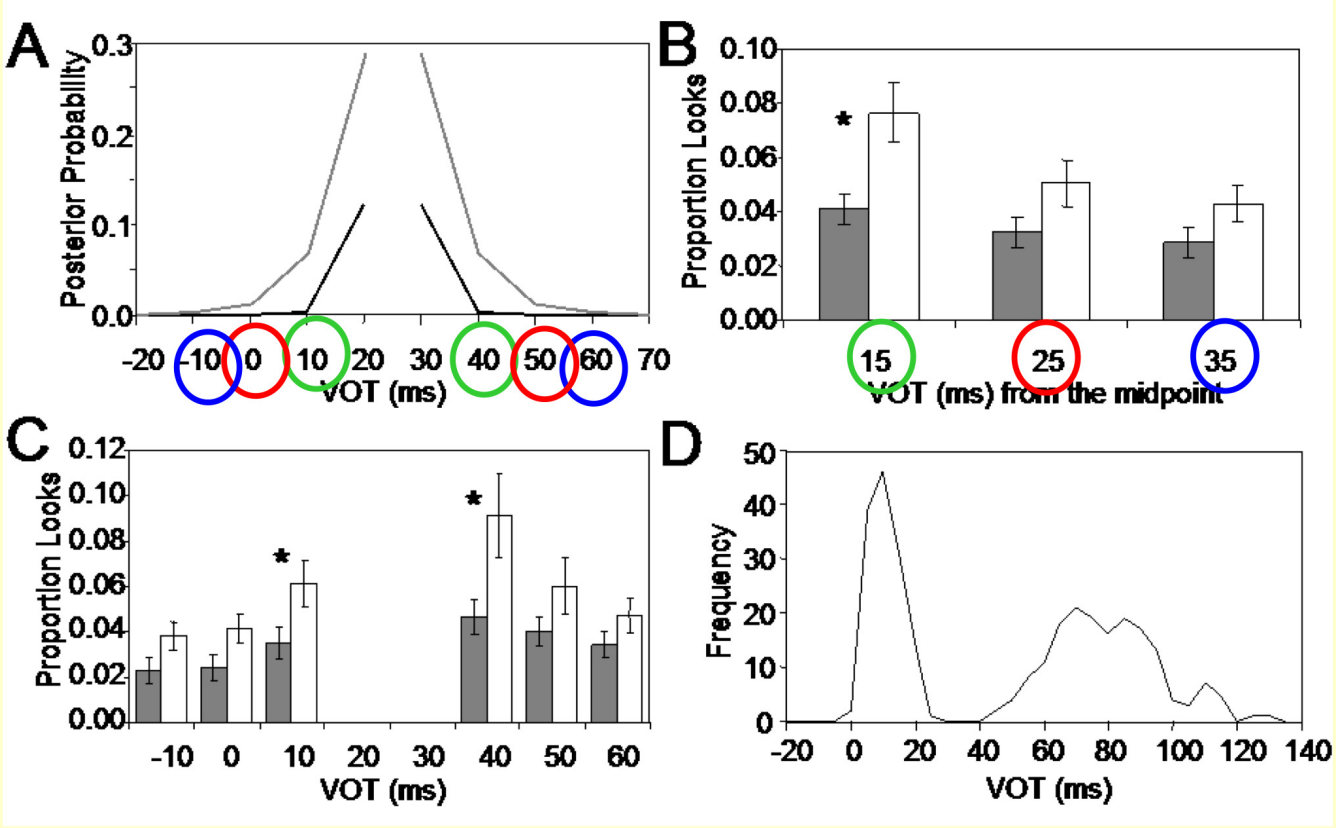


Exposed people to VOTs from wide or narrow distributions



Identification functions for (A) narrow and B (wide subjects

Predicted and actual group identification functions

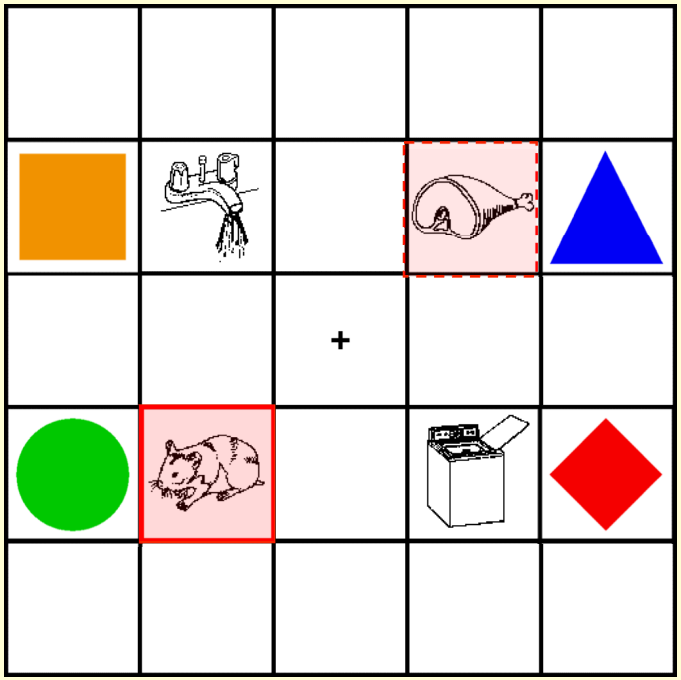


Bigger effect on “p”-side reflects effect of priors-- “p” s are more narrowly distributed than Ps in natural language stimuli.

Vowels

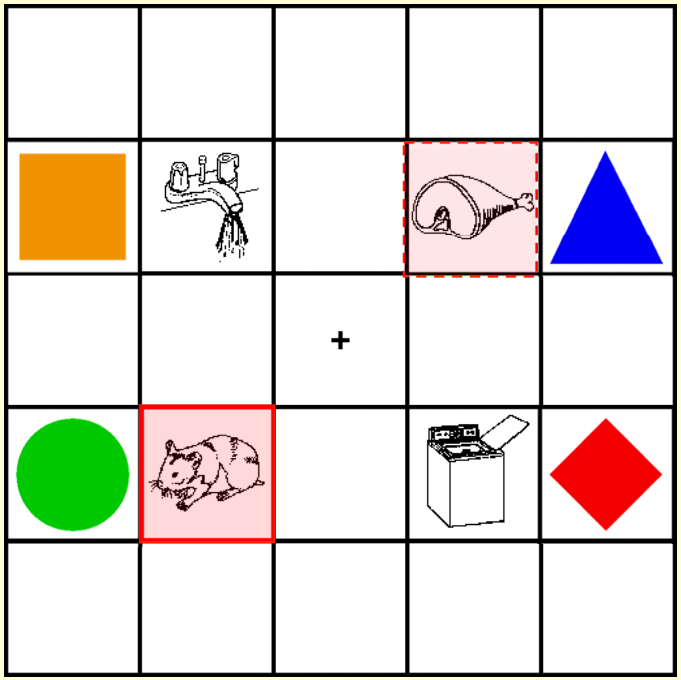
- Coarticulatory information (subcat mismatch)
- Vowel duration (ham/hamster; cap/captain)
  - Is there a ham in hamster?

Experiment 1.1 Salverda, Dahan & McQueen (2003, *Cognition*)



Ze dacht dat die hamster verdwenen was

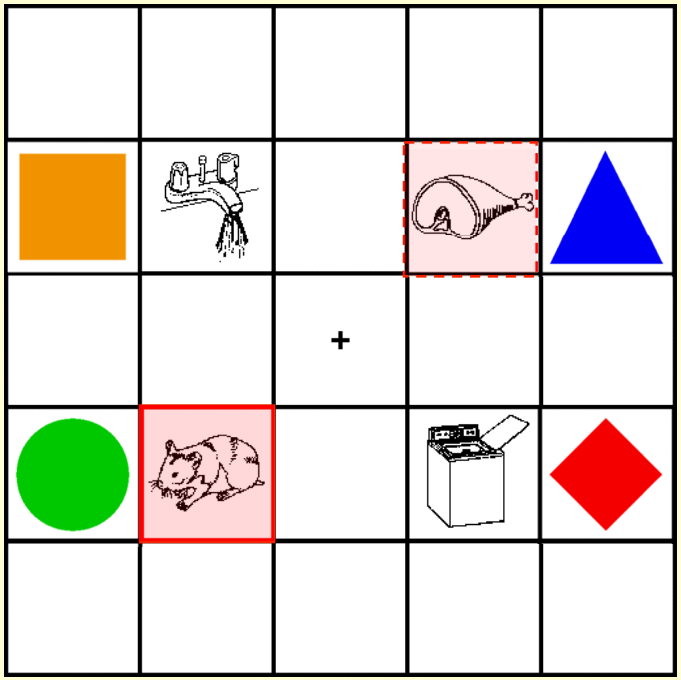
Experiment 1.1 Salverda, Dahan & McQueen (2003)



Ze dacht dat die hamster verdwenen was

Ze dacht dat die hamster verdwenen was

Experiment 1.1 Salverda, Dahan & McQueen (2003)

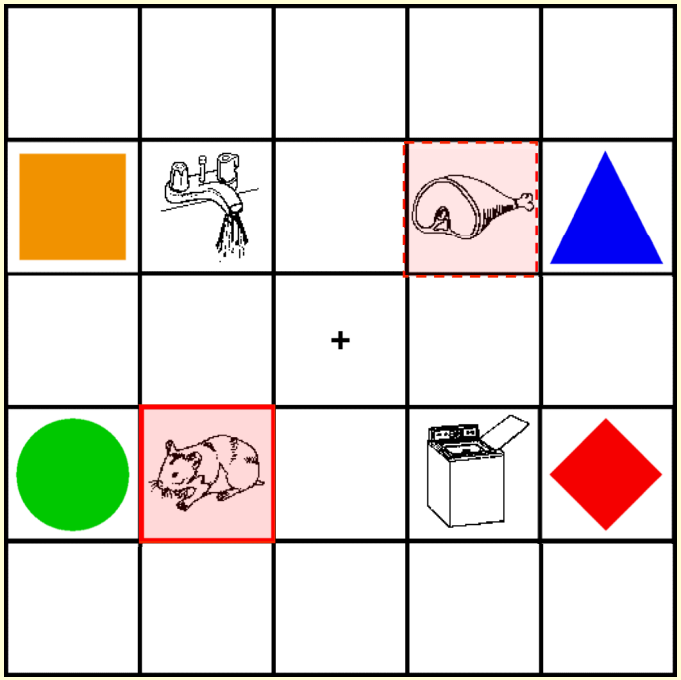


Ze dacht dat die hamster verdwenen was

Ze dacht dat die hamster verdwenen was

Ze dacht dat die ham stukgesneden was

Experiment 1.1 Salverda, Dahan & McQueen (2003)

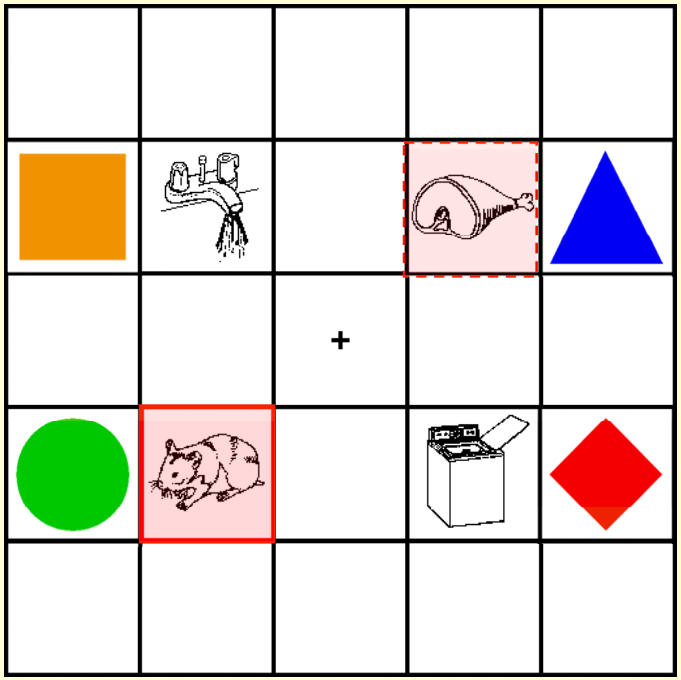


Ze dacht dat die hamster verdwenen was

Ze dacht dat die hamster verdwenen was

Ze dacht dat die ham stukgesneden was

Experiment 1.1 Salverda, Dahan & McQueen (2003)

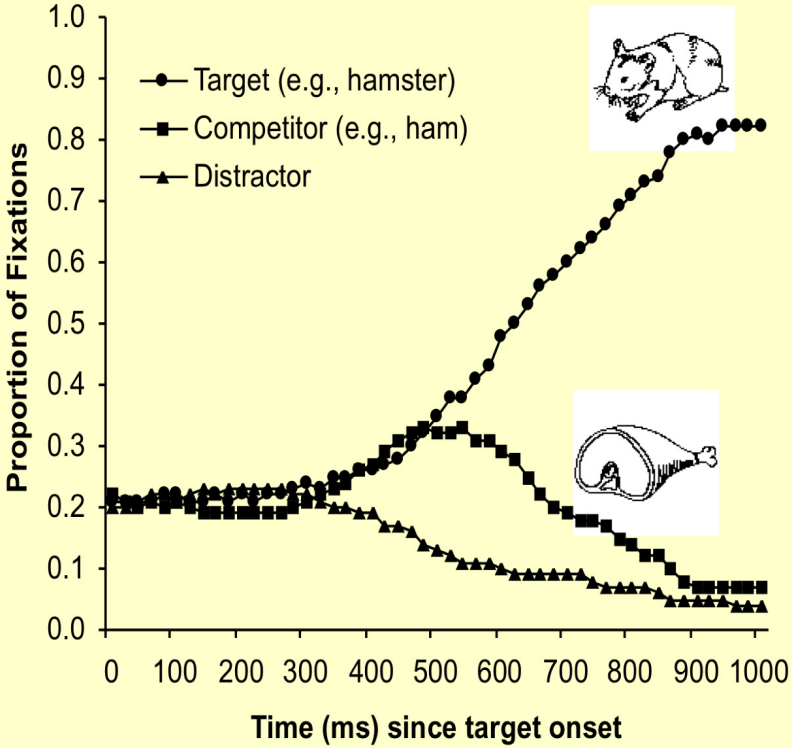


Ze dacht dat die hamster verdwenen was

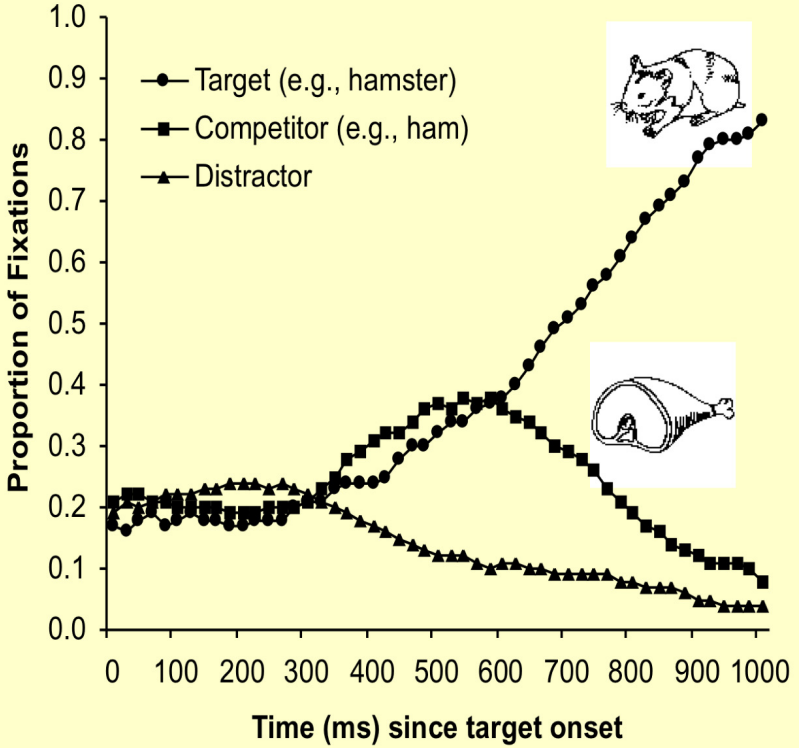
Ze dacht dat die ham ster verdwenen was

Ze dacht dat die ham stukgesneden was

Experiment 1.1 Salverda, Dahan & McQueen (2003)



Ze dacht dat die hamster verdwenen was



Ze dacht dat die ham ster verdwenen was

Conclusions Experiment 1.1

Listeners can distinguish a short word (e.g. ham) from the onset of a longer word (e.g. hamster)

Size of effect correlated with difference in duration between short word and first syllable of longer word

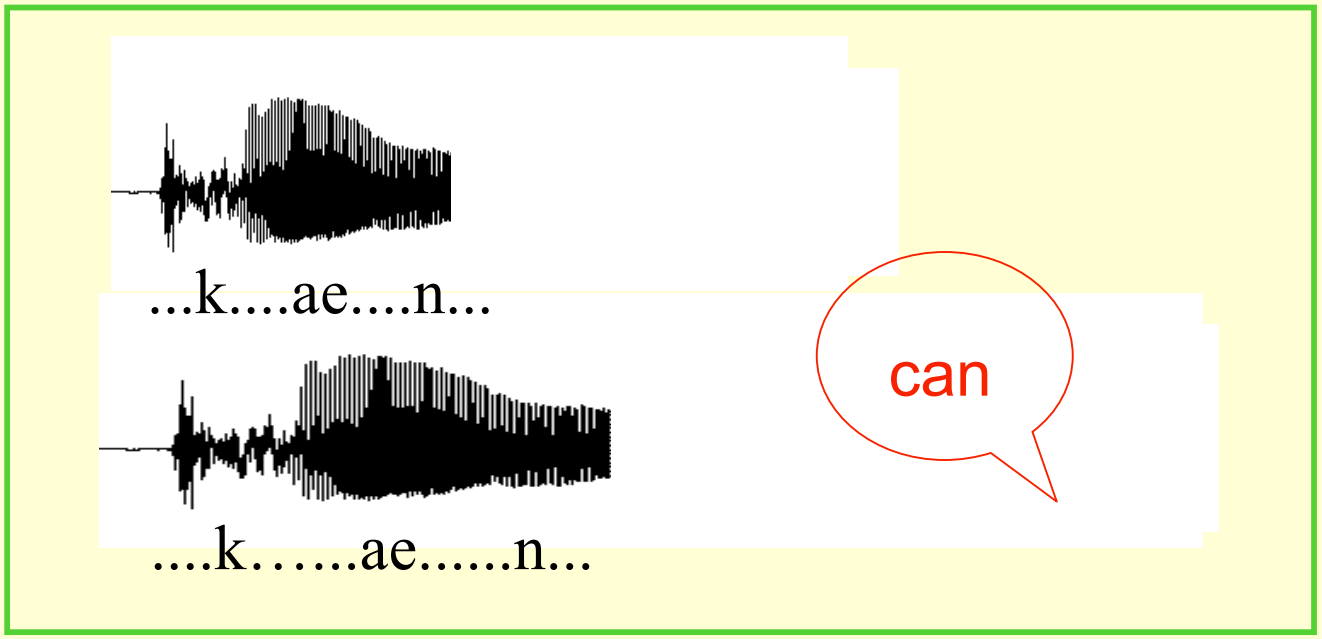
Initial segments of *ham* are affected more strongly by following prosodic boundary (pre-boundary lengthening) than the initial segments of the first syllable of *hamster*

Prosodically-conditioned detail affects lexical activation

Prosodic domains & lexical neighborhoods

- (Salverda et al., in press, *Cognition*)

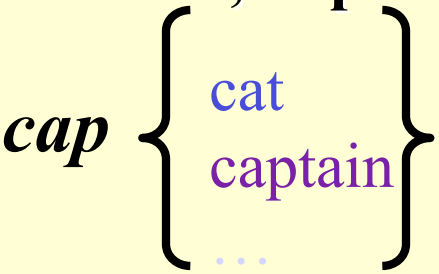
Word-lengthening



- Degree of lengthening depends upon strength of position in a prosodic domain

Does the composition of a lexical neighborhood change across prosodic domains?

(Salverda et al., in press, *Cognition*)

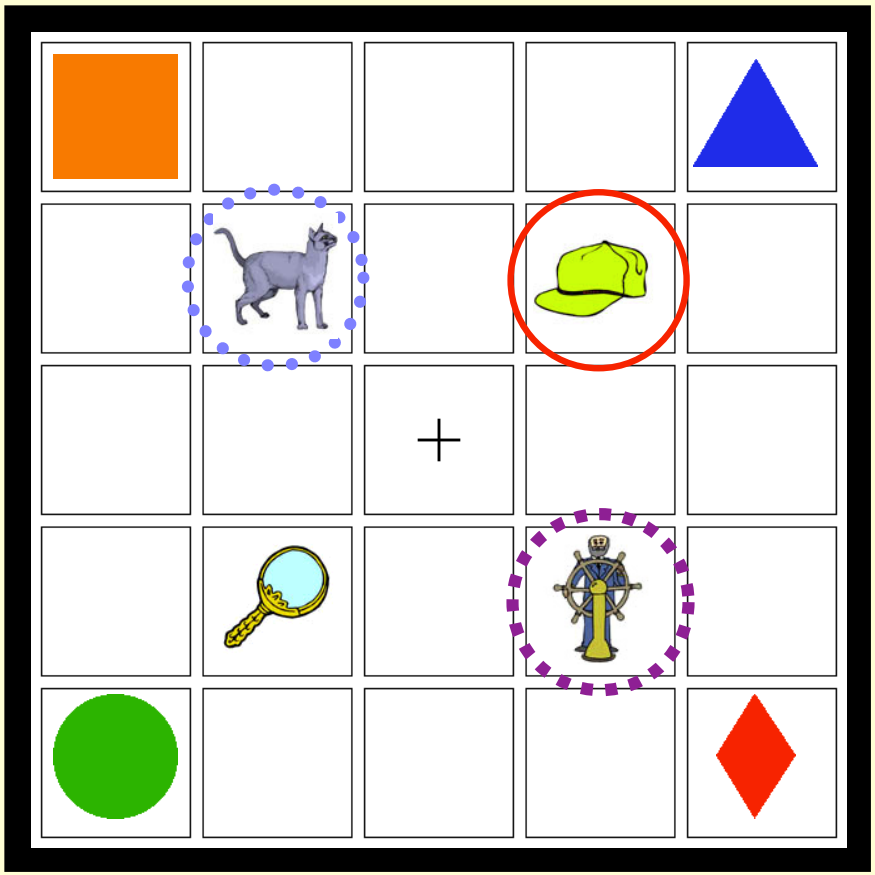


Articulatory strengthening and/or boundary tones **increase** vowel duration and **reduce** coarticulation, especially for monosyllabic words.

Cohort competition (e.g., cat) may be **stronger** in utterance final position compared to medial position:

Carrier word competition (e.g.,captain) may be **stronger** in utterance medial position compared to medial position:

Triples Experiment:

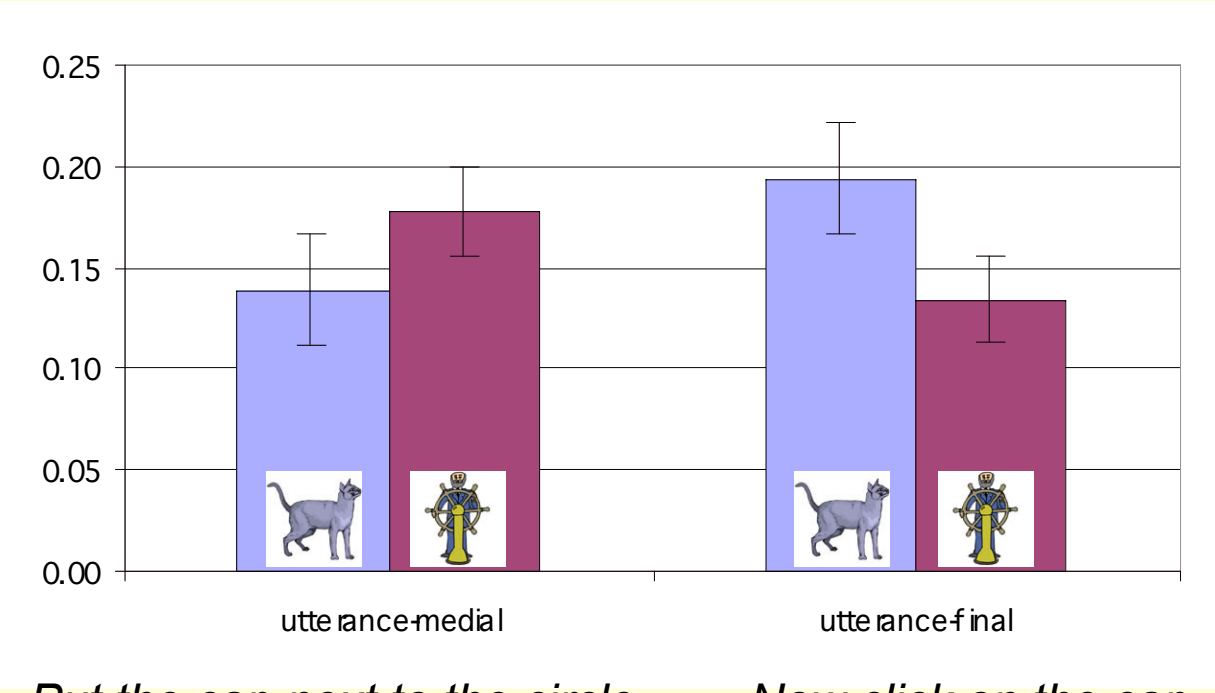


Put the cap next to the circle

or

Now click on the cap

Triples Experiment
Fixation proportions [300,900]



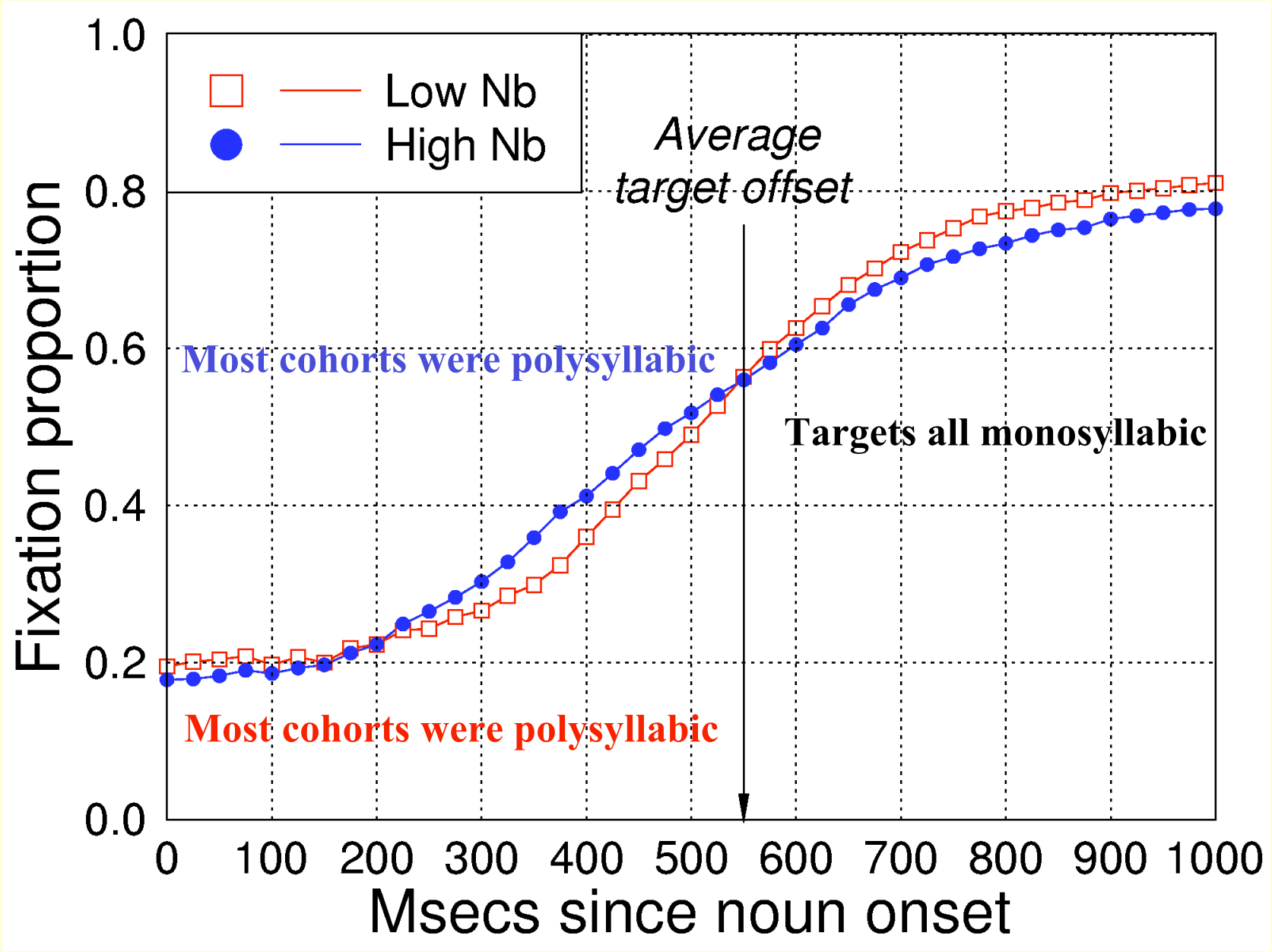
Put the cap next to the circle

Now click on the cap

Conclusions: Triples

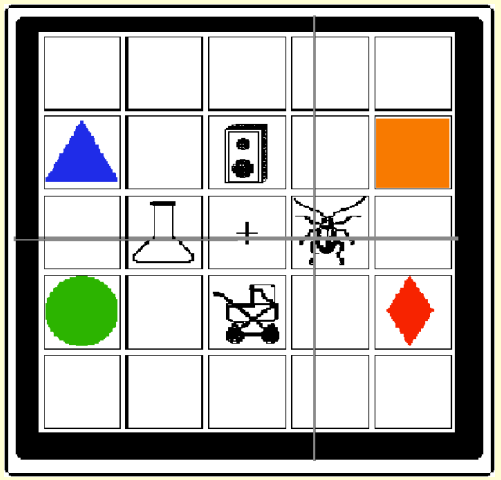
- The match between a lexical candidate and a spoken word is contingent on its prosodic context
- The prosodic context of a word has an impact on the relative activation of different *types* of competitor words.
- Studying neighborhood effects for words in isolation can be misleading (effects will generalize best for strong positions in a prosodic domain)

Neighborhood density: paradox resolved



Examples of using cohort manipulations as tools

Pitch accents: Can use cohorts to isolate effects from the vowel (Dahan, Tanenhaus & Chambers, JML 2002)



1. Put the beetle below the square.

   Now put the BEAKER/beaker….

2. Click on the beaker and the carriage

   Now put the BEAKER (H*/L+H*) below the triangle.

Other uses:

3. Speaker specific information (e.g., prior exposure to beaker and beetle in same of different voices (Creel, Aslin & Tanenhaus, in press, *Cognition*)

4. Context and lexical access (Dahan & Tanenhaus, 04, JEP:LMC) uses cohorts and cross-spliced misleading coarticulatory information.

5. McMurray and Gow (coronoal assimilation, LabPhon in press)

Examples of using competitor artificial lexicon language manipulations

Lexical learning

Creel et al., Shatzman & McQueen (lexical learning and representation)

Context effects (create well-defined contingencies, well-controlled neighborhoods)

(Pirog, Tanenhaus & Aslin, ms; Magnuson, Tanenhaus & Aslin, ms)

Does word recognition automatically activate perceptual/motor areas of cortex?
(Pirog, Aslin & Tanenhaus, ms)

Experiment Details

- Lexicon
  - 8 CVCV items refer to novel shapes
  - 8 CVCVCV items refer to changes that happen to the shapes
    - 4 motion (horizontal oscillation, vertical oscillation, contract/expand, rotate)
    - 4 surface (black, white, marble, speckle)
  - CVCVCV items form 4 cohort pairs
    - Motion-motion, surface-surface, motion-surface (x2)

Experiment Details

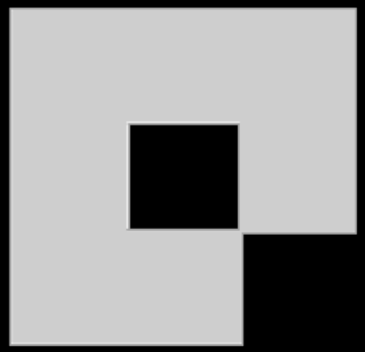
- Training
  - 1 hr/day for 3 days prior to scan
    - "show" task: participant sees shape + change, hears name of shape and change
    - "tell" task: participant hears name of shape and change, sees shape and change, must indicate match / no match (shown correct pairing as feedback)
  - All participants at ceiling levels of accuracy (> 98%) on day 3 (inside mock magnet)

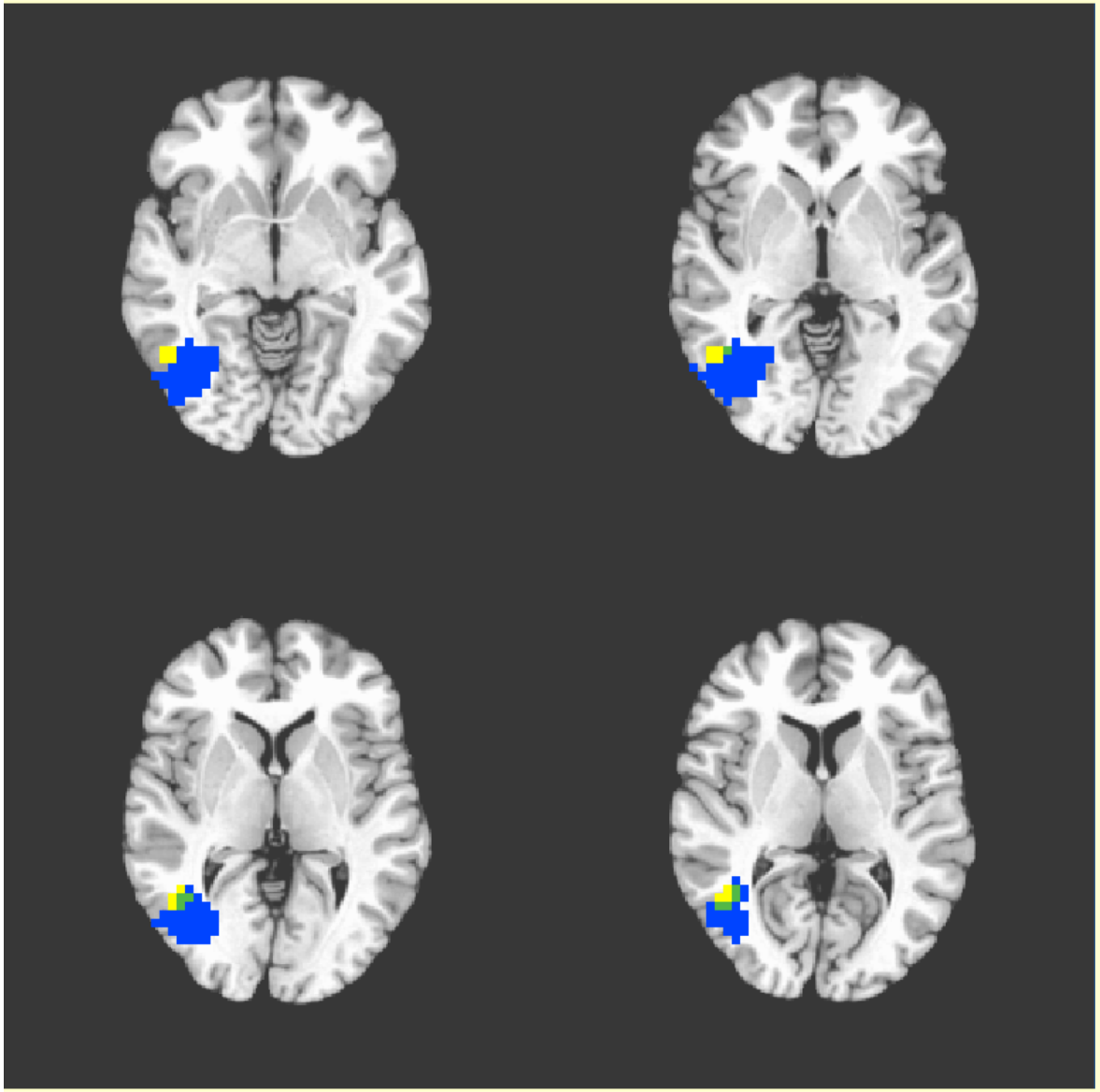
Experiment Details

- Test Session
  - Modified version of “tell” task
    - Computer-paced
    - Reduced feedback
    - Participants still highly accurate (95% correct)
  - MT localizer scan
  - Structural scan

start auditory visual response interval + feedback

3 – 7 s 2 – 4 s 3 s 1 – 9 s





- Blue: more active for motion events than texture events
- Yellow: more active for motion words than texture words
- Green: area of overlap
- Competitor modulates degree of activation
  - MM>MN>NM>NN

Conclusions: General

- Eye movements, coupled with contingent analyses, are a powerful tool for studying time-course of speech/lexical processing
- The system is exquisitely sensitive to fine-grained variation in the signal (embarrassment of riches)
- Variation within phonetic categories/features is signal not noise
  - Promising for a Bayesian approach to asynchronous cue integration--like those used in vision
- Cohort manipulations useful for studying pitch accents
- Artificial lexicons/languages in visual world a useful tool for range of questions