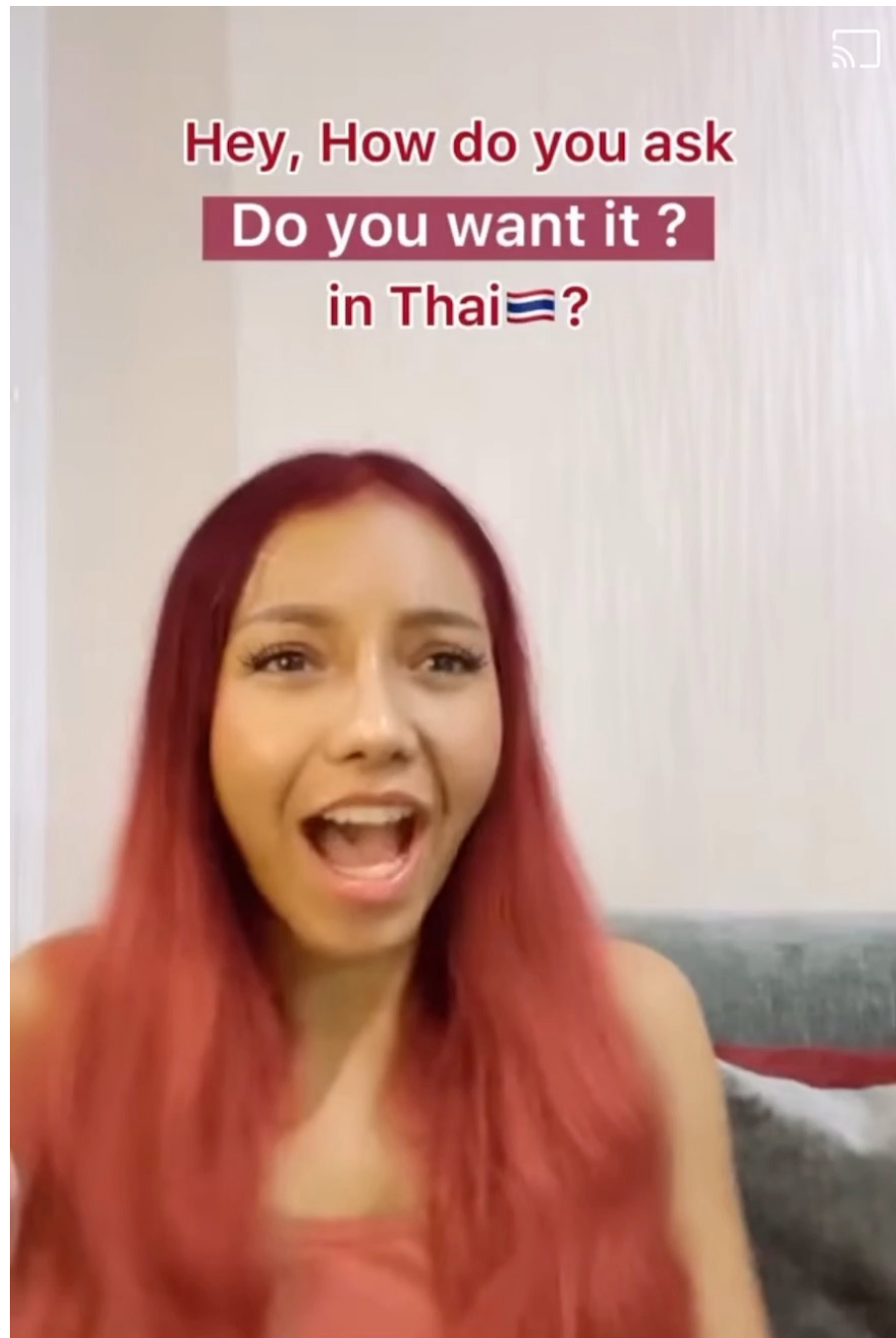


Lecture 7:

Speech Perception & Word Recognition

COGS 153

Speech Perception



Thai is a tonal language, it uses tones to create meaningful contrasts! (i.e. differentiate between words)

*Sorry, I can't find credit for the creator 😞

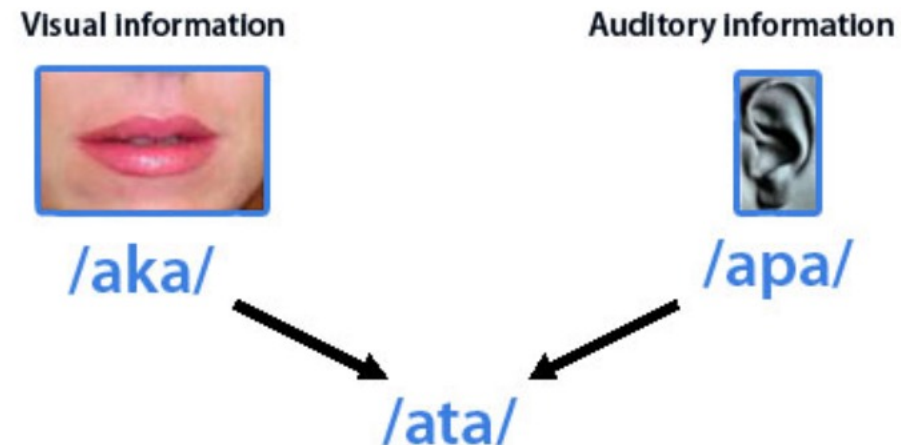
Challenges of Speech Perception

- *Lack of Invariance* problem.... speech sounds can sound different:
 - Depending on the other sounds surrounding it
 - e.g., hamster is actually pronounced “hampster”
 - When uttered by different speakers
- Speech Segmentation – speech stream is continuous and yet we are able to distinguish word boundaries
- What helps listeners disambiguate between sounds?
 - Visual information
 - Context
 - Attention



Vision Helps Disambiguate

- **McGurk effect:** a perceptual phenomenon where speech sounds are often miscategorized when the auditory cues in the stimulus conflict with the visual cues from the speaker's face
 - i.e., if we hear an audio recording of one sound & watch a video recording of someone making a different sound, we experience a blend of the two input streams
- Last video example:
 - Visual: ga ga ga
 - Auditory: ba ba ba
 - Combined: da da da



Context

- Context is the problem!
 - Speech sounds sound different depending on other neighboring speech sounds
- ... context is the solution!
 - Knowing the way sounds can be combined in your language and knowing words helps to disambiguate phonemes
- e.g., If you obscure one sound with a cough, people “hear” it anyway & what they hear depends on context
 - It was found that the _eel was on the axle → wheel
 - It was found that the _eel was on the shoe → heel
 - It was found that the _eel was on the orange → peel
 - It was found that the _eel was on the table → meal


Phonemic Restoration Effect


- **Phonemic restoration effect:** a perceptual phenomenon where under certain conditions, sounds actually missing from a speech signal can be restored by the brain and may appear to be heard.
- The effect occurs when missing speech sounds in an auditory signal are replaced with a noise
 - The brain fills in the absent speech sounds
- Example 1: Can you tell which sound is missing?
- Example 2: Did it get easier to understand?



Ganong Effect

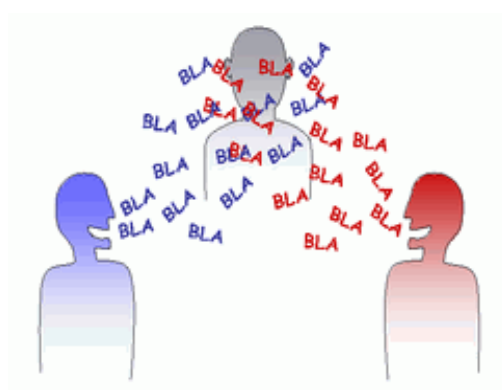
- **Ganong effect:** a perceptual phenomenon where a sound that is ambiguous between two speech sounds is perceived differently depending on the word that contains it

- Example 1: 

- Example 2: 

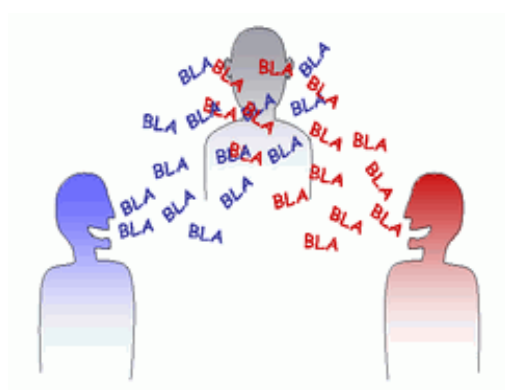
- What words did you hear?
- The first sound is the same in both words!

Attention modulates perception



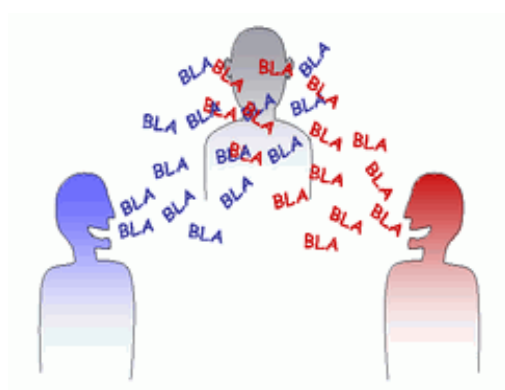
- In a crowded room with a lot of simultaneous conversation, we are usually able to filter out the background chatter and focus our attention on a single conversation... but if someone says your name in the background, it is consciously perceived and attention is refocused

Attention modulates perception



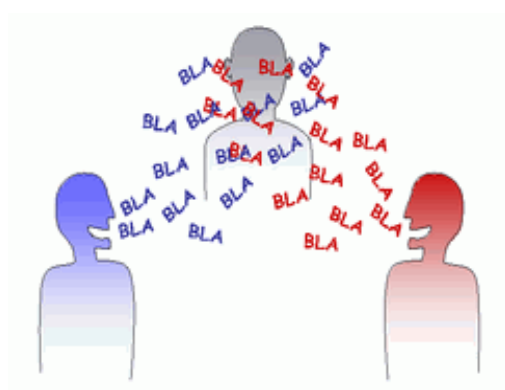
- In a crowded room with a lot of simultaneous conversation, we are usually able to filter out the background chatter and focus our attention on a single conversation... but if someone says your name in the background, it is consciously perceived and attention is refocused
 - This ability is known as the **cocktail party effect** (Cherry, 1953)

Attention modulates perception



- In a crowded room with a lot of simultaneous conversation, we are usually able to filter out the background chatter and focus our attention on a single conversation... but if someone says your name in the background, it is consciously perceived and attention is refocused
 - This ability is known as the ***cocktail party effect*** (Cherry, 1953)
 - We can **selectively attend to a single stream of speech even when hearing more than one**

Attention modulates perception



- In a crowded room with a lot of simultaneous conversation, we are usually able to filter out the background chatter and focus our attention on a single conversation... but if someone says your name in the background, it is consciously perceived and attention is refocused
 - This ability is known as the ***cocktail party effect*** (Cherry, 1953)
 - We can **selectively attend to a single stream of speech even when hearing more than one**
- Listeners are also aware of the likelihood of certain sounds occurring in certain contexts because of their experience (Ballas & Mullins, 1991)
 - e.g., a cow moo-ing would be much more salient in an office than in a barnyard

Word Recognition

dollar

How do we recognize words?

- **Lexical access:** the process by which we produce a specific word from our mind / recognize it when see or hear it

This lecture:

- Models of lexical access
 - Cohort
 - TRACE
- Lexical ambiguity and some empirical evidence for how word representations may be activated

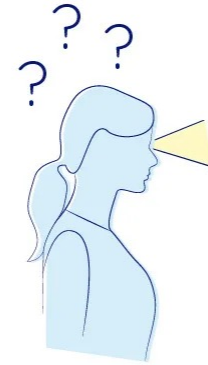
Bottom-up vs. Top-down processing

- Bottom-up

- Raw sensory data taken from the world (e.g., speech sounds)
- “stimulus-driven” processing

- Top-down

- Previous knowledge, models, ideas, and expectations applied to sensory data
- “context-driven” processing



What am I seeing?

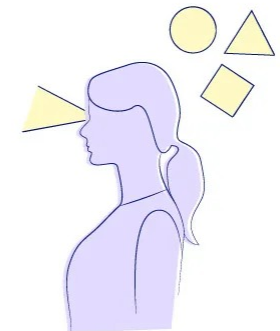
Bottom-up processing:

taking sensory information and then assembling and integrating it

Top-up processing:

using models, ideas, and expectations to interpret sensory information

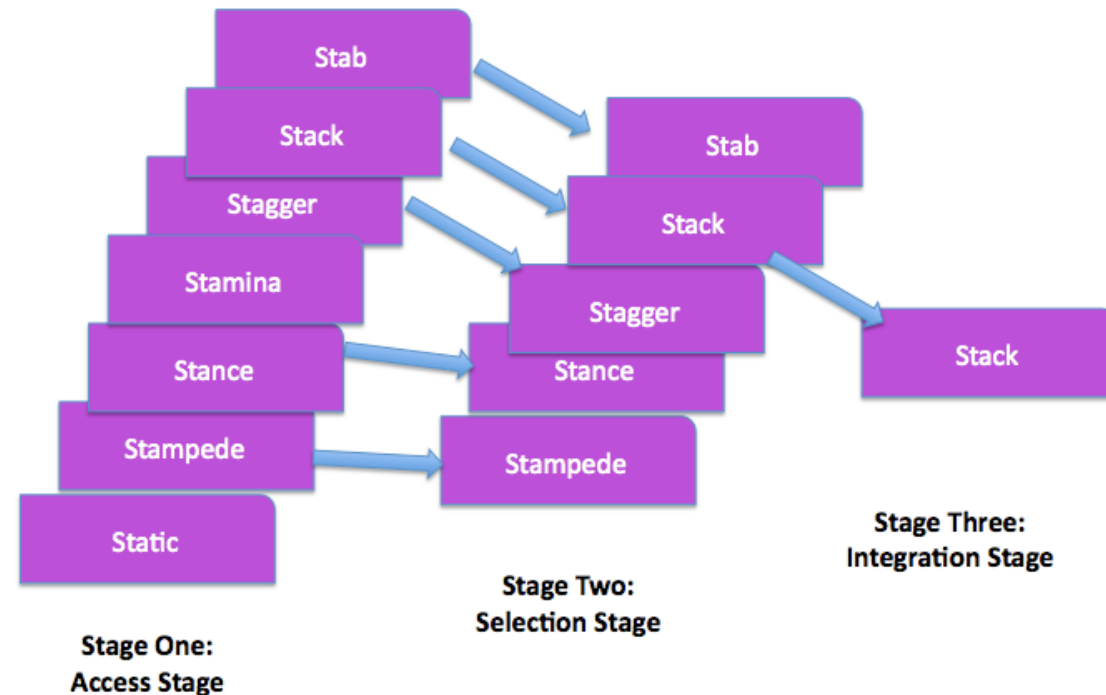
Is that something I've seen before?



Jack Westin

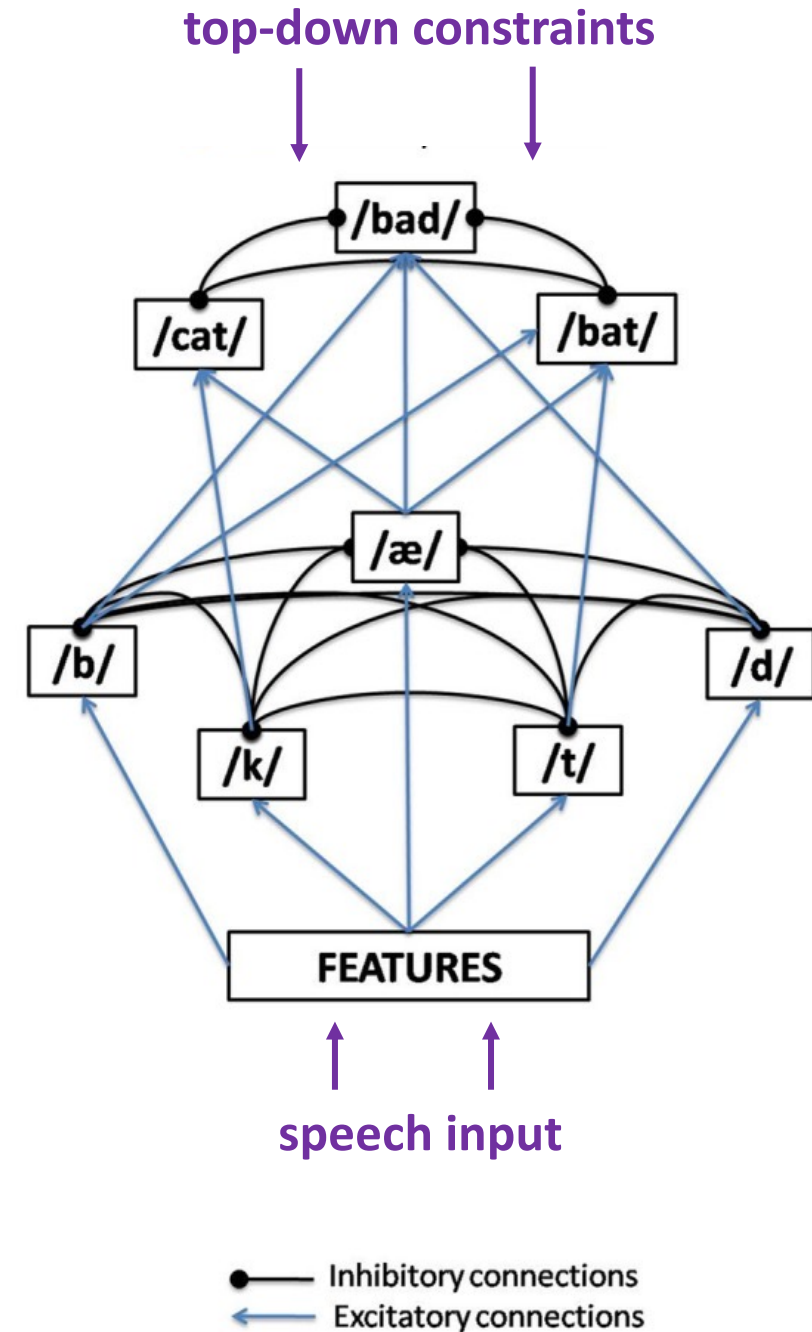
Cohort model

- 3 stages: *access*, *selection* (pre-lexical), and *integration* (post-lexical)
 - Key idea: processing of the word starts at the first sound + context is used later
- **1. ACCESS:**
 - Access to potential 'candidates' of 'cohorts' is based on the speech signal, which will then will activate all words that begin with the same sounds.
- **2. SELECTION:**
 - During the selection stage, one 'cohort' of words becomes activated.
 - Continuing incoming speech signal will cause mismatches between all the words in the cohort.
 - Mismatches are then eliminated sequentially until only the target word remains.
 - This is referred to as the ***recognition point / uniqueness point***
- **Note:** only sensory input is used in selecting one candidate from the generated cohort
 - Bottom-up processing
 - NO CONTEXT used in first two stages, i.e. no top-down interference
- **3. INTEGRATION:**
 - Context is finally accessed
 - Now word properties are used to integrate the selected word into the context of the phrase in which it is being used



TRACE model

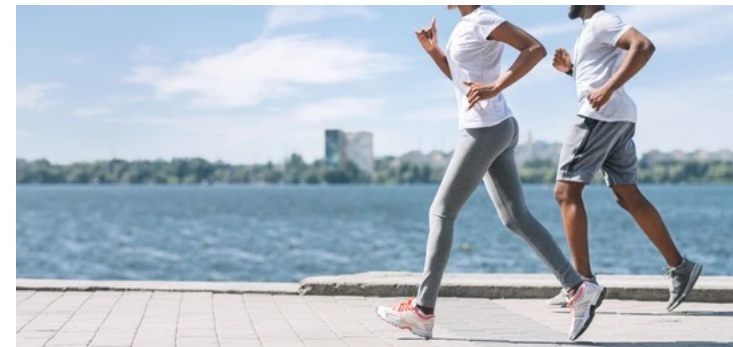
- Interactive network model of speech perception
 - Idea: perception occurs through a network of neurons in which different levels of speech units (e.g., features, speech sounds, words) are represented at different levels + activation can move between layers (bidirectional)
- Input level
 - Features of speech sounds are accessed + activated
 - Speech sounds are recognized, and an excited speech sound will then excite the 'word units' it has connections to
 - e.g., /b/ will activate the words /bad/ and /bat/
 - Excitatory activation across levels + inhibitory effect in each level
 - i.e., at the speech sound level, the activation of one sound inhibits the activation of other competing sounds (activated /b/ will inhibit /k/)
- TRACE can 'recover' if elements of the speech signal are missing
 - **does not rely on the initial sounds** of the input (unlike Cohort!)
 - bidirectional connections can use **top-down processing**
 - → integrates context to facilitate speech perception



What happens in the case of
lexical ambiguity?

Lexical ambiguity

- **Lexical ambiguity:** two or more possible meanings for a single word
- Homonyms
 - Two or more *unrelated* words that sound and/or are spelled the same
 - Homographs: spelled the same but mean different things
 - Bass, bass
 - Homophones: pronounced the same but mean different things
 - Knight, night
- Polysemous words and phrases
 - Individual words or phrases that have multiple *related* meanings or senses
 - She has a run in her stockings.
 - There was a run on the banks this week.
 - She went out for a morning run.
 - You can run your fingers through my hair.
 - Let's run through various options.
 - He had a run of bad luck.

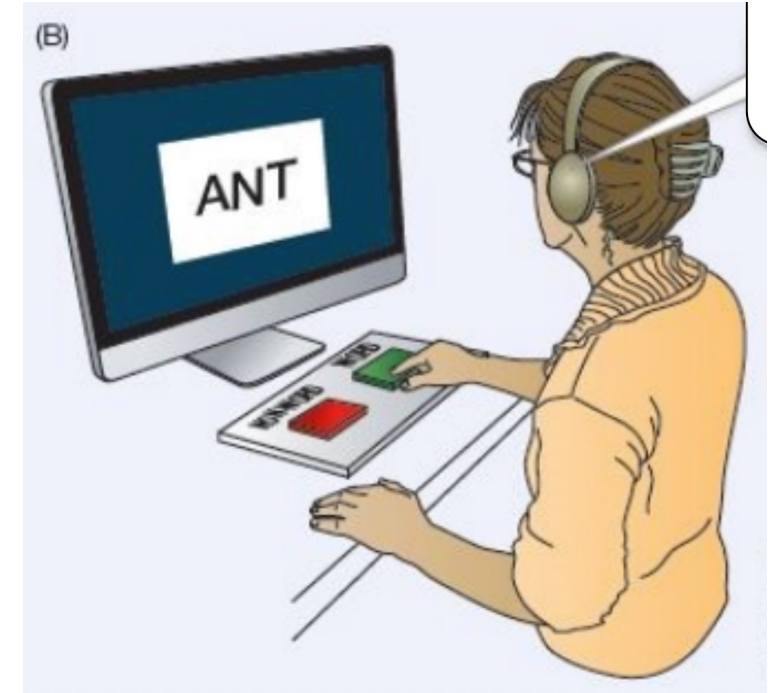


Some questions we can ask

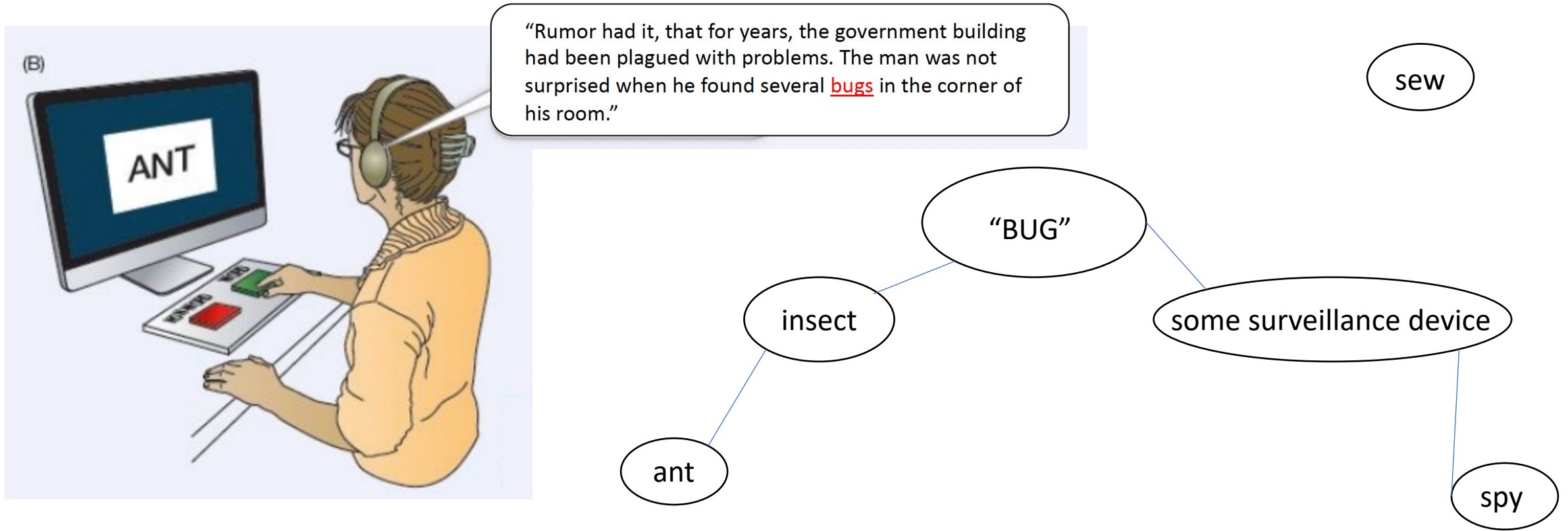
- How is lexical ambiguity resolved?
- Can studying ambiguity help us understand lexical access?
 - Is the process top-down or bottom-up, or both? What is the role of context?
 - Bottom-up model
 - Multiple meanings activated until context is recruited to resolve ambiguity
 - Top-down model
 - Context generates expectations (predictions) and 'pre-activates' particular meanings

Do we activate *simultaneous* word meanings?

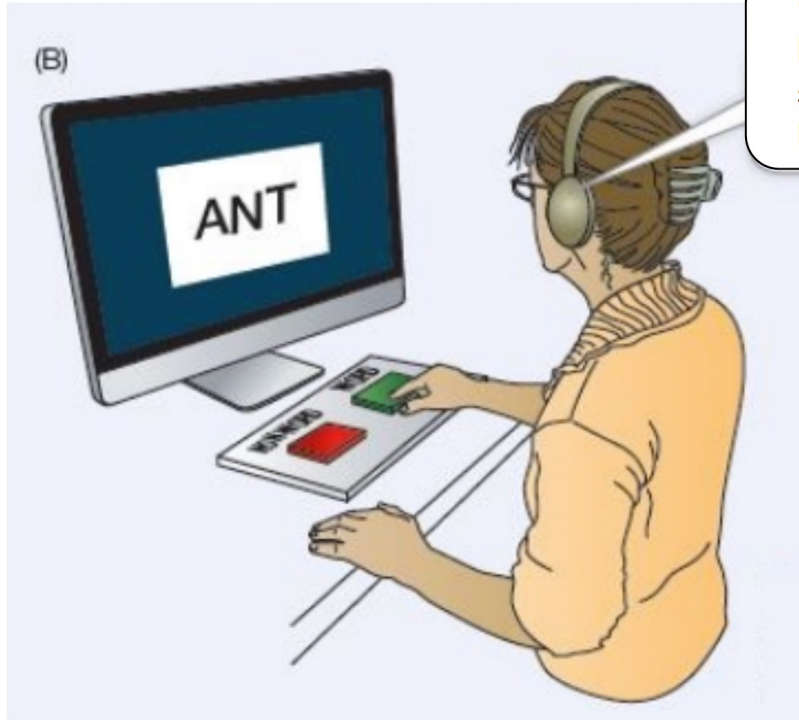
- **Method:** cross-modal (visual/audio) priming task + lexical decision task
- **Procedure:**
 - Participant listens to sentences with a priming word
 - The prime is either an ambiguous or unambiguous word
 - The sentences provide either neutral context or biasing context for the prime
 - Participant is shown a visual target (word) on the screen and asked to press a key to indicate if it's a word or nonword
 - Words can either be relevant, irrelevant, or unrelated to prime
 - Nonwords = not real words in English
 - **Dependent variable:** response time to push button



Condition 1: Neutral context + ambiguous prime

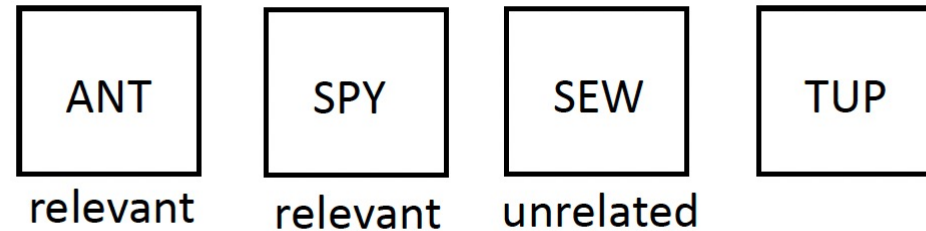


Condition 1: Neutral context + ambiguous prime



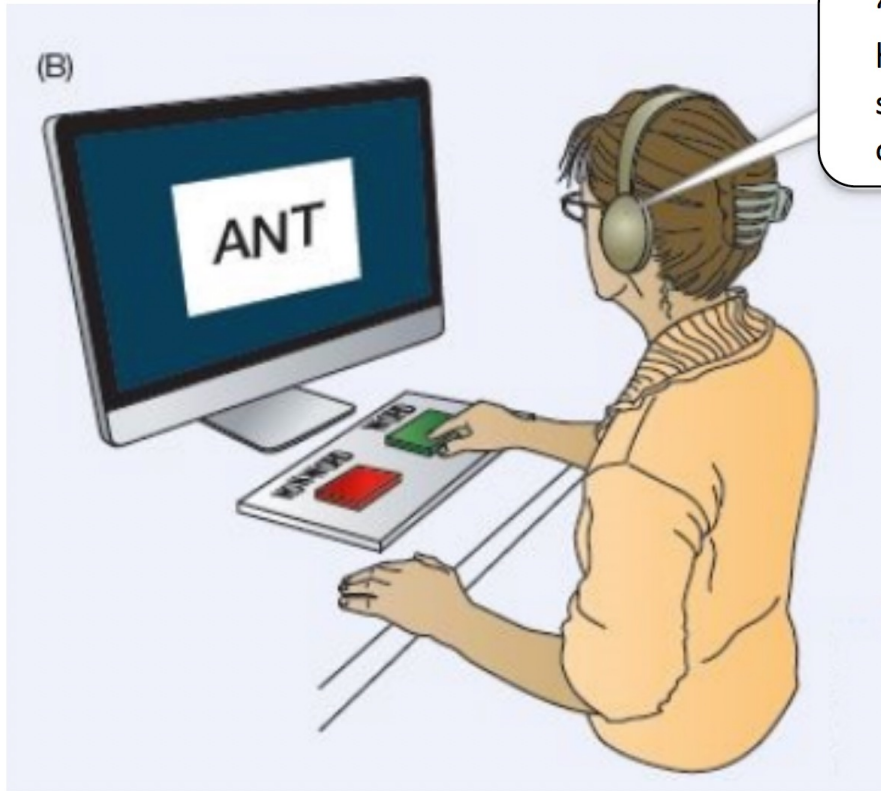
"Rumor had it, that for years, the government building had been plagued with problems. The man was not surprised when he found several bugs in the corner of his room."

- **Audio Context:** Neutral
- **Prime:** ambiguous ("bugs")
- **Visual target** immediately after prime:



- **Results:**
 - Similar response times for both relevant visual targets (ANT and SPY)
 - Faster response times to relevant visual targets compared to unrelated word
 - RTs: relevant = relevant < unrelated

Condition 2: Neutral context + unambiguous prime



“Rumor had it, that for years, the government building had been plagued with problems. The man was not surprised when he found several insects in the corner of his room.”

- **Audio Context:** Neutral
- **Prime:** unambiguous (“insects”)
- **Visual target** immediately after prime:

ANT

relevant

SPY

irrelevant

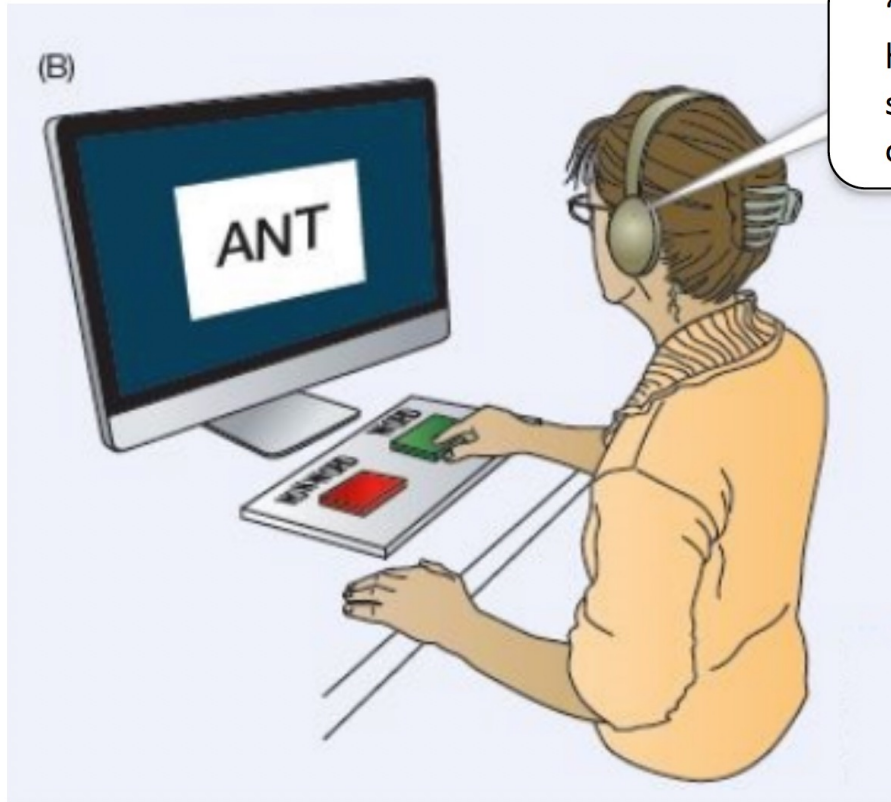
SEW

unrelated

TUP

- **Results:**
 - Similar response times for both irrelevant and unrelated visual targets (SPY and SEW)
 - Faster response times to relevant visual targets compared to irrelevant and unrelated
 - RTs: relevant < irrelevant = unrelated

Condition 3: Biasing context + ambiguous prime



“Rumor had it, that for years, the government building had been plagued with problems. The man was not surprised when he found several spiders, roaches, and other bugs in the corner of his room.”

- **Audio Context:** Biasing
- **Target word:** ambiguous (“bugs”)
- **Visual target** immediately after target word:

ANT

relevant

SPY

irrelevant

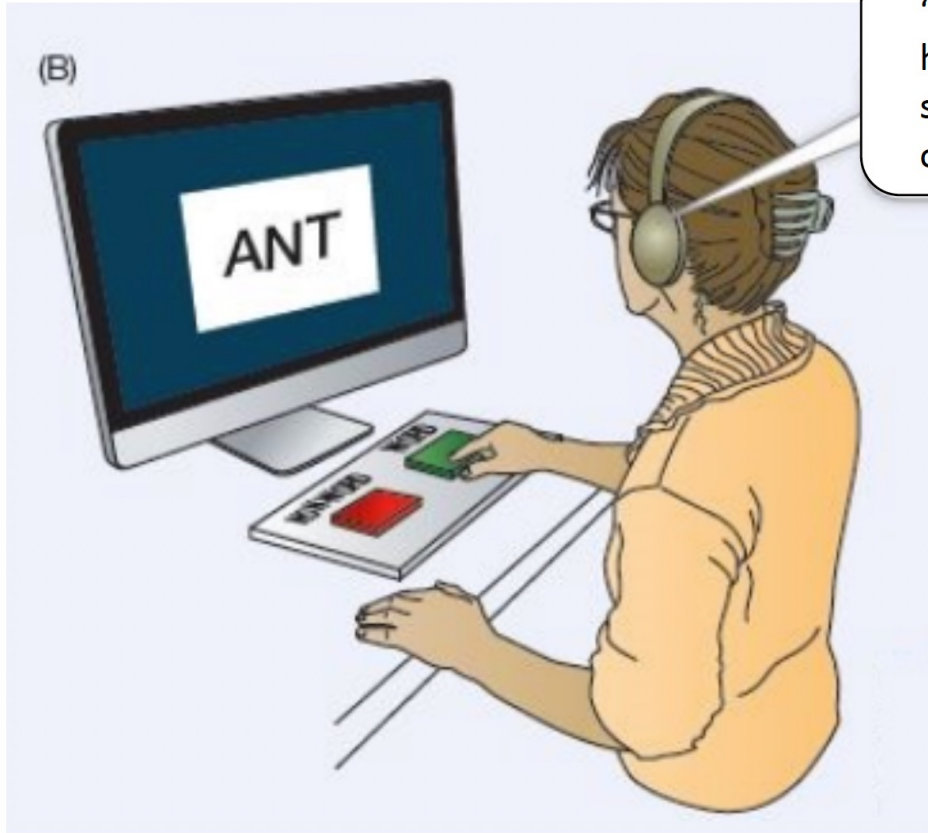
SEW

unrelated

TUP

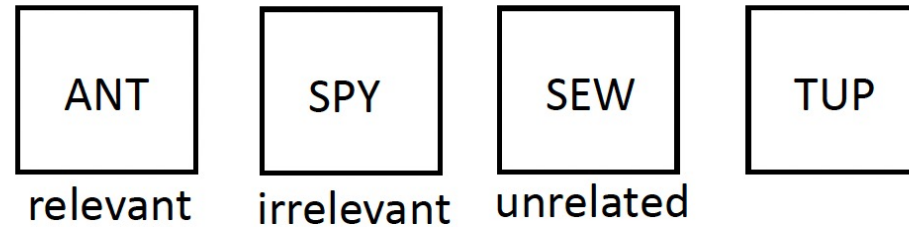
- **Results:**
 - Similar response times for both relevant and irrelevant visual targets (ANT and SPY)
 - Faster response times to relevant and irrelevant visual targets compared to unrelated word
 - RTs: relevant = irrelevant < unrelated

Condition 4: Biasing context + unambiguous prime



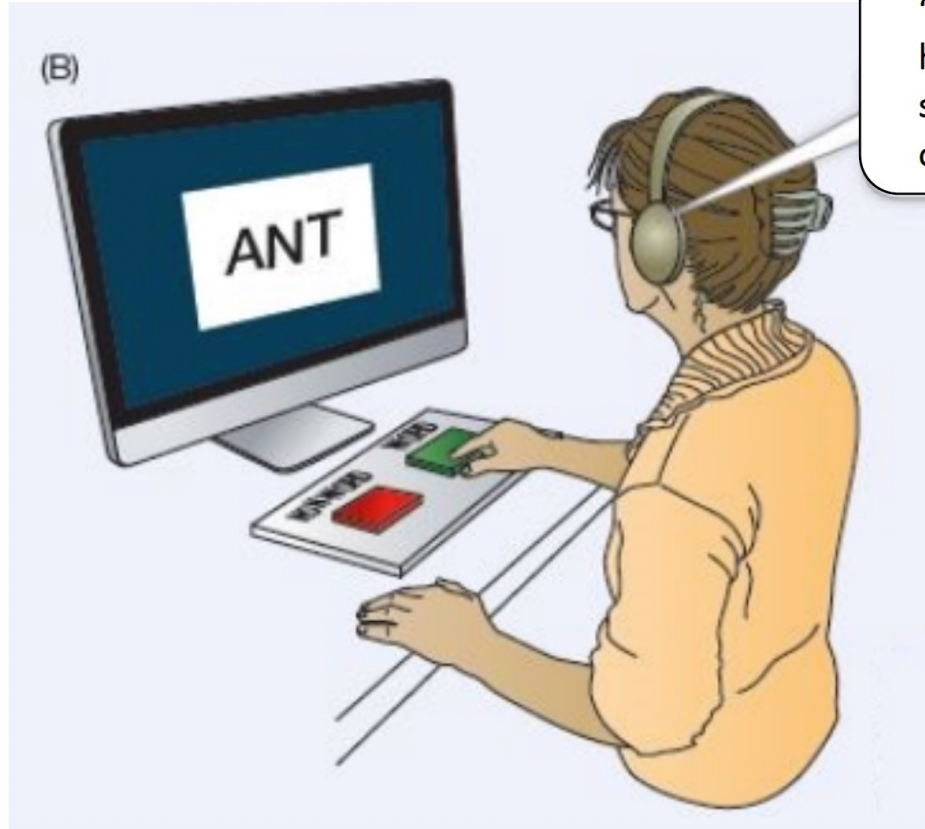
"Rumor had it, that for years, the government building had been plagued with problems. The man was not surprised when he found several spiders, roaches, and other insects in the corner of his room."

- **Audio Context:** Biasing
- **Prime:** unambiguous ("insects")
- **Visual target** immediately after prime:



- **Results:**
 - Similar response times for both unrelated and irrelevant visual targets (SEW and SPY)
 - Faster response times to relevant visual targets compared to irrelevant and unrelated words
 - RTs: relevant < irrelevant = unrelated

Experiment 2: Does timing matter?



“Rumor had it, that for years, the government building had been plagued with problems. The man was not surprised when he found several spiders, roaches, and other **bugs** in the corner of his room”

- **Audio Context:** Biasing
- **Prime:** ambiguous (“bugs”)
- **Visual targets** *three syllables after prime* (“cor”):

ANT

relevant

SPY

irrelevant

SEW

unrelated

TUP

- **Results:**
 - Similar response times for both unrelated and irrelevant visual targets (SEW and SPY)
 - Faster response times to relevant visual targets compared to irrelevant and unrelated words
 - RTs: relevant < irrelevant = unrelated

Ambiguity: do we activate *simultaneous* word meanings?

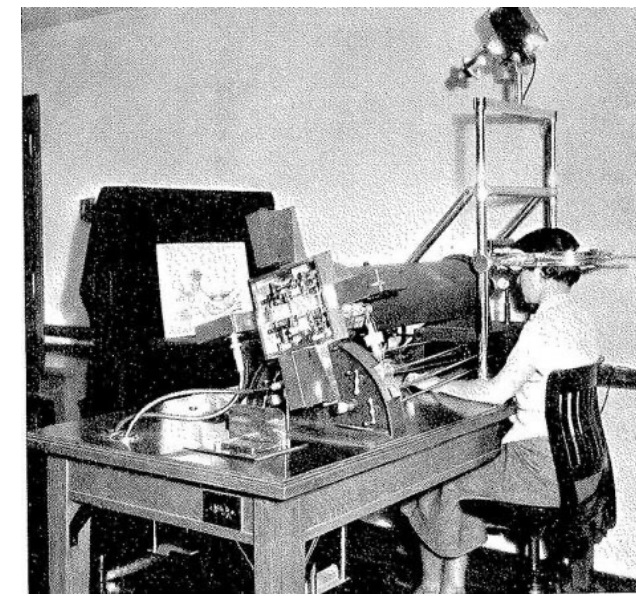
- When the visual target was presented immediately after hearing the the prime, the responses times were similar for the relevant and irrelevant targets (and faster than the unrelated target)
 - → BOTH MEANINGS of the word are activated
- But when the visual target was presented three syllabus after the prime, only the relevant target showed sped up response times (faster than both the irrelevant and unrelated targets)
 - → irrelevant meaning was activated in parallel with the intended meaning, but was quickly suppressed
- Are irrelevant meanings always getting activated?

Frequency of meaning in ambiguous words

- Some ambiguous words have roughly equal distributions of meaning
 - Bark (dog \approx tree)
 - Pitcher (baseball \approx drink)
 - Straw (hay \approx drink)
 - Chest (storage \approx body)
- Other ambiguous words differ in their frequency of meaning
 - Port (ships $>$ alcohol)
 - Cabinet (furniture $>$ council of advisors)
 - Yarn (thread $>$ story)
 - Mint (candy $>$ (herb) $>$ building)

Does the *frequency of meaning* influence simultaneous activation?

- **Method:** eye-tracking paradigm
- **Procedure:**
 - Participants read sentences with a word that could vary in 3 ways:
 - Unequal Frequency Ambiguous Target Word (e.g., mint: a place where money is coined < peppermint candy)
 - Equal Frequency Ambiguous Target Word (e.g., pitcher: baseball ≈ drink)
 - Unambiguous Control Word (e.g., jail)
 - The context of the sentence supports one meaning of an ambiguous word over another
 - Dependent Variable: Looking time at target word/reading time
- **Results:**
 - When the context favored less frequent meanings of ambiguous words, participants read SLOWER (compared to unambiguous control words in the same sentence)
 - e.g., slower when sentence says MINT compared to when it says JAIL
 - → the slowdown suggests competition from alternative meaning (peppermint candy)
 - For equal frequency ambiguous words in sentences with context that favored one of the meanings, there was NO DIFFERENCE in reading times between ambiguous target word and unambiguous control word
 - e.g., same reading time for pitcher vs whiskey
 - → suggests no competition from alternative meaning (baseball pitcher)
- The frequency of meaning can influence simultaneous activation



Unequal Frequency:

“Although it was by far the largest building in town, the (**mint**/**jail**) was seldom mentioned.”

Equal Frequency:

“Because it was kept on the back of a high shelf, the (**pitcher**/**whiskey**) was often forgotten.”

Summary of ideas

- Both frequency of meanings and contextual expectations can affect the activation levels of word representations
 - When both frequency + context factors activate the same meaning, that meaning representation becomes very activated
 - → leading to quickly inhibiting the other competitors!
 - When frequency and context factors mismatch (frequency activating one and context favoring the other), maybe both meanings are equally activated
 - → leading to competition between them!