Using free-view eye-tracking to study spoken language processing

Instructors: (LSA 363)
  Michael K. Tanenhaus
  (mktanenhaus@gmail.com)*
  John C. Trueswell

Graduate Assistants:
  Katie Carbary
  Austin Frank
  Natalie Klein

Course Website:
  http://www.ircs.upenn.edu/igert/LSA363
Overview*

- We review some of the “visual world” literature with an eye towards:
  - highlighting how visual world paradigm can be used to address issues in:
    - phonetics
    - spoken word recognition
    - parsing,
    - reference resolution
    - production
    - experimental pragmatics and interactive conversation
  - considering methodological issues that arise when using eye movements to examine spoken language processing
  - providing students with some hands-on experience designing visual world experiments

* all bad puns should be attributed to JCT unless otherwise noted
## Lectures

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic and lecturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/5</td>
<td>intro, eye movements (JCT, MKT)</td>
</tr>
<tr>
<td>7/9</td>
<td>speech, words, pitch accents (MKT)</td>
</tr>
<tr>
<td>7/12</td>
<td>sentence processing, expectation (JCT)</td>
</tr>
<tr>
<td>7/16</td>
<td>referential domains, implicature (MKT)</td>
</tr>
<tr>
<td>7/19</td>
<td>parsing to learn and vice versa (JCT)</td>
</tr>
<tr>
<td>7/23</td>
<td>production, perspective-taking (JCT, MKT)</td>
</tr>
<tr>
<td>7/25</td>
<td>conversation, future directions (MKT, JCT)</td>
</tr>
</tbody>
</table>
Labs and Discussion

• Times to be arranged
  – Tuesday and Wednesday (We need three 2 hour blocks)*
  – Discussion and demos with head-mounted ISCAN, TOBII portable

• Labs:
  Lab 1: Referential ambiguity and point of disambiguation (T, W July 10-11th)
  Lab 2: Speaker eye gaze in verb learning and parsing (T, W July 17-18th)
  Lab 3: Common ground in spoken conversation (T, W July 24-25th)

* Each student attends one section each week (your assigned section)
Why use eye movements to study spoken language in a “Visual World”? 
Why spoken language?

- Talking and listening are primary
  - All societies have spoken language
  - All children learn to converse, initially by talking about the world.

- Other forms of language are derivative
  - Most languages don’t have writing systems
Ignoring speech can encourage scientific balkanization, making it easier to ignore duality of patterning:

speech perception--> **word recognition**--->syntactic processing--> sentence interpretation--> **discourse** representation

**Duration:**

*cap/captain* (in strong position, vowel in *cap* is longer)
But information structure of discourse can affect duration.

*Put the cap above the captain. Now put the cap/captain…*
*Put the CAPtain*

Moral: “high” and “low” level subsystems can share the same input data
Why use saccadic eye movements to study spoken language?

• Consider what spoken language is like from two perspectives
  – Language unfolding over time as a sequence of transient acoustic events.
  – Conversation as a joint activity between two or more people.
The unfolding signal

1. Language unfolds over time.

2. Processing is closely time-locked to the unfolding utterance.

3. Requires monitoring moment-by-moment comprehension with careful control of the signal.

Put the apple on the towel...

beetle, beacon, beak, beep...
Listeners are NOT sensitive to small VOT differences.

- Sharp identification of speech sounds on a continuum
  - ID (%/pa/)
  - Discrimination poor within a phonetic category

Example 1: Categorical Perception
Example 2: Sentence processing

Paradigms have required de-contextualized language
e.g., cross-modal priming

spy

The lawyer represented the doctor who testified that the bug frightened him
Interactive conversation
Gibberish?

1  *ok, ok I got it* ele...ok
2  alright, *hold on*, I got another easy piece
1  *I got a* well wait I got a green piece RIGHT above that
2  above this piece?
1  well not exactly right above it
2  it can’t be above it
1  it’s to the...it’ doesn’t wanna fit in with the cardboard
2  it’s to the right, right?
1  yup
2  w- how? *where*
1  *it’s* kinda line up with the two holes
2  line ‘em right next to each other?
1  yeah, vertically
2  vertically, meaning?
1  up and down
2  up and down

A curtain separates the subjects.

Eye-tracker

Resource areas
Participants replace stickers / with blocks 
This is mid-way through the task:

Subject 1’s Board

Subject 2’s Board
1: you’re gonna use your other yellow block
2: long one?
1: yep
2: ok.... 8 lines over to the left
1: yep
Interactive conversation
Gibberish?

1  *ok, ok I got it* **ele...ok**
2  alright, *hold on*, I got another easy piece
1  *I got a* well wait I got a green piece RIGHT above that
2  above this piece?
1  well not exactly right above it
2  **it can’t be above it**
1  it’s to the…it’ doesn’t wanna fit in with the cardboard
2  it’s to the right, right?
1  yup
2  **w- how? *where***
1  *it’s* kinda line up with the two holes
2  line ‘em right next to each other?
1  yeah, vertically
2  vertically, meaning?
1  **up and down**
2  **up and down**
Spoken Language in a Different Tradition

Prototypical experiment requires participants to interact in goal-driven task

Director

Matcher

Take the shape with thee, uh, the circle above the triangle and, uh.. Yeah, the dancer

You mean the one that looks like a dancer with a fat leg
Reasons for separate traditions

**Methodological:**
It has been difficult to study language in “natural” contexts with precision necessary to examine time-course.
Why eye movements?

- ballistic measure
- can be used with continuous speech
- natural response measure
- low threshold response
- subject is unaware
- does not require meta-linguistic judgment
- closely time-locked to speech
- plausible linking hypothesis:

  Probability of eye movement at time (t) is a function of activation of possible alternatives plus some delay for programming and execution (~200ms)
Eye movements allow us to study spoken language in rich contexts with precision necessary to examine time-course.

Is this a good thing?

No: (confuses language with non-linguistic stuff)

We should be studying how we construct/generate linguistic representations. In visual world, we introduce task-specific strategies, etc.

Yes:

There is always a context, it always matters; understanding how the system works in a rich environment can be more informative about basic principles than studying the system in an impoverished environment (analogy with vision; bottom-up or top-down).

We can manipulate action/goals/language/display
Cards Video
Allopenna, Magnuson & Tanenhaus (1998)

Eye camera

Scene camera

Pick up the beaker
Target = beaker
Cohort = beetle
Unrelated = carriage

Look at the cross. Click on the beaker.
Fixation Proportions summed over an interval

Trials

1
2
3
4
5

200 ms

Proportion of fixations

Time

Proportion of fixations

cohort unrelated
Fixation time over an interval

Trials

1
2
3
4
5

Proportion of fixations

Time spent fixating (ms)

cohort
unrelated
Trials

Fixation time over an interval

Proportion of fixations generated

Time

Proportion of fixations

1
2
3
4
5

Trials

cohort
unrelated
Allopenna et al. results

- Target, e.g. beaker
- Cohort, e.g. beetle
- Rhyme, e.g. speaker
- Unrelated, e.g. carriage

**Average target offset**

200 ms after coarticulatory information in the vowel
Activation converted to probabilities using the Luce (1959) choice rule

\[ S_i = e^{k a_i} \]

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

\[ L_i = \frac{S_i}{\sum S_i} \]

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:

\[ d_i = \frac{\text{max}_act_i}{\text{max}_act_{overall}} \]

\[ R_i = d_i L_i \]

Linking hypothesis

- Activation converted to probabilities using the Luce (1959) choice rule
- \( S_i = e^{k a_i} \)
- \( S \): response strength for each item
- \( a \): activation
- \( k \): free parameter, determines amount of separation between activation levels (set to 7)

\[ L_i = \frac{S_i}{\sum S_i} \]

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:

\[ d_i = \frac{\text{max}_act_i}{\text{max}_act_{overall}} \]

\[ R_i = d_i L_i \]
Put the apple on the towel in the box

Put the apple on the towel in the box.
Put the apple on the towel in the box.

Put the apple that’s on the towel in the box.
Other measures/terms

Contingent analyses:

response contingent:

speech (McMurray et al., 2002, Cognition)
reference (Runner et al., 2003, Cognition)

look contingent:

time to target from picture X after point in the signal
(Dahan & Gaskell, in press, JML)

Point of disambiguation (POD):

Eberhard et al., 1995, JPR,

Anticipatory eye movements:

Altmann & Kamide, 1999, Cognition
The boy ate the cake
* Assumes listeners will immediately interpret prenominal adjectives such as empty contrastively (i.e., there should be one empty X and one not empty X). See Sedivy et al. (1999) *Cognition* for supporting evidence.
Late Match

Early Match

Pick up the empty martini glass ...

Time since onset of determiner (ms)

Fixation Probability

Target
TargetO
Comp
CompO
Other
Late Match

Pick up the empty martini glass ...

Early Match

Pick up the empty martini glass ...
Late Match

Pick up the empty martini glass ...

Early Match

Pick up the empty martini glass ...

Time since onset of determiner (ms)

Fixation Probability

- Target
- TargetO
- Comp
- CompO
- Other