Prosody and Visual-World Eye Tracking

Annie Tremblay
University of Kansas

Kiwako Ito
Ohio State University

Aix Summer School on Prosody

September 9, 2016

Some History

- Cooper (1974)
  - Listeners looked at visual displays while listening to short narratives
  - Listeners were told they could look anywhere on the screen, that only their pupil size was recorded
  - Listeners automatically fixated words that had been mentioned in the speech signal, with fixations being closely time-locked to speech
- It took another 20 years before this method began to be used in psycholinguistic research **(Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995)**

Visual-World Eye-Tracking

- Participants' eye movements to objects in a visual display or printed words on the screen are recorded as participants listen to speech

Apparatus

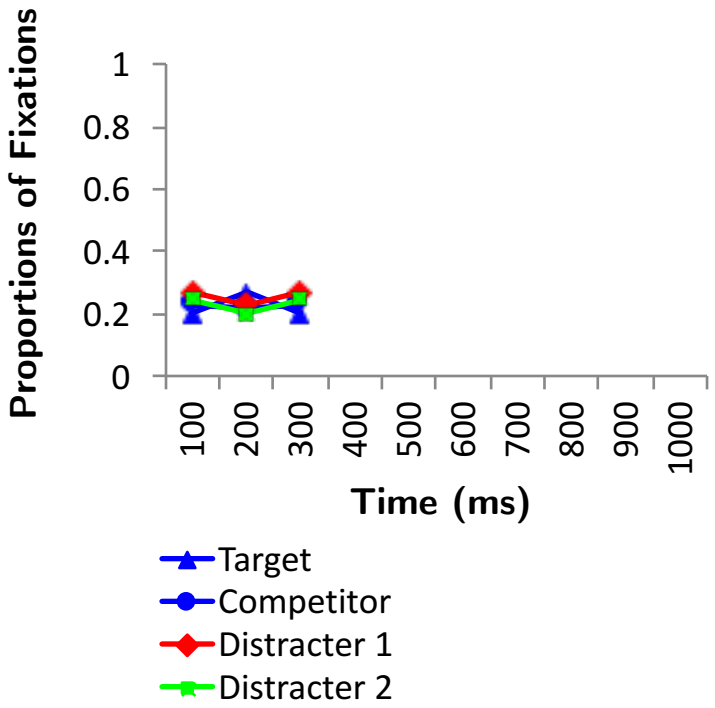
Desktop-mounted Eyelink

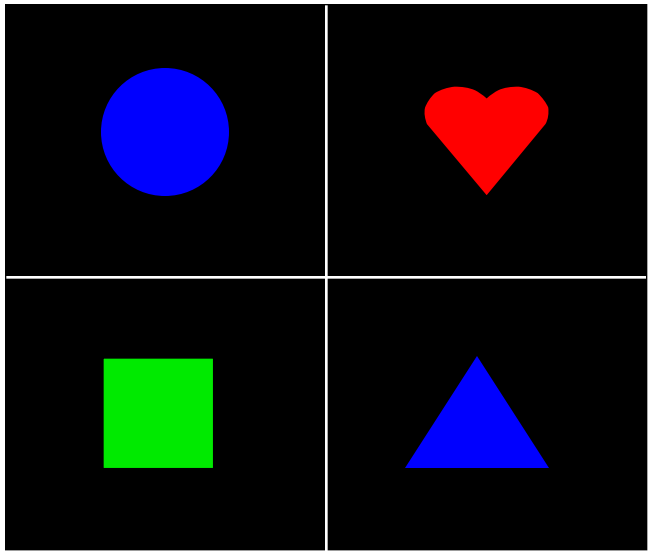
Tower-mounted Eyelink

Head-mounted Eyelink

Illustration

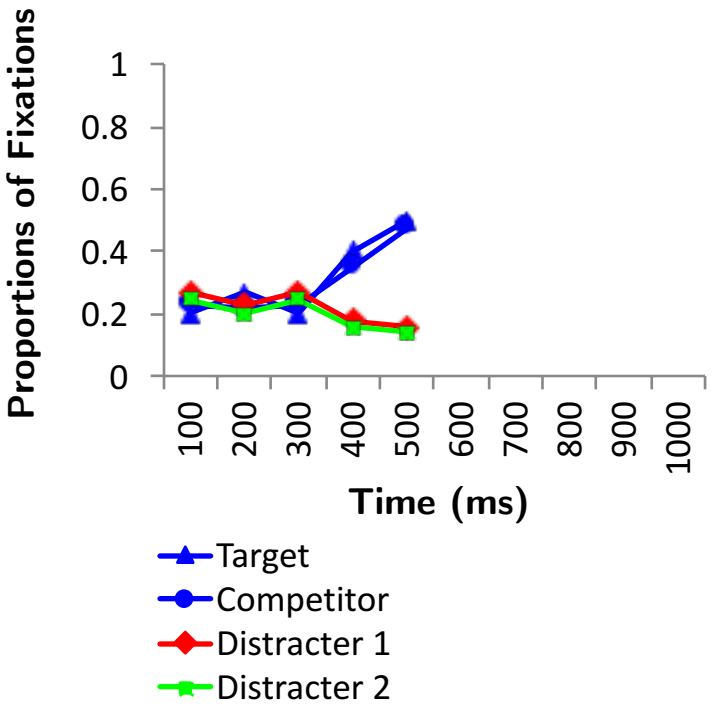
Click on the...

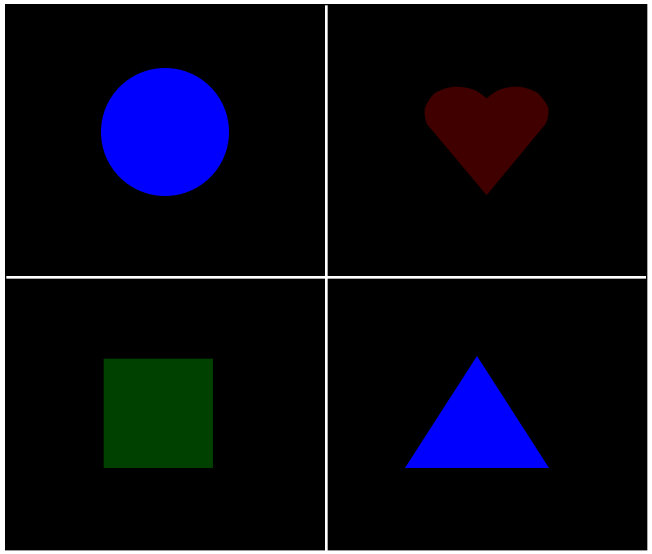




Illustration

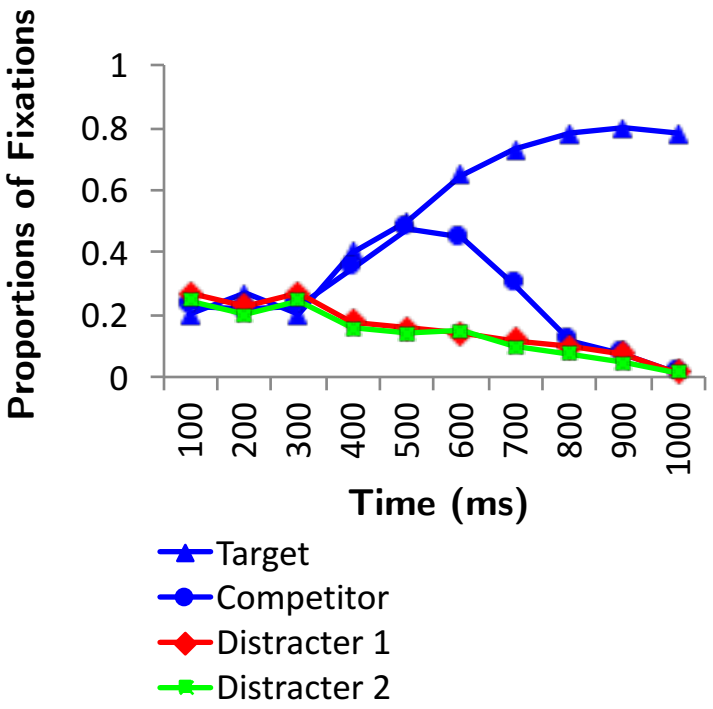
Click on the blue...

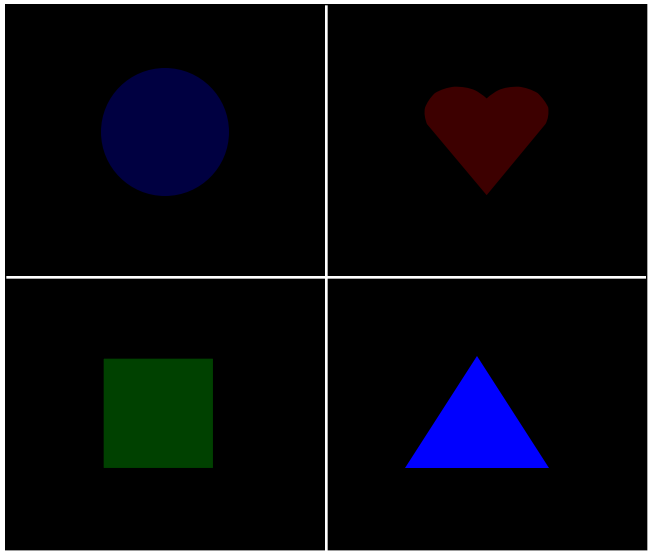




Illustration

Click on the blue triangle





Visual-World Eye-Tracking

- Started being used in psycholinguistics research in the mid 1990s (e.g., Tanenhaus et al., 1995)
- Prevalent method to investigate spoken word recognition and sentence comprehension as the speech signal unfolds
  - Provides a continuous measure of word activation/ sentence interpretation over time

Central Premises of Visual-World Eye-Tracking Research

- Lexical candidates compete for word recognition
- Lexical competition among word candidates co-varies, such that greater activation of one word will result in lower activation of another

What Visual-World Eye-Tracking Research Must Determine

- The circumstances under which particular words compete with one another (e.g., phonological overlap)
- The types of information that modulate lexical activation (e.g., fine-grained phonetic information)
- The time course of lexical activation
  (i.e., at what point in time different types of information are used)

Why Visual-World Eye-Tracking?

- More sensitive and more ecologically valid measure of speech processing than other 'online' measures (e.g., RTs) that require an explicit or conscious decision (e.g., lexical decision)
- Can provide information about the **time course** of lexical activation; no other behavioral measure can do this with the same degree of precision

Why Visual-World Eye-Tracking?

- Can shed important light on how prosodic information modulates spoken word recognition and sentence comprehension as the speech signal unfolds

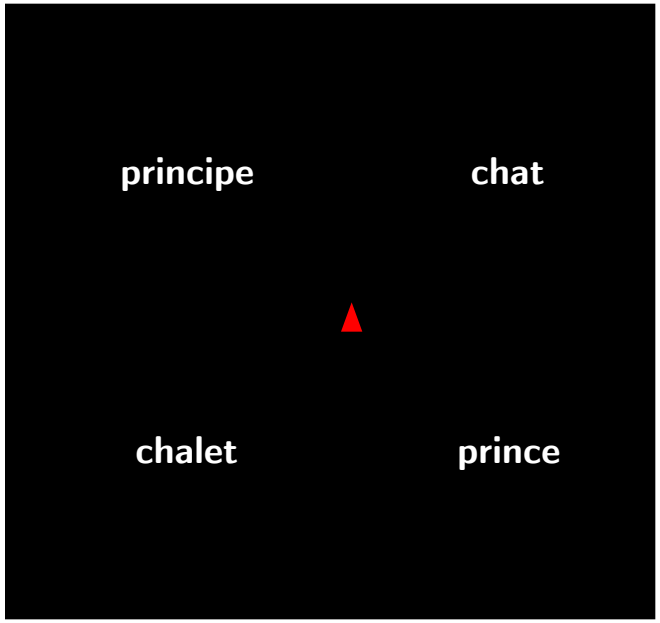
Example Prosodic Studies

- **Ito & Speer (2008, *JML*, Exp. 2)**
  - Example stimulus: *Hang the green ball/sock. Next, hang the BLUE/blue ball.*
  - Task: Participants hung the objects mentioned in the instructions
  - General results: Higher proportions of target fixations when the contrastive accent was felicitous than when it was not

Example Prosodic Studies

Example Prosodic Studies

- **Tremblay, Broersma, Coughlin, & Choi (2016, *Frontiers*)**
  - Example stimulus: *Le chat lépreux s'endort doucement* 'the leprous cat is slowly falling asleep,' where *chat* contained or did not contain an F0 rise
  - Task: Participants click on the word they heard
  - General results: Higher proportions of target fixations (*chat*) and lower proportions of competitor fixations (*chalet*) when *chat* contained an F0 rise than when it did not



Present Workshop

Methodological considerations in the creation of a visual-world eye-tracking (with focus on prosody)

Roadmap

1. Methodological validity
   - Linking hypothesis
   - Design issues
2. Stimuli preparation
   - Auditory stimuli
   - Visual stimuli
3. Data analysis
   - Dependent variables
   - Types of analyses
4. Practice
   - Brainstorming session about hypothetical study

1. Methodological Validity

What Drives Eye-Movements?
(Huettig, Rommers, & Meyer, 2011)

- "... the spoken words do not pull the eyes to certain objects on the screen; instead, the speech-eye link arises because the verbal information affects the listeners' allocation of attention, which in turn governs the direction of their gaze." (p. 166)

What Drives Eye-Movements?

(Huettig, Rommers, & Meyer, 2011)

- What listeners pay attention to depends in part on
  - The auditory stimuli they hear and the properties of the objects or printed words in the display
  - Their understanding of and compliance to task demands

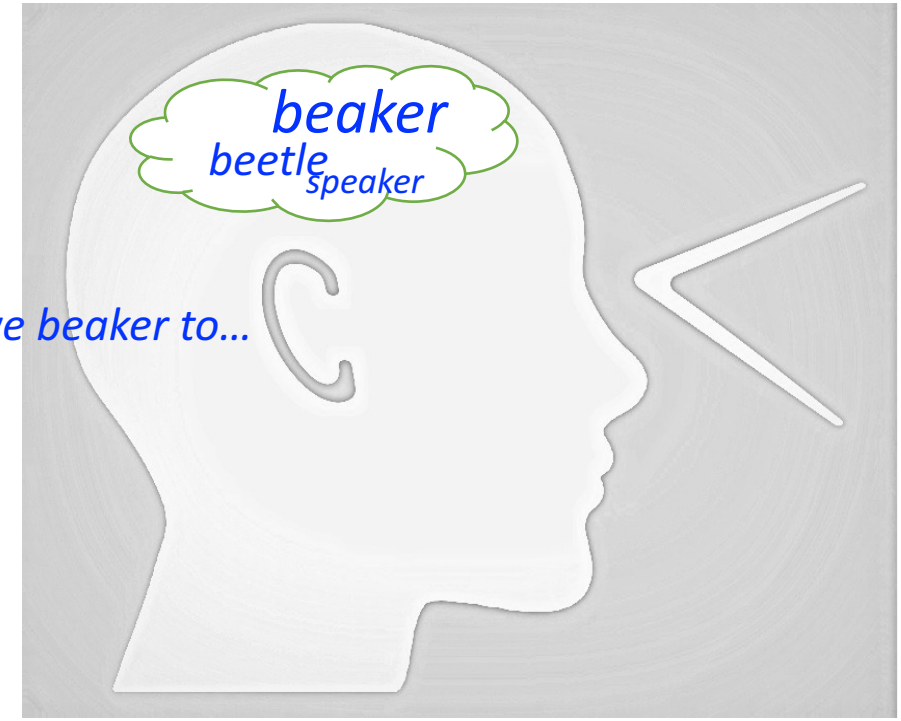
What Drives Eye-Movements?

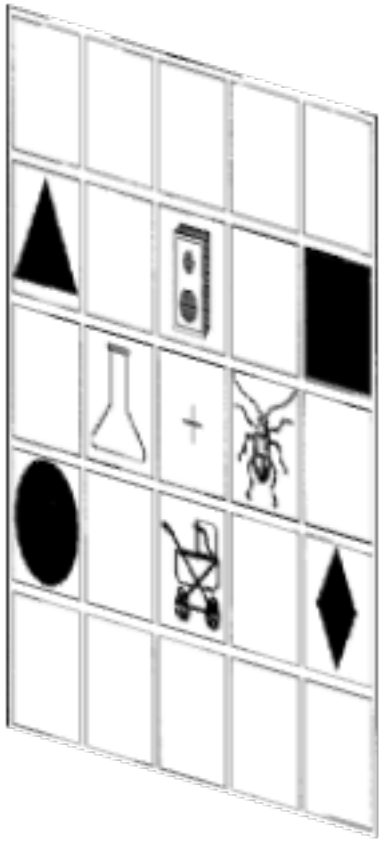
(Huettig, Rommers, & Meyer, 2011)

- What listeners pay attention to depends in part on
  - **The auditory stimuli they hear and the properties of the objects or printed words in the display**
  - Their understanding of and compliance to task demands

Linking Hypothesis (Allopenna et al., 1998)

- A hypothesis about how word activation levels map onto eye fixations



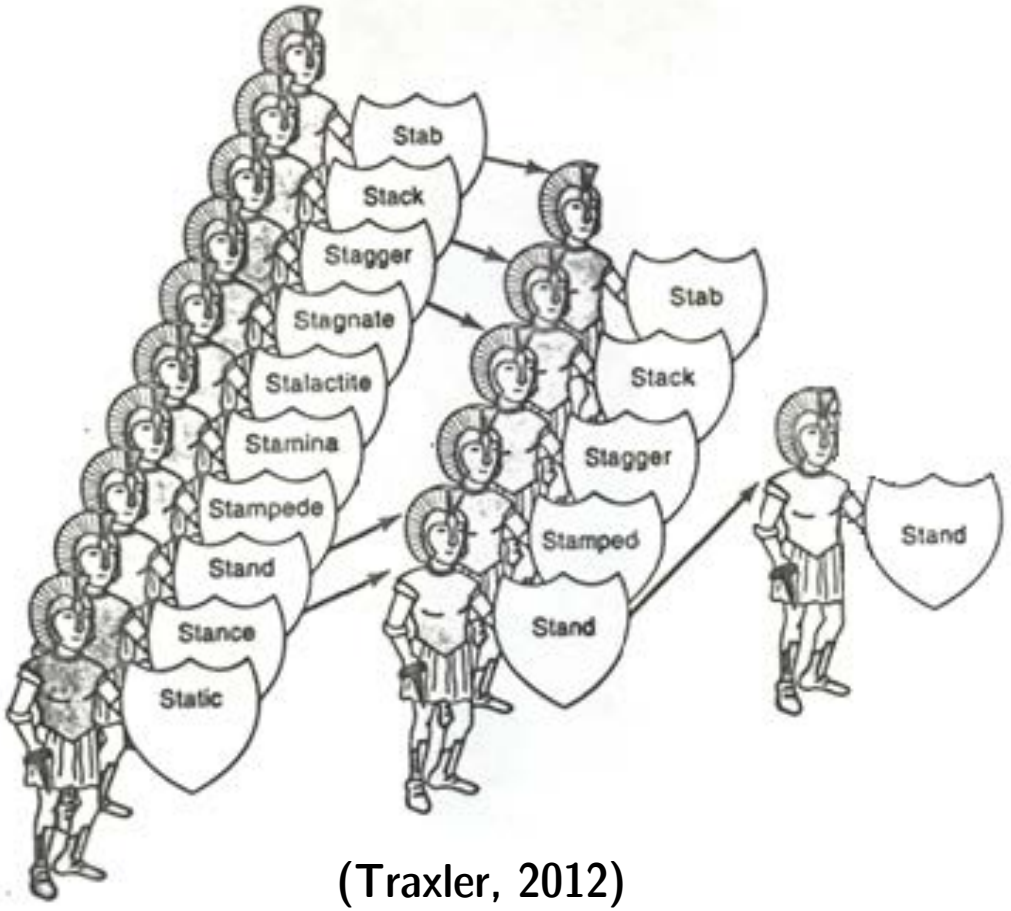


Activation Levels?

- What are activation levels? How do we identify them?

Cohort Model (Marslen-Wilson, 1987)

- Verbal algorithm model
  - Stage 1: Activation (bottom-up)
  - Stage 2: Selection (bottom-up AND top-down)
  - Stage 3: Integration



(Traxler, 2012)

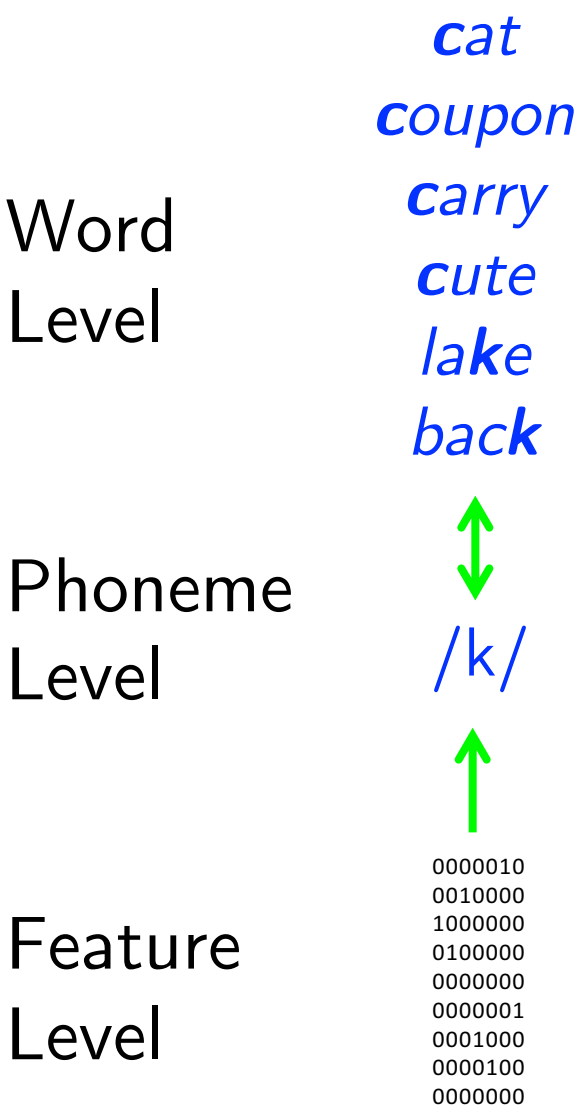
Cohort Model (Marslen-Wilson, 1987)

- The activated words are those of the Cohort
- Degree of fit with the input is instantiated in the removal of non-matching representations over time (**strict left-to-right matching**)
- Lexical competition takes place within the selection stage of the model

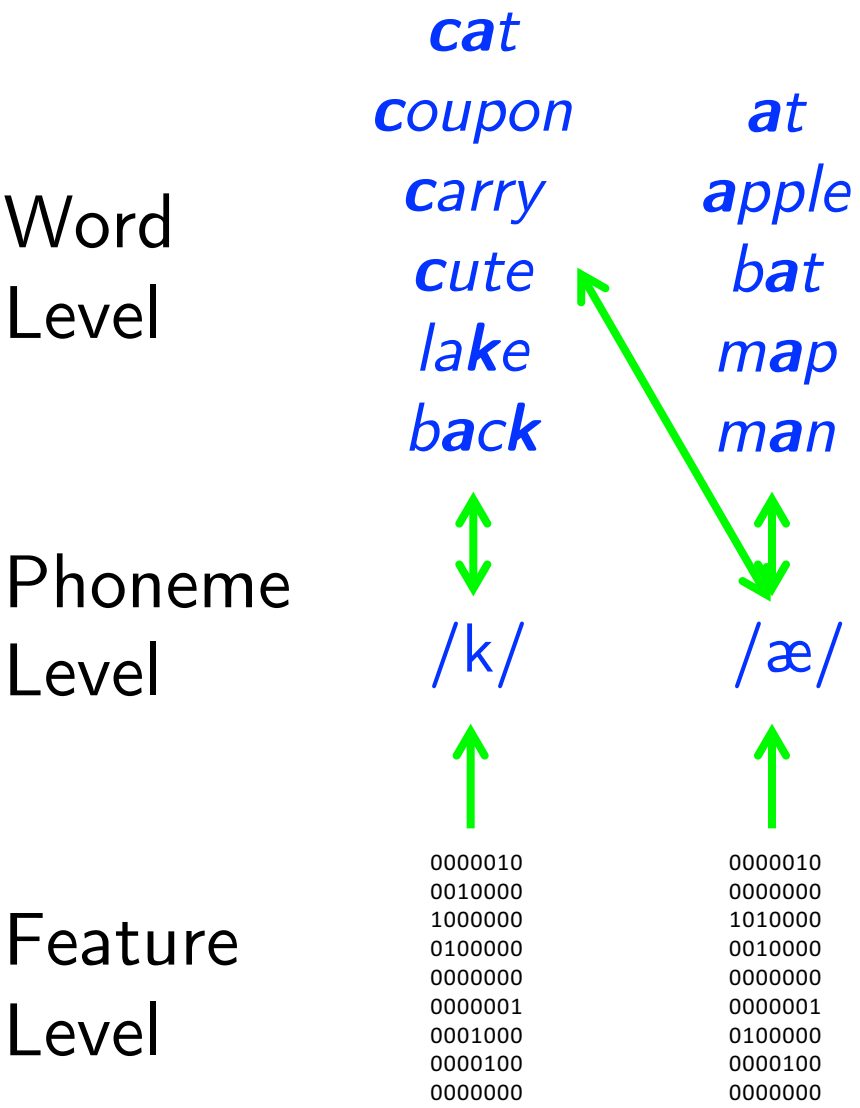
TRACE **(McClelland & Elman, 1986)**

- Connectionist model of spoken-word recognition
  - Three layers (**word, phoneme, feature**)
  - Input: Features identified from the signal
  - Output: **Word activation levels**
  - Connections: bottom-up AND top-down, cascaded
    - Excitatory between items at different levels
    - Inhibitory between items at the same level

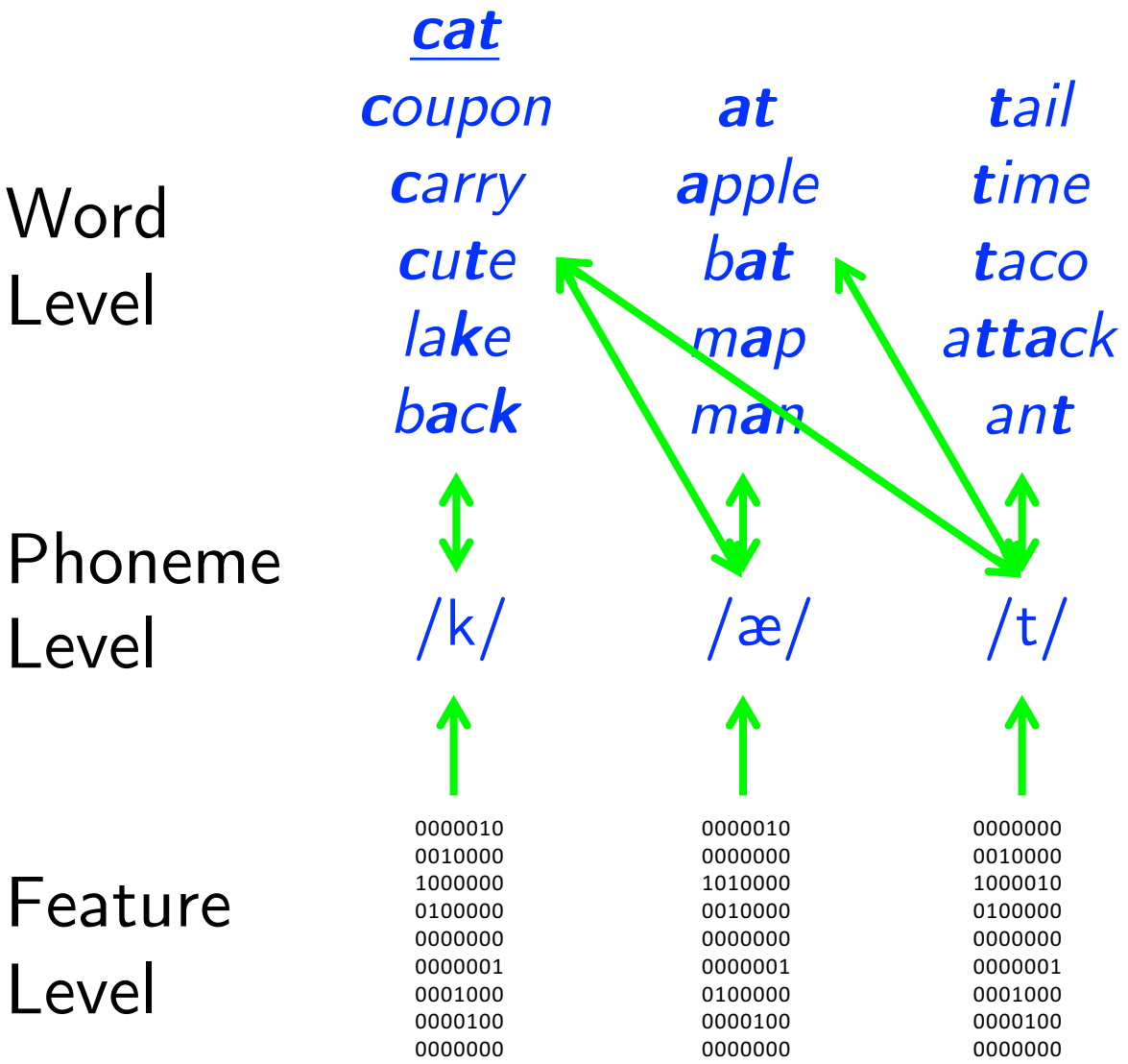
TRACE (McClelland & Elman, 1986)



TRACE (McClelland & Elman, 1986)



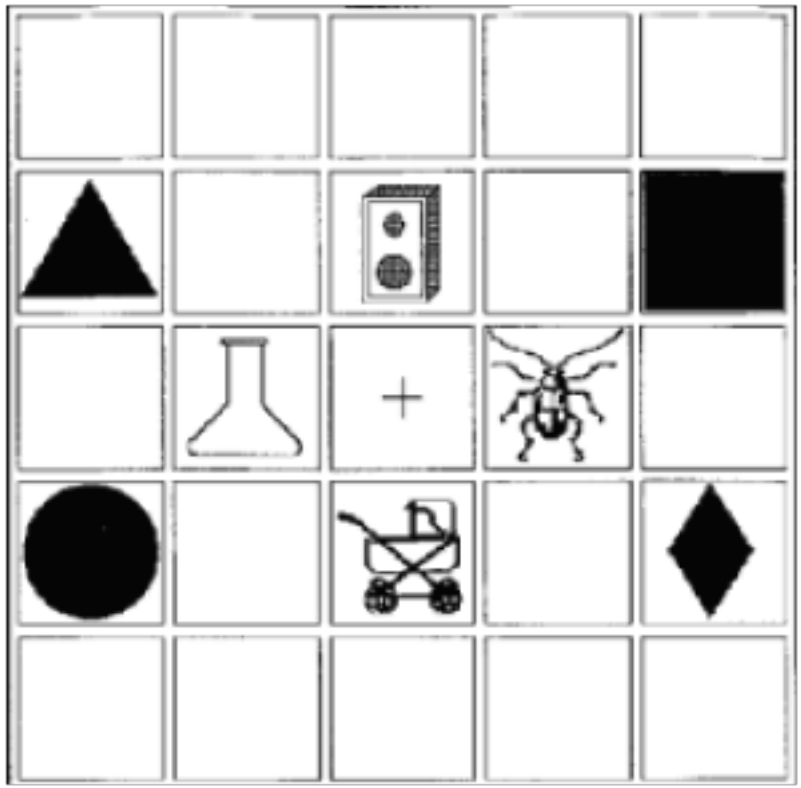
TRACE (McClelland & Elman, 1986)

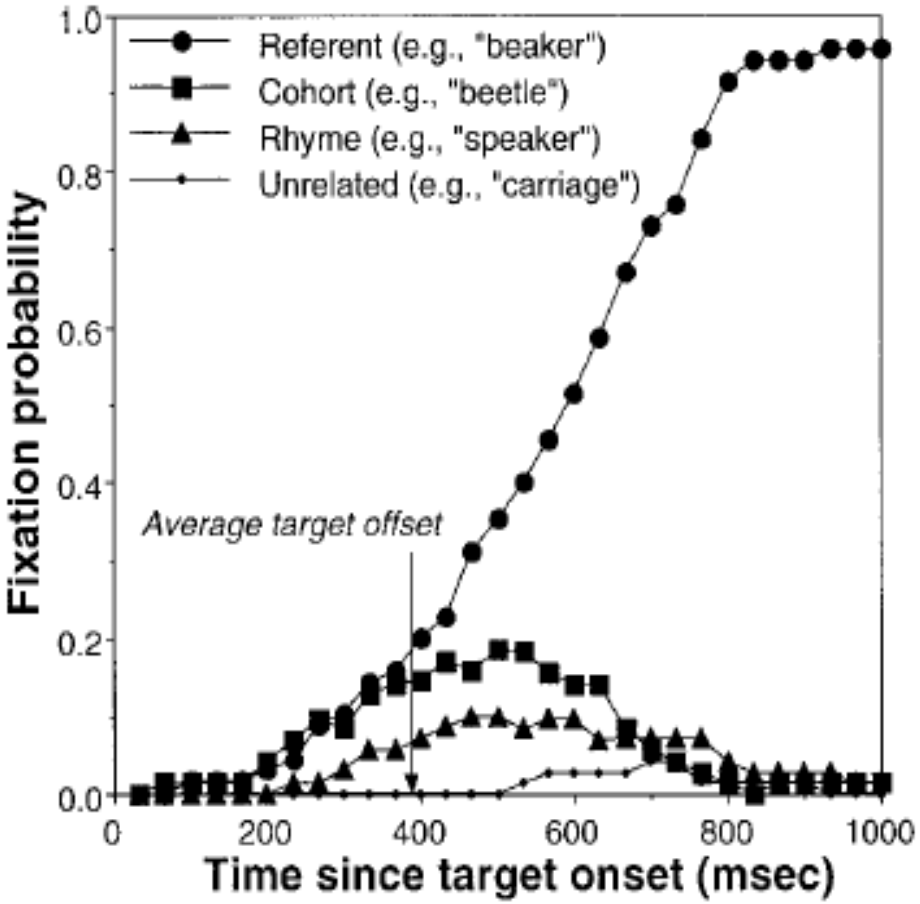


TRACE (McClelland & Elman, 1986)

- All words in the lexicon that contain one of more of the phonemes from the input are activated
- Degree of fit between the input and the activated words is determined by matching phonemes
- Word recognition is achieved though competition among the activated lexical representations

Allopenna et al. (1998)





Linking Hypothesis **(Allopenna et al., 1998)**

- The probability of initiating an eye movement to fixate the target object at any given time is a direct function of
  - The activation level of the target in relation to the activation level of other potential targets (as determined by TRACE)
  - The probability that any object on the screen is the target given the speech input

Linking Hypothesis (Allopenna et al., 1998)

TRACE Simulations

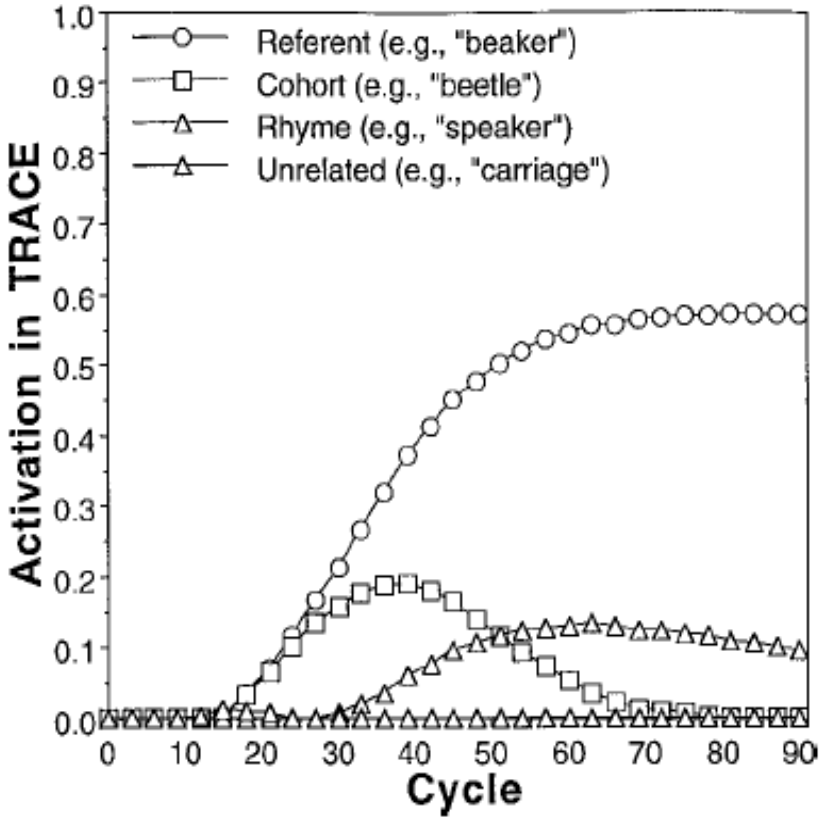


FIG. 1. Average activations from eight TRACE simulations with both cohort and rhyme competitors.

Predicted Fixations Based on Linking Hypothesis

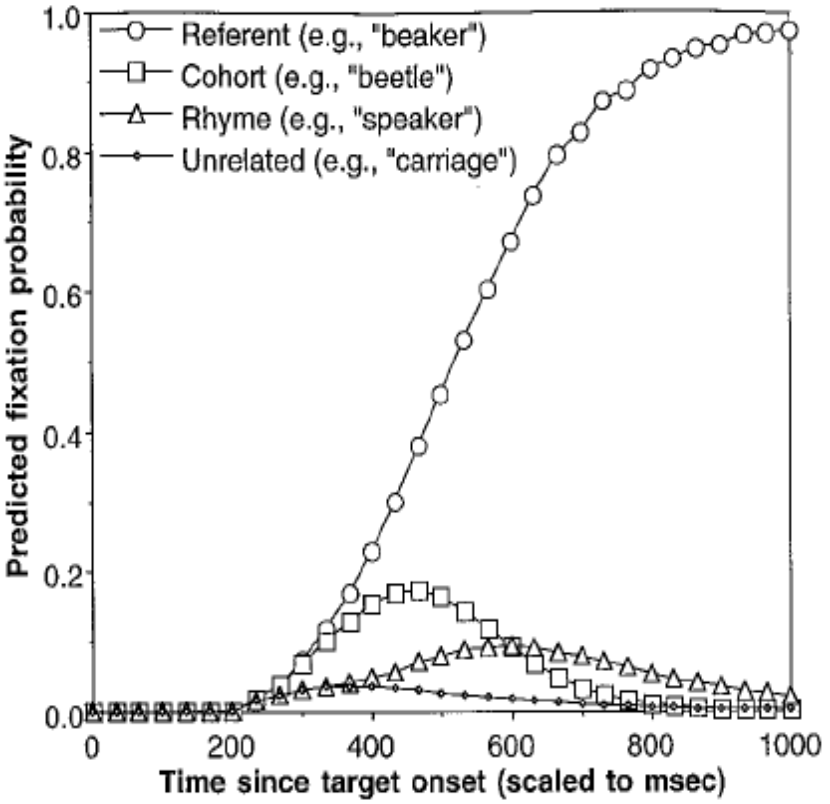


FIG. 2. Predicted response probabilities converted from TRACE using the scaled Luce choice rule.

Linking Hypothesis (Allopenna et al., 1998)

Predicted Fixations Based on Linking Hypothesis

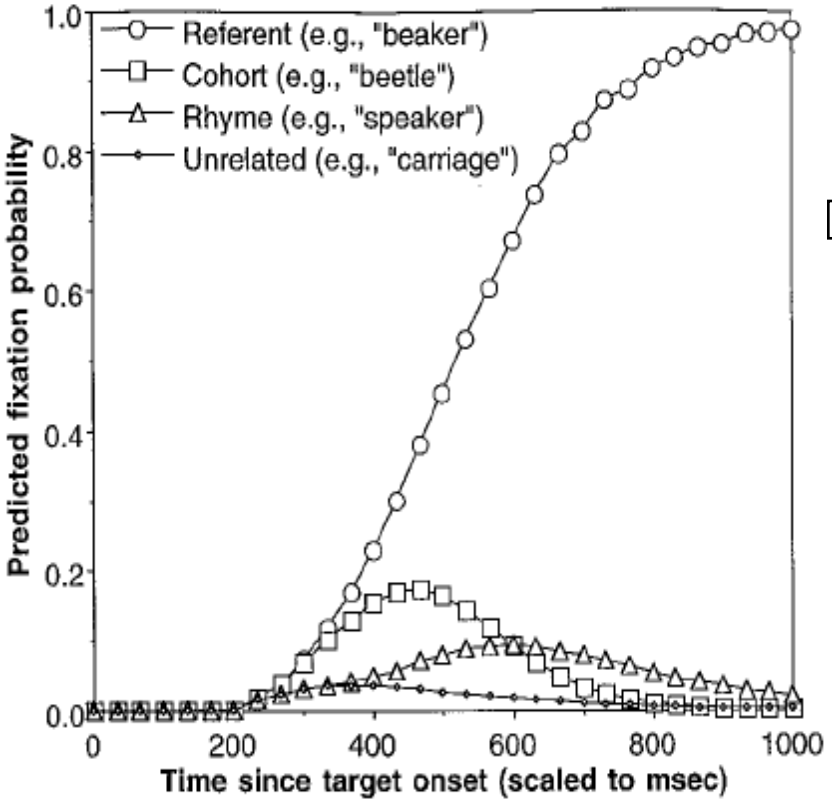


FIG. 2. Predicted response probabilities converted from TRACE using the scaled Luce choice rule.

Listeners' Actual Fixations

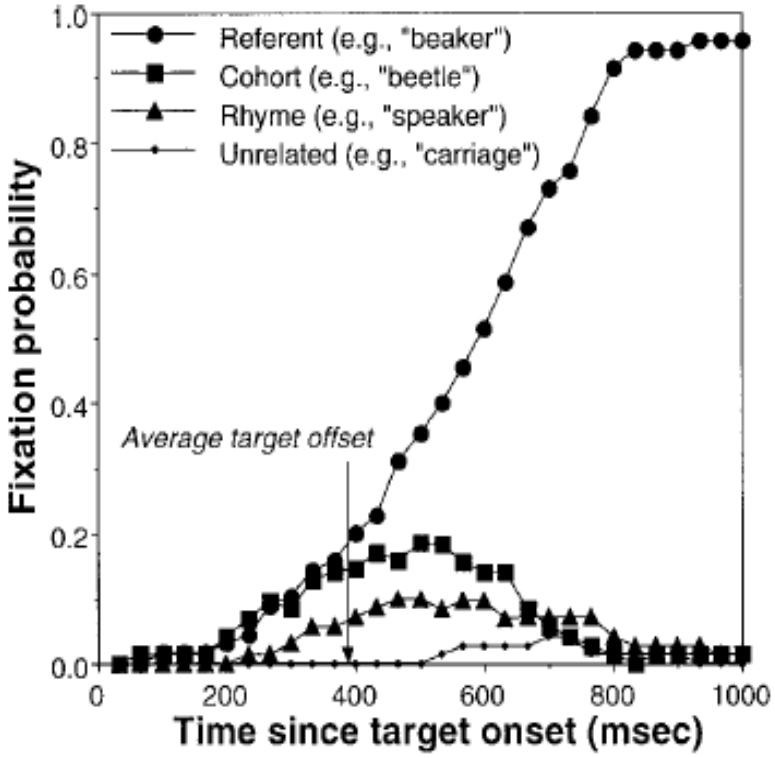


FIG. 4. Probability of fixating each item type over time in the full competitor condition in Experiment 1. The data are averaged over all stimulus sets given in Table 1; the words given in the figure are examples of one set.

Linking Hypothesis

- The exact link between word activation levels and eye fixations also depends on the task that listeners perform **(for discussion, see Pyykkönen-Klauck & Crocker, 2016; Salverda, Brown, & Tanenhaus, 2011)**
  - **Goal-oriented tasks** (e.g., *Click on X; Move X to Y*): More control over listeners' attention, so more accurate estimation of the linking hypothesis
  - **Look-and-listen tasks**: Less control over listeners' attention, so less accurate estimation of the linking hypothesis

Design Issues

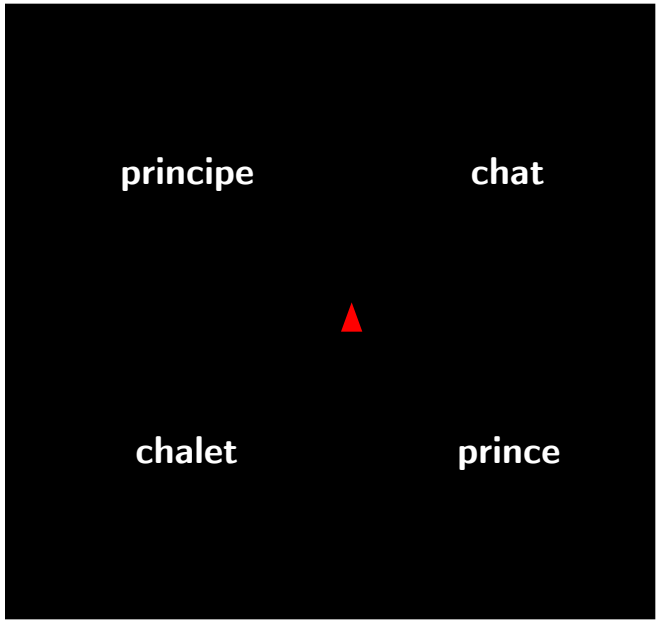
- Researchers must decide
  - Whether the visual display should contain the auditory target
  - Whether the visual display should also contain one or more competitors (determined by the research questions)
  - Whether the visual display should also contain one or more distracters (more distracters is usually better)

Design Issues

- **Ito & Speer (2008, *JML*, Exp. 2)**
  - Example stimulus: *Hang the green ball/sock. Next, hang the BLUE/blue ball.*
  - The display contains
    - The target (blue ball)
    - Several competitors (all blue objects)
    - Several distracters (all non-blue objects)

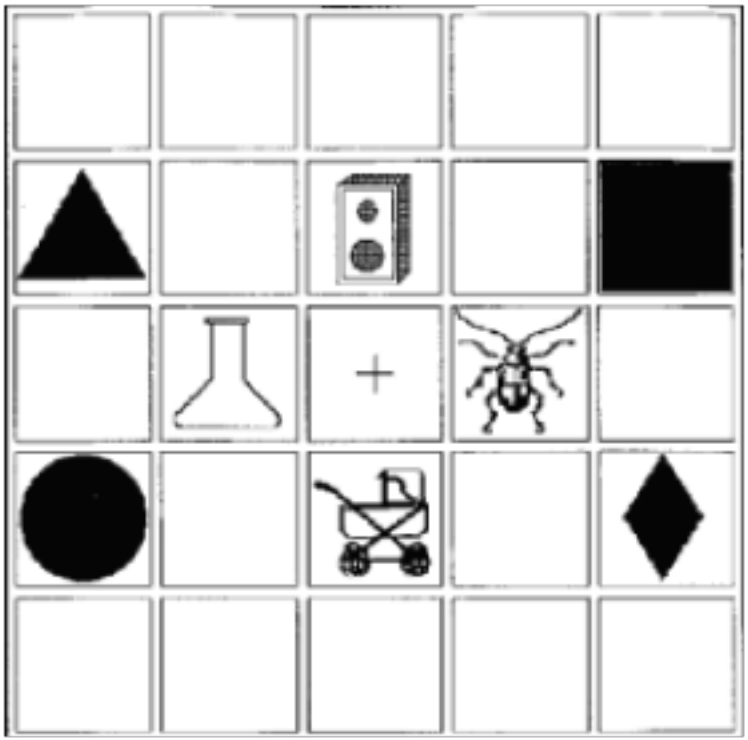
Design Issues

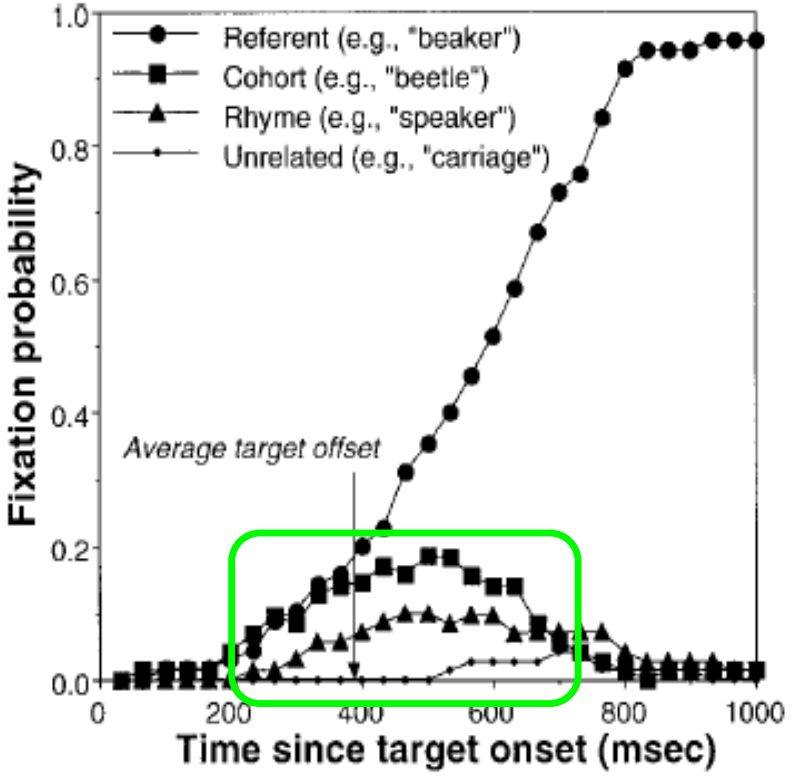
- **Tremblay et al. (2016, *Frontiers*)**
  - Example stimulus: *Le chat lépreux s'endort doucement* 'the leprous cat is slowly falling asleep,' where *chat* contained or did not contain an F0 rise
  - The display contains
    - The target (*chat*)
    - One competitor (*chalet*)
    - Two distracters (*prince*, *principe*)



Design Issues

- Distracters are needed in nearly all designs
- Illustration: **Allopenna et al. (1998)**





2. Stimuli Preparation

Auditory Stimuli

- Auditory target words in isolation or embedded in a carrier phrase or sentence
- In prosodic studies, the target word is usually presented in context
  - **Ito & Speer (2008, *JML*, Exp. 2)**
    - *Hang the green ball. Next, hang the BLUE ball.*
  - **Tremblay et al. (2016, *Frontiers*)**
    - *Le chat lépreux s'endort doucement.*
      'The leprous cat is slowly falling asleep.'

Auditory Stimuli

- If context adds too much complexity (e.g., processing of lexical tones in Chinese), listeners can instead go through an auditory word familiarization phase before the experiment
  - That way, listeners can become familiar with the talker's pitch range without needing an immediate context in the task

Visual Stimuli

Arrays of objects

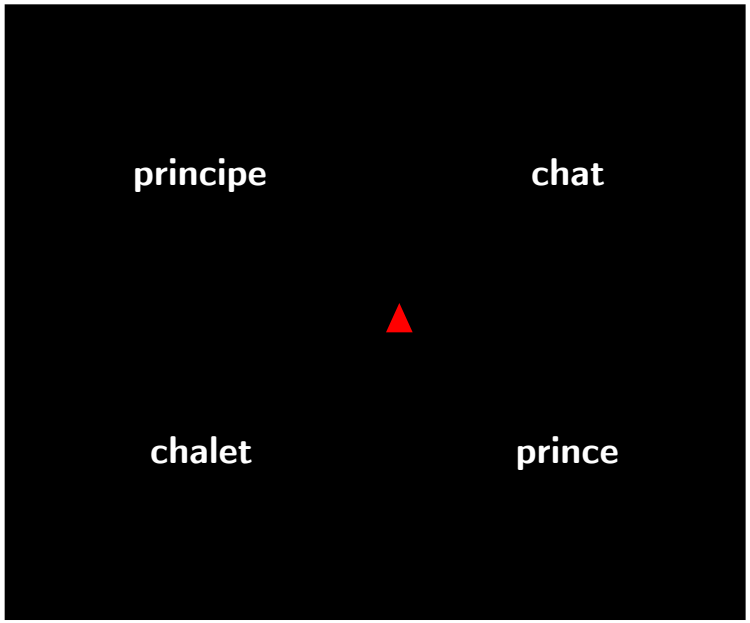
- Better for examining the recognition of individual words
- Sensitive to both phonological and semantic information
  (e.g., Huettig & McQueen, 2007)

(Ito & Speer, 2008, *JML*)

Visual Stimuli

Arrays of printed words

- Better for examining the recognition of individual words that may not be easily imageable
- More sensitive to phonological information, less sensitive to semantic information
  (e.g., Huettig & McQueen, 2007)

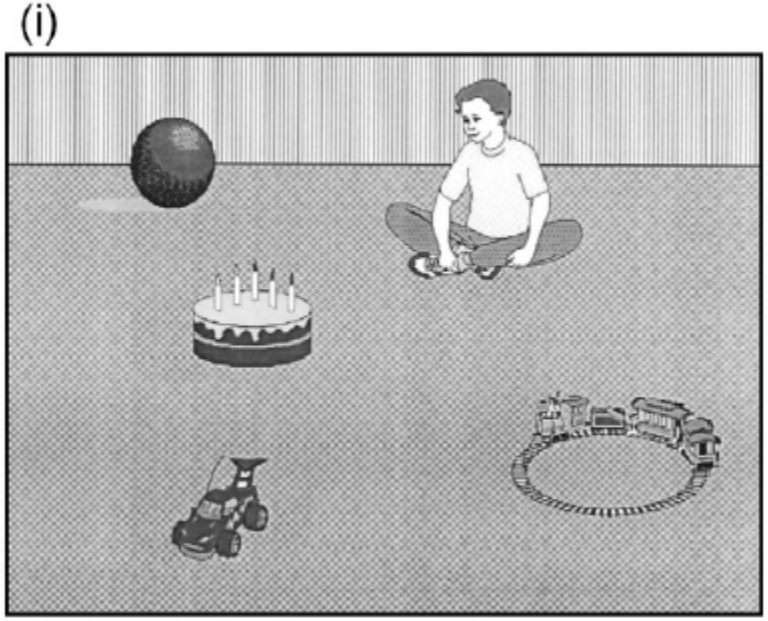


(Tremblay et al., 2016, *Frontiers*)

Visual Stimuli

Semi-realistic visual scenes

- Better for examining how listeners knowledge of real-world scenes and events affect their understanding of spoken utterances



(Huettig, Rommers, & Meyer, 2011)

Visual Stimuli

- Timing: The visual display appears **before** the speech stimulus is heard and remains on the screen until listeners' response
  - Words **more** likely to be phonologically activated prior to hearing the speech stimulus
  - Thus, **earlier and more robust** effects of phonological competition **(Huettig & McQueen, 2007)**

Visual Stimuli

- Timing: The visual display appears **as** the speech stimulus is heard and remains on the screen until listeners' response
  - Words **not** phonologically activated prior to hearing the speech stimulus
  - Thus, **later and less robust** effects of phonological competition **(Huettig & McQueen, 2007)**

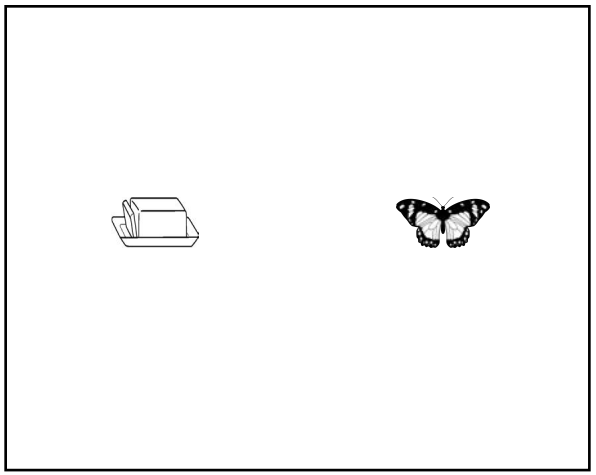
Visual Stimuli

- Lexical factors to control for when selecting **words** to be represented in the display
  - Phonological overlap with other words in the display
  - Semantic overlap with other words in the display
  - Orthographic overlap with other words in the display (for arrays of printed words)
  - Phonological length
  - Orthographic length (for arrays of printed words)
  - Imageability (for arrays of objects)
  - Type and token frequency
  - Neighborhood density
  - etc.

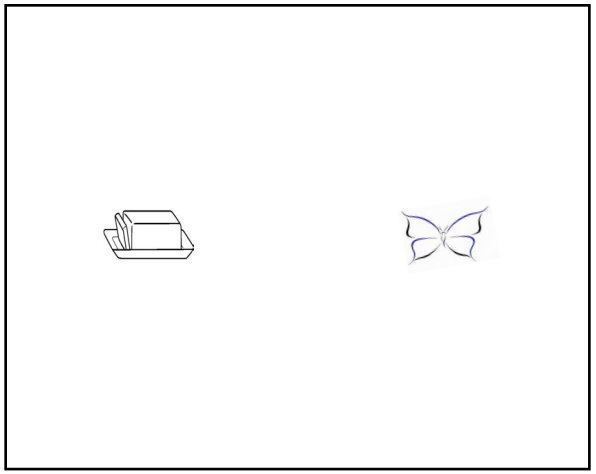
Visual Stimuli

- Visual factors to control for in arrays of objects
  - **Visual salience:** No object should be visually more salient than the others unless the salience is intended
  - Using pre-normed picture sets may help solve this issue
    (e.g., Snodgrass & Vanderwart, 1980)

Not Great

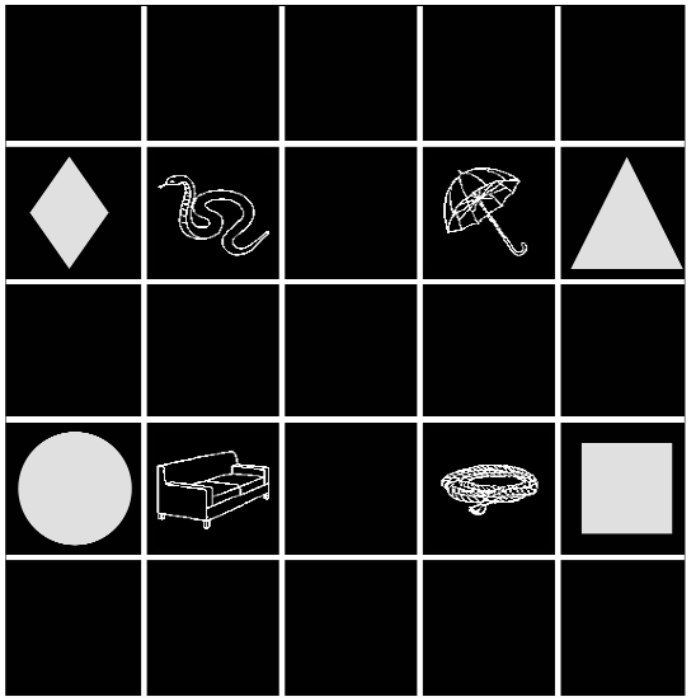


Better



Visual Stimuli

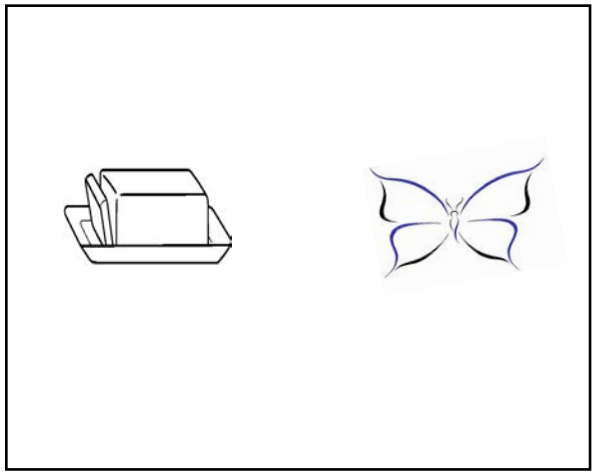
- Visual factors to control for in arrays of objects
  - **Perceptual similarity:** Objects not intended to compete should not be perceptually similar
  - **Dahan & Tanenhaus (2005)**: *rope* competes with *snake*



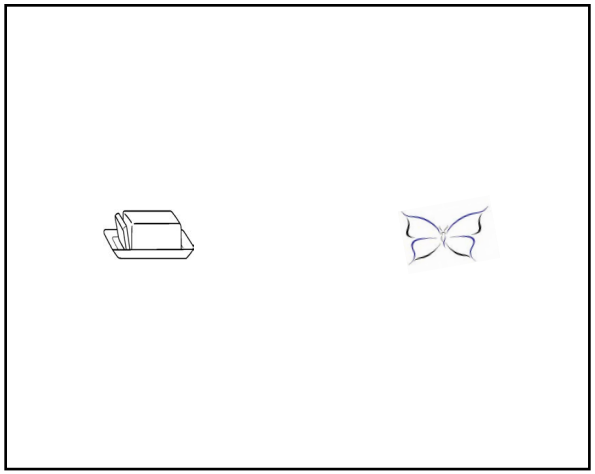
Visual Stimuli

- Visual factors to control for in arrays of objects
  - **Object size**: When using arrays of objects, ensure that the objects are not so big that listeners can do the task with their peripheral vision

Not Great

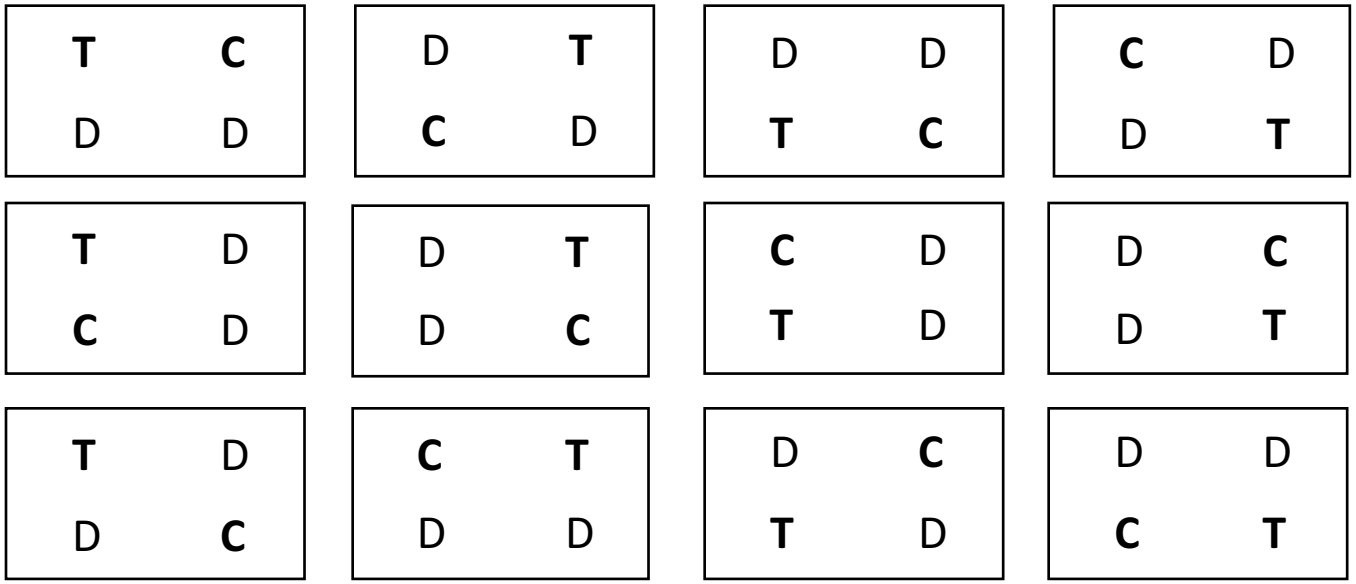


Better



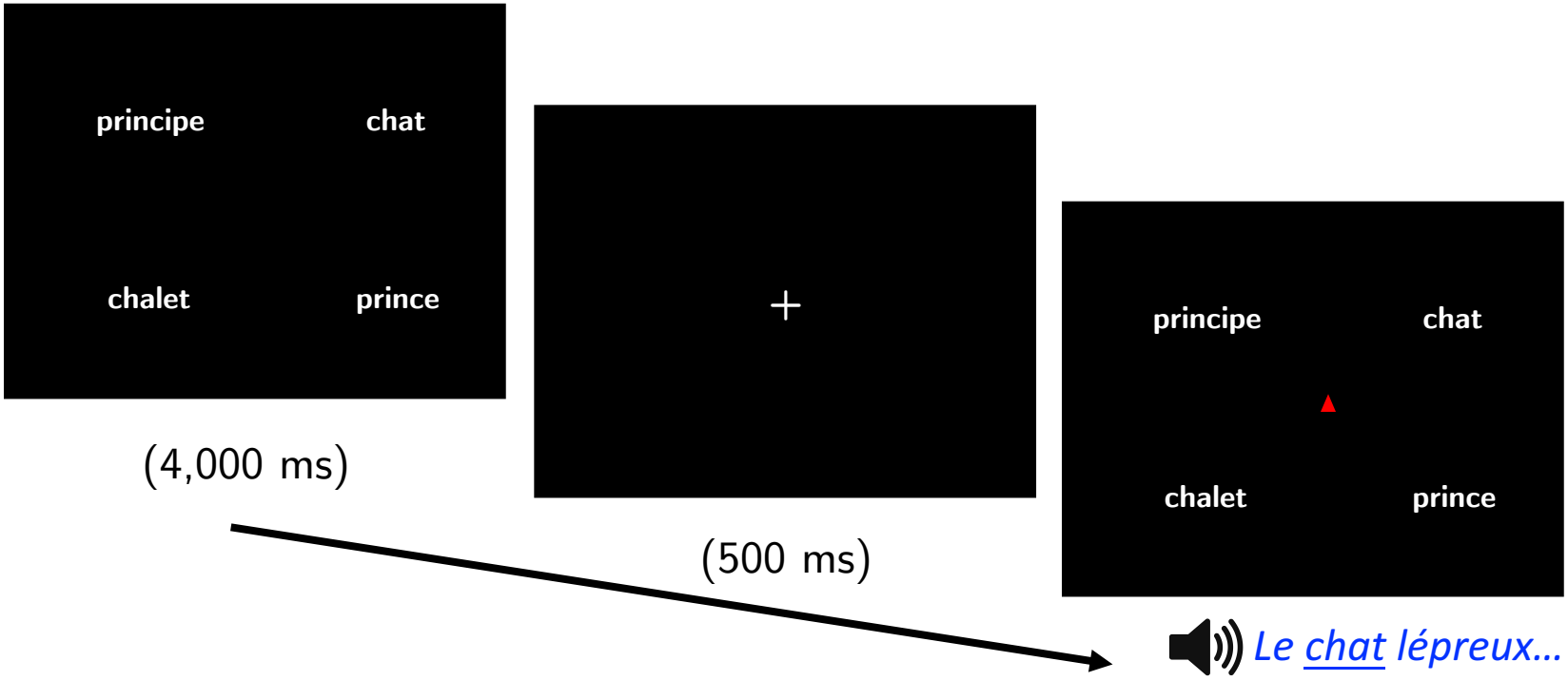
Visual Stimuli

- Visual factors to control for in arrays of objects
  - **Randomization of target and competitor visual locations** across experimental trials and throughout the experiment



Visual Stimuli

- One additional way to help reduce early visual biases is to have a **fixation cross** appear on the screen right before the speech stimulus is heard



(Tremblay et al., 2016, *Frontiers*)

Time-Locking Fixations to the Speech Signal

- Fixations in different auditory conditions can be compared only if the signal segmentally disambiguates between the target and competitor words at the same time in relation to when eye fixations are time-locked to the speech signal
- Illustration: **Tremblay et al. (2016, *Frontiers*)**

Tremblay et al. (2016): Materials

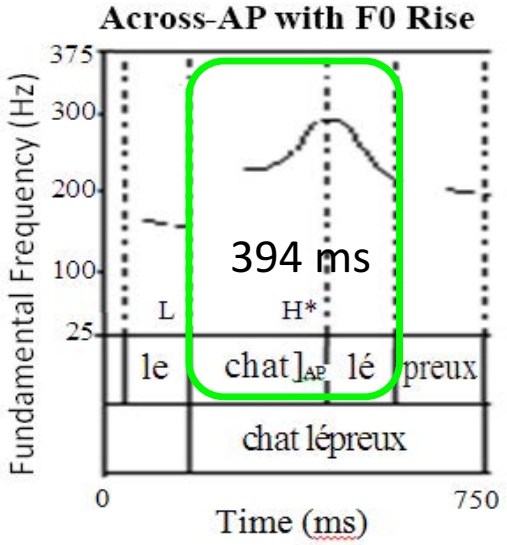
	Across-AP	**Within-AP**
F0 Rise	H* L [*Le chat*]$_{AP}$ [l*épreux*... (natural)	H L [*Le chat lépreux*]$_{AP}$ ... (resynthesized)
No F0 Rise	L H [*Le chat*]$_{AP}$ [*lépreux*... (resynthesized)	L H* [*Le chat lépreux*]$_{AP}$ ... (natural)
	'the leprous cat'	

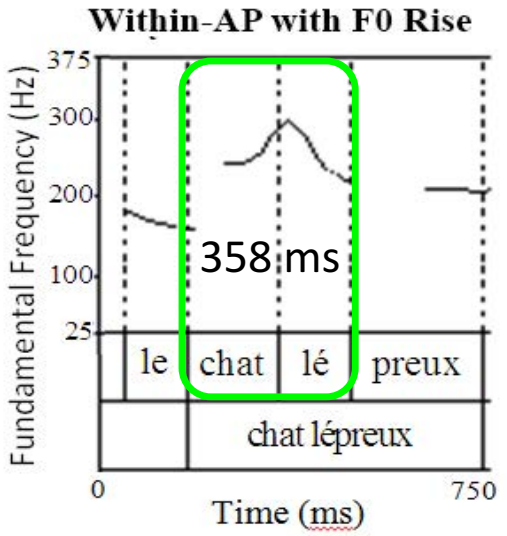
Target: *chat* 'cat'

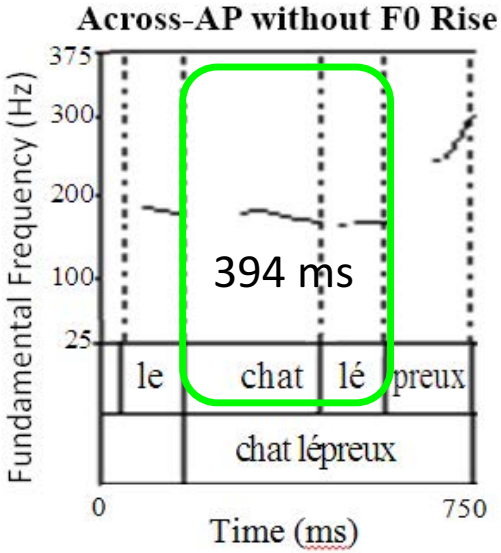
Competitor: *chalet* 'cabin'

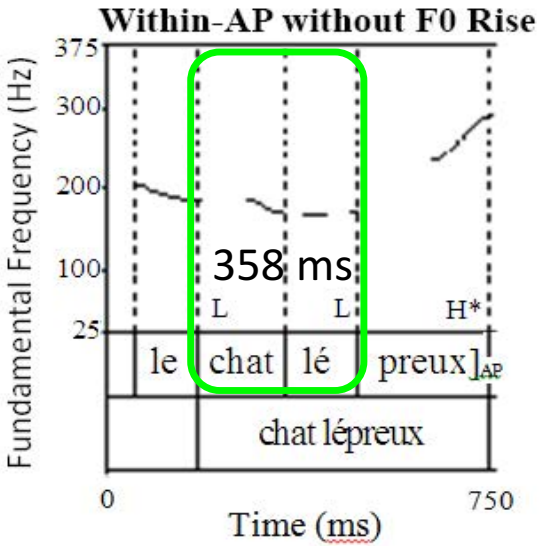
Fixations time-locked to the onset of the target: *chat* 'cat'

Tremblay et al. (2016): Stimuli







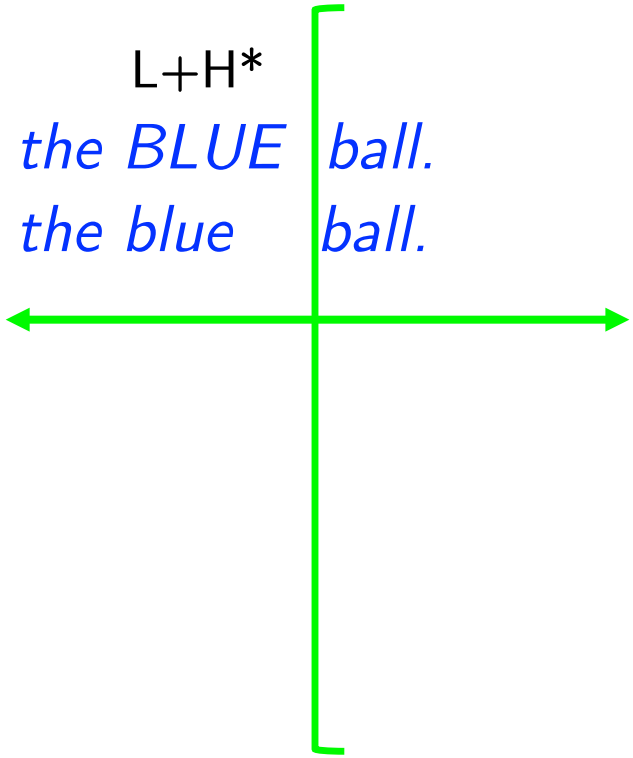


Time-Locking Fixations to the Speech Signal

- Splicing/resynthesis of the auditory stimuli can help control for duration
- One alternative solution can be to time-lock eye movements from the segmental disambiguation point (e.g., *-preux)* rather than from target-word onset (e.g., *chat*) and analyze eye fixations prior to and after the disambiguation point

Time-Locking Fixations to the Speech Signal

- This is what **Ito & Speer (2008, *JML*, Exp. 2)** did in their study
  - *Hang the green ball. Next, hang the BLUE* | *ball.*
  - *Hang the green ball. Next, hang the blue* | *ball.*

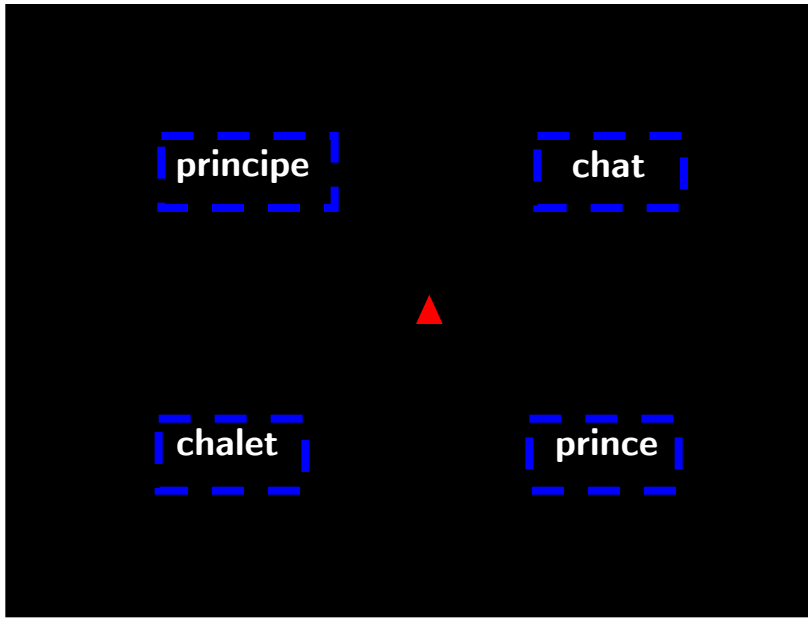


3. Data Analysis

Dependent Variables

- What word(s) do participants look at?
- Is usually automated (interest areas)

(Ito & Speer, 2008, *JML*)



(Tremblay et al., 2016, *Frontiers*)

Dependent Variables

Dependent Variables

- Fixations to target, competitor, and distracter words
  - Binomial (1 = looking; 0 = not looking)
  - Recorded for each sample (1 every 1-4 ms)
- Proportions of target, competitor, and distracter fixations (fixations averaged over some time unit)
  - Continuous and bounded (ranges from 0 to 1)
  - Time unit must be identified
- Difference between proportions of target and competitor fixations
  - Continuous and bounded (ranges from -1 to 1)

Dependent Variables

- Number of target, competitor, and distracter fixations (averaged over some time unit, weighed for fixation duration)
  - Continuous and unbounded
- (Cumulative) proportions of trials in which the target, competitor, and distracter words were fixated (averaged over some time unit)
  - Continuous and bounded
- Timing/latency of first fixation to the target word, or timing/latency of disambiguation between target and competitor words
  - Continuous and unbounded

Dependent Variables

- Baseline
  - It takes approximately 150-200 ms to launch an eye-movement reflecting the intake of the speech signal **(cf. Altmann, 2011)**
  - Up until then, fixations to any of the words should be at chance
  - Statistical analyses within that time window can test whether fixations differ from chance (they should not)
    - If differences are found, statistical analyses should account for them / distinguish them from effects attributed to the speech signal **(e.g., Barr, Gann, & Pierce, 2011)**

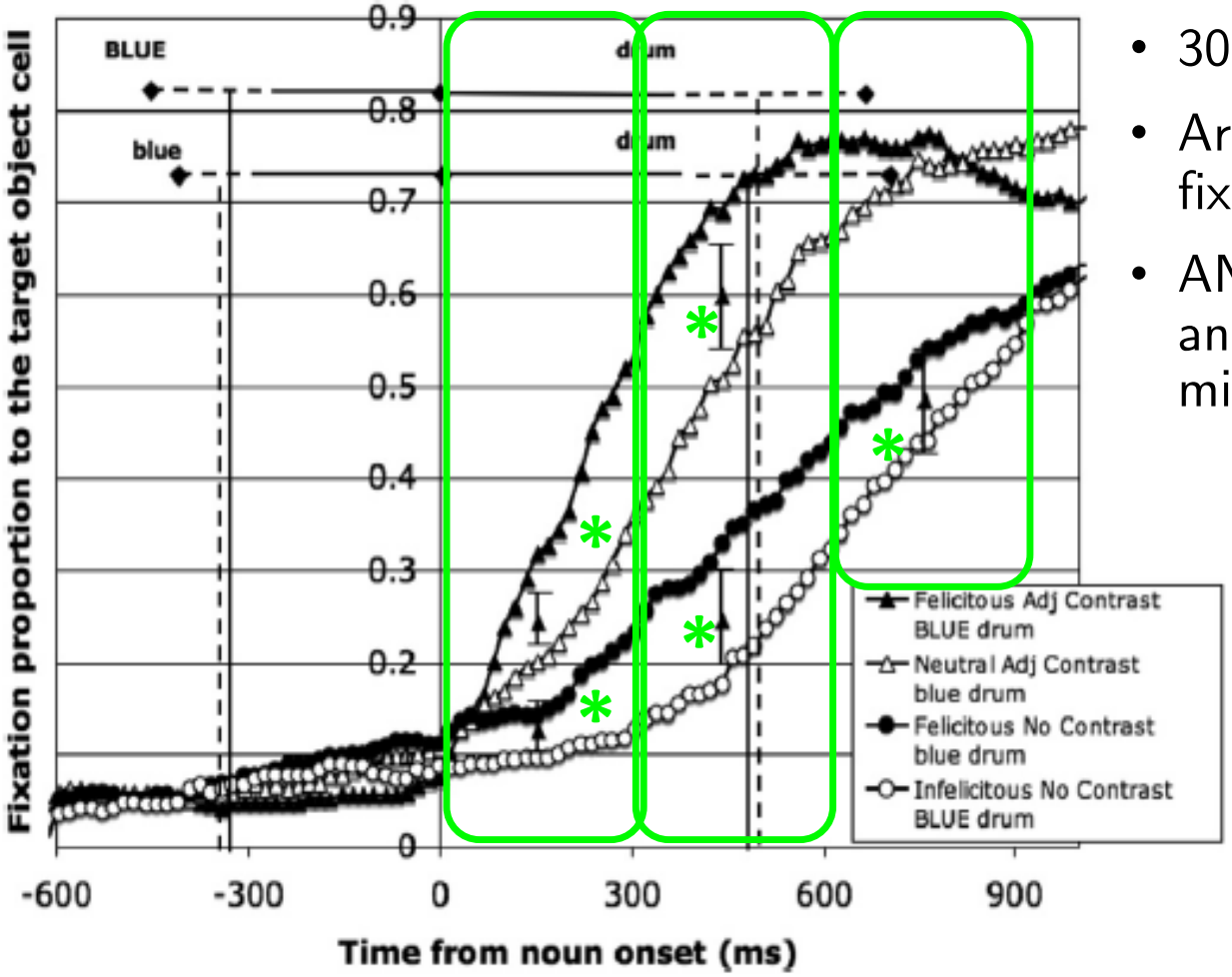
Dependent Variables

- Issues of concern
  - Bounded dependent variables become less normally distributed as they approach 0 or 1
    - **Transformations may be required**
  - The probability of fixating Object A at Time X is not independent from the probability of fixating Object A at Time X + 1
    - **Time should not be in the analysis if ANOVAs (and other analyses with similar assumptions) are used**
  - The probability of fixating Object A is not independent from the probability of fixating Object B
    - **Fixations to different objects within the same trial should not be directly compared**

Types of Analyses

- Time window analyses (traditional)
  - Listeners' proportions of fixations are **averaged** within a particular time window (sometimes with a 150-200-ms delay from the target-word onset), with this average being the dependent variable
  - What time window?
    - Target-word onset to offset vs. post-target-word offset
    - Target word onset to disambiguation vs. post-disambiguation
    - Target word onset to time increment (e.g., every 300 ms)
    - Target-word onset to end of trial
  - Example: **Ito & Speer (2008, *JML,* Exp. 2)**

Types of Analyses:
Ito & Speer (2008, *JML*, Exp. 2)



- 300-ms time windows
- Arcsine-transformed fixation proportions
- ANOVAs (can also be analyzed with linear mixed-effects models)

Types of Analyses

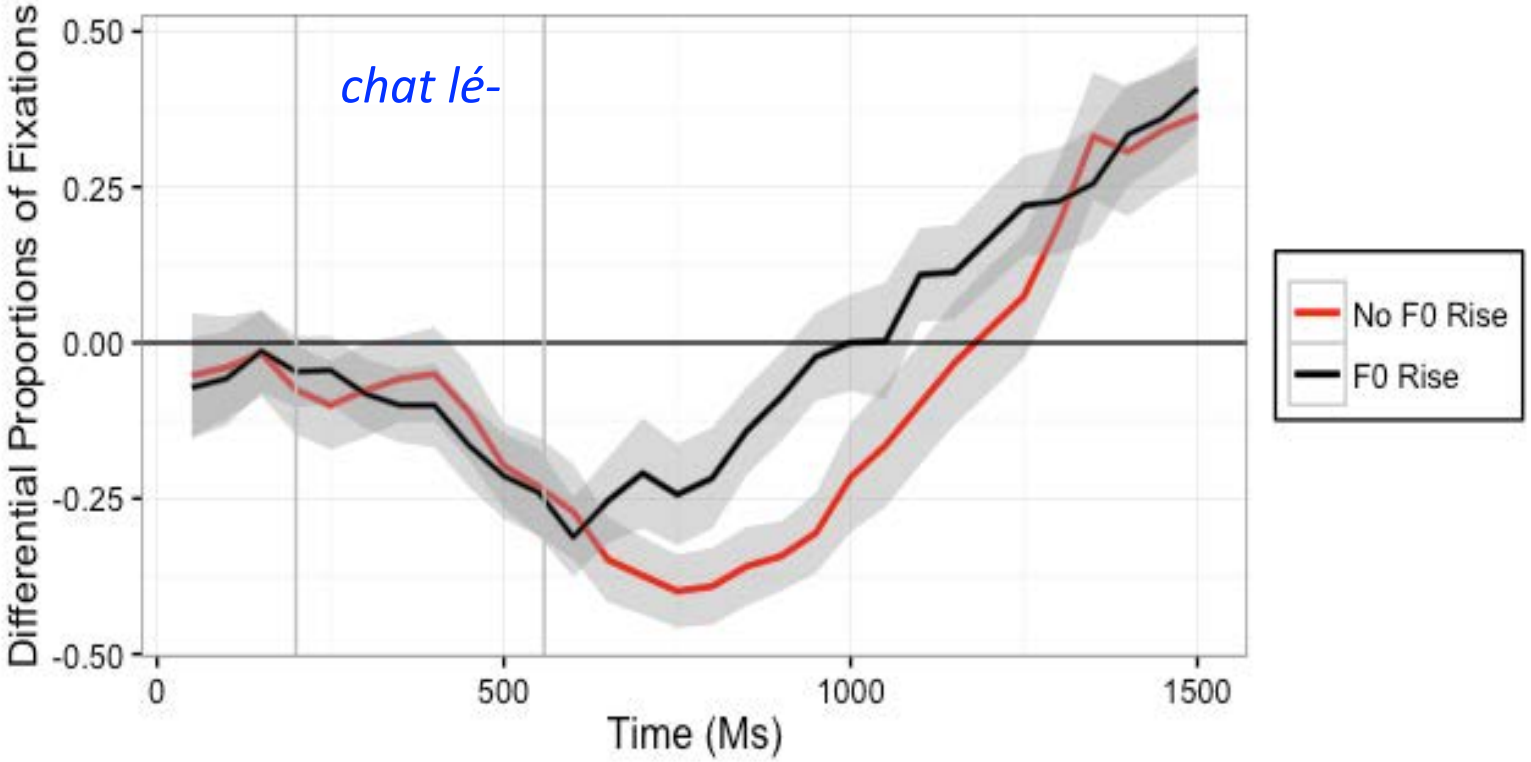
- Advantages of time window analyses
  - Analysis and interpretation are relatively simple
- Limitations of time window analyses
  - The number and size of time windows can be subjective and/or arbitrary
  - Less likely to capture subtle acoustic effects
  - Can miss effects when fixation lines cross over (see previous graph)
  - Transformations of proportions are often necessary

Types of Analyses

- Growth Curve Analyses (GCA)
  (e.g., Mirman, 2014; Mirman, Dixon, & Magnuson, 2008)
  - Listeners' proportions fixations are **averaged** within very brief time windows (e.g., in 20-ms time windows) and modeled over time with orthogonal time polynomials (typically: linear, quadratic, cubic)
  - Fixations can be modeled from target-word onset
  - Effects of conditions must interact with time for them to be attributable to the speech signal
  - Analyses typically include time polynomials as random slope for the participant variable
  - Example: **Tremblay et al. (2016, *Frontiers*)**

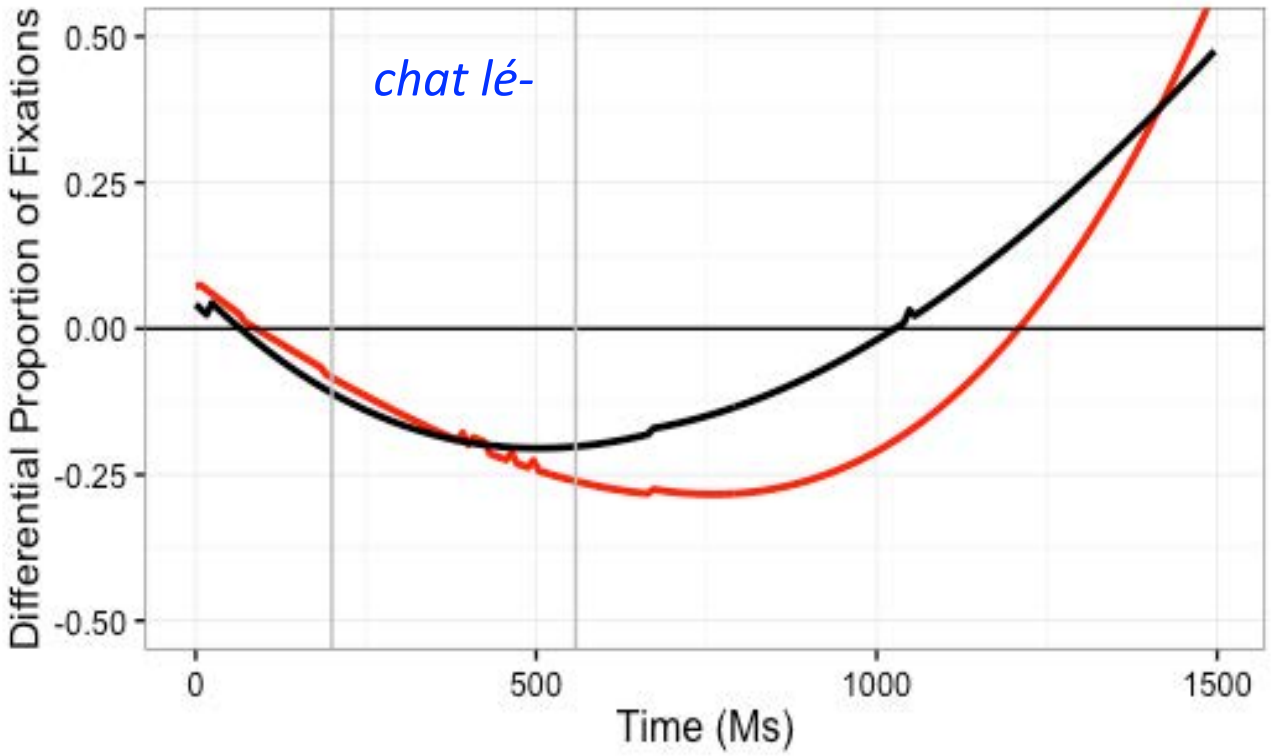
Types of Analyses: **Tremblay et al. (2016, *Frontiers,* French listeners, Within-AP Condition)**

- **Actual** difference between proportions of target and competitor fixations



Types of Analyses: **Tremblay et al. (2016, *Frontiers,* French listeners, Within-AP Condition)**

- **Modeled** difference between proportions of target and competitor fixations



- Proportions not transformed

Effects:

- F0
- F0 x time (linear)
- F0 x time (quadratic)
- F0 x time (cubic)

Types of Analyses

- Advantages of GCA
  - Can model the shape of the fixation line
  - Can capture subtle acoustic effects, even when lines cross over
- Limitations of GCA
  - Can be complex to interpret, especially when analyses involve more than one categorical variable
  - Cannot pinpoint to the exact point in time where fixations become different
  - Transformations of proportions are often necessary

Types of Analyses

- Logit-based GCA **(see also Barr, 2008)**
  - Whether or not listeners fixate the word at each sample recorded (binomial) is modeled over time using orthogonal time polynomials (typically: linear, quadratic, cubic)
- Advantages of logit-based GCA
  - Same as GCA, plus no need to transform the data
- Limitations of logit-based GCA
  - Same as GCA other than for the possible need to transform the data
  - Models less likely to converge?

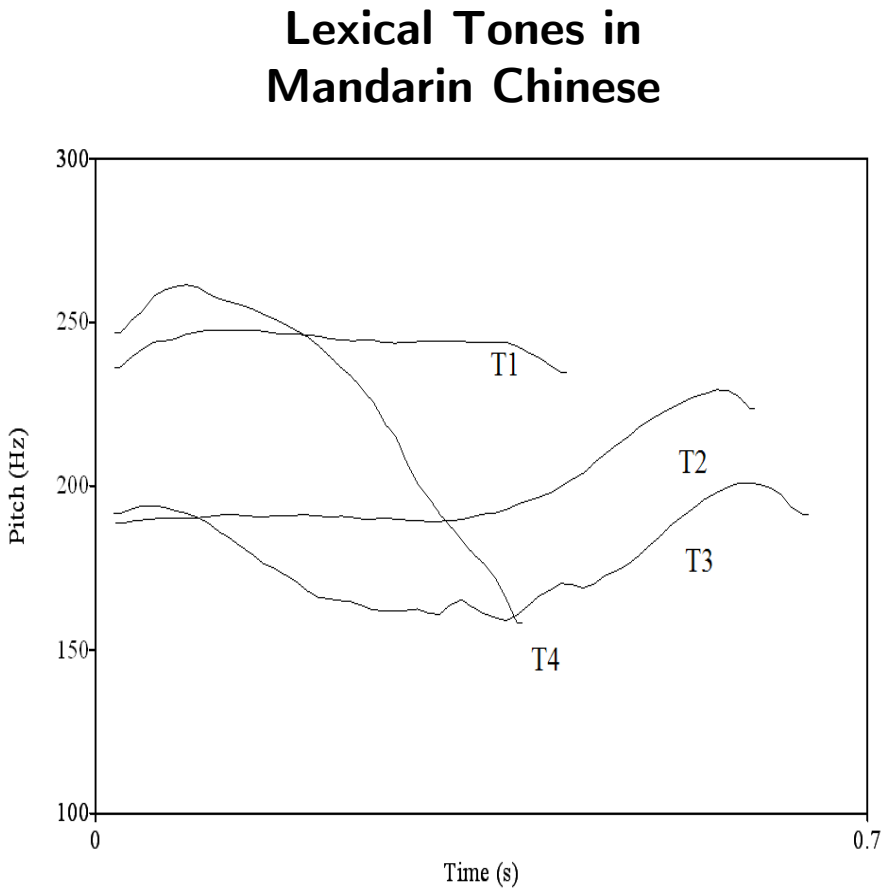
Types of Analyses

- Ultimately, the research questions and the data patterns should dictate the type of analysis used
- Fancier analyses are not necessarily better (they can be more difficult to interpret, which can make it more difficult for researchers to sell a particular story)

4. Practice

Brainstorming Session

- You are interested in testing the effect of **early pitch height** on lexical competition in Mandarin Chinese
  - **Target word**: T1 words
  - **Competitor words**: T2 words and T4 words
- How would you set up such a study?
- What issues would you need to worry about?



References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Altmann, G. T. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psycholica, 137*, 190-200.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

Barr, D. J., Gann, T. M., & Pierce, R. S. (2011). Anticipatory baseline effects and information integration in visual world studies. *Acta Psychologica, 137*, 201-207.

Cooper, R. 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perceptton, memory and language processing. *Cognitive Psychology, 6*, 84-107.

Dahan, D., & Tanenhaus, M. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review, 12*, 453-459.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*, 460-482.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica, 137*, 151-171.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language, 58*, 541-573.

References

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71-102.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: Taylor & Francis.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*, 475-494.

Pyykkönen-Klauck, P., & Crocker, M. W. (2016). Attention and eye movement metrics in visual world eye tracking. In P. Knoeferle, P. Pyykkönen-Klauck, P., M. W. Crocker (Eds.), *Visually situated language comprehension* (pp. 67-82).

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psycholica, 137*, 172-180.

Snodgrass, J. G., & and Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174-215.

Tanenhaus, M., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Traxler, M. J. (2012). *Introduction to psycholinguistics: Understanding language science*. Malden, MA: Blackwell.

Tremblay, A., Broersma, M., Coughlin, C. E., & Choi, J. (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology, 7*. http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00985/full