A new generation of data:
Crowdsourcing and language studies

Tyler Schnoebelein, Robert Munro, Steven Bethard, David Clausen, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, and Harry Tily
What’s this presentation about?

- Chairs
What’s this presentation about?

- Chairs
- Giving you new chairs
What’s this presentation about?

• Chairs
• Giving you new chairs
• Giving you fuller chairs
Transparency of phrasal verbs

Segmentation of an audio speech stream

Contextual predictability

Judgment studies of fine-grained probabilistic grammatical knowledge

Post-hoc metaphorical frequency analysis of electrophysiological responses

Filler informativity and island constraints

BONUS! Screening for linguistic attentiveness

EXTRA BONUS IF YOU STAY!
Best practices!!!!
Crowdsourcing

• First use of crowdsourcing:
  – Get everyday folks to annotate and translate sentences
  – (These are still great uses for crowdsourcing)

• Other uses:
  – Generating targeted corpora
  – Checking linguistic intuitions
  – Learning about language ideology

• Focus here today:
  – Crowdsourcing offers faster and cheaper results with broader demographics than most lab studies
  – But are the results reliable?
Amazon Mechanical Turk

- Launched at the end of 2005
- Originally for “artificial artificial intelligence”
  - Stuff that’s easy for people to do and SHOULD be easy for computers to do...but isn’t
  - “What’s the focus of this photograph?”
- Lots of alternative services to consider:
  - CrowdFlower is especially geared towards academics
  - clickworker.com may be useful in Europe
  - See also http://www.crowdsortium.org/ and http://crowdsourcing.org
(1) Transparency of phrasal verbs

- **Transparent** phrasal verbs are easy to figure out: *lift up* entails something being lifted and going up.
- **Opaque** phrasal verbs are more confusing: when you *give up* what do you give? What goes up?
- 3 Turk experiments, 2 undergrad experiments
  - 215 participants
  - 96 phrasal verbs
  - With and without context (examples of phrasal verbs in sentences)

On a scale of 1–7, how similar are ‘cool’ and ‘cool down’?
Super-cheap and fast

- The more you pay, the faster results will come in. Even if you’re a cheapskate:
  - Most studies have results done in under a week, usually just a couple days
- To get 20 responses for each of 96 phrasal verbs:
  - $11.60
- So obviously, if it’s reliable that’s terrific. But can we trust the results?
Other measures show reliability

<table>
<thead>
<tr>
<th></th>
<th>StudentContext and TurkContext</th>
<th>All tests but StudentLong</th>
<th>StudentLong</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC consistency</td>
<td>0.899</td>
<td>0.780</td>
<td>0.0934</td>
</tr>
<tr>
<td>ICC agreement</td>
<td>0.900</td>
<td>0.740</td>
<td>0.0854</td>
</tr>
</tbody>
</table>

- High intra-class correlation coefficient (ICC) values—except for the wayward students in StudentLong.
- Cohen’s kappa (weighted):
  - $\kappa = 0.700$, for all tests excluding StudentLong
  - $\kappa = 0.835$ for StudentContext and TurkContext
  - $\kappa > 0.8$ is “almost perfect agreement” (Landis and Koch 1977)
(2) Segmentation of an audio speech stream

- Lab experiments use more than just text
  - Adobe Flash: plugin enabling display of multimedia content in most users' browsers
- Custom written applets allow rich multimedia content in a Turk task
  - Developed using Adobe Flex (freely available, Javascript-like environment)
- Here: a classic speech segmentation task (after Saffran et al 1996)
Statistical speech segmentation

- Audio presentation of “words” from a fixed vocabulary (no pauses between words):
  tibibagukipaladazuzivatugi
  piduvatugitibibaguladazuzi
  ladazuzikipalukavutibibagu

After training, participants complete forced choice task

Click on the word that was in the language you heard:
  tibiba  kavuti
Results

- Lab subjects
  - 71% correct (n=12);
- Turkers
  - 66% (n=24)
- Turkers perform better significantly better than chance but are indistinguishable from lab subjects
(3) Contextual predictability

- To build models of sentence processing, linguists need to understand the effect of context.
- One way to do with is with a Cloze task, which is like Mad Libs.

Stimuli 1: Margie ___
Stimuli 2: Margie moved ___
...
Stimuli 5: Margie moved into her new ____
Data:
- 48 sentences, each w/ 10+ words

Participants:
- 488 on Turk
- 20 undergrads in Reichle et al’s lab at UMass

Example:
- *Margie moved into her new apartment* – 17/34 Turkers
  - 17/34 Turkers
  - 15/20 undergrads
Cloze predictability results

- Strong correlation between Turkers and students
  - Spearman’s rank correlation, rho=0.823 (p<0.0001)

- "Shared zeros" are those words in stimuli sentences that no one—in the lab or on Turk suggested.
  - To be extra conservative, we take them out of consideration.
  - The correlation is still terrific: rho=0.759 (p<0.0001)
<table>
<thead>
<tr>
<th></th>
<th>Lab predictability</th>
<th>Turk predictability</th>
<th>Surprisal</th>
<th>Word position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab predictability</td>
<td></td>
<td>0.759</td>
<td>0.087</td>
<td>0.346</td>
</tr>
<tr>
<td>Turk predictability</td>
<td>&lt;0.001</td>
<td></td>
<td>0.076</td>
<td>0.388</td>
</tr>
<tr>
<td>Surprisal</td>
<td>0.096</td>
<td>0.150</td>
<td></td>
<td>0.281</td>
</tr>
<tr>
<td>Word position</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>
(4) Judgment studies of fine-grained probabilistic grammatical knowledge

- Information-theoretic approach to syntax:
  - Greater probability of relative or complement clause (RC or CC) -> more likely to omit ‘that’

- In corpus-derived models of speaker production the most significant factor is predictability of embedded clause, given head noun (RC) or verb (CC) lemma

We \(_v\)\[\text{hope}\] \((\text{that})_{\text{COMPL}}\) you enjoy
\(\text{NP}\)\[\text{the talk}\] \(_i\) \((\text{that})_{\text{REL}}\) we give \(\_\_\_\_\_\_\_\) today
### Really quick background first:

1. What is your gender?  
   - Male  
   - Female

2. Was English your first language?  
   - Yes  
   - No

3. What is your age?

4. What country do you live in?

### Now the question:

*Here's a bit of actual recorded telephone conversation...*

S1: The longevity of the house is not, uh, is not worth it. How about in your case?
S2: Well, in my case my husband is not a carpenter, but in fact, he's in electronics...

*How likely is each of the following to have been the next sentence?  
Assign 100 points between the two boxes (80/20, 23/77, 51/49, etc.).*

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>...but he knew that the only way we'd ever have a new home is if he would build it.</td>
<td>...but he knew the only way we'd ever have a new home is if he would build it.</td>
</tr>
</tbody>
</table>

*Remember: It has to add up to 100!*

**Encourage Turker honesty:**
- Pay all
- Post-hoc filter for adult native English speakers living in U.S.
Results: Judgment = production ... and Turk matches Lab!

- Alignments (Spearman’s $\rho$)
  - Lab/Corpus 0.638***
  - Turk/Corpus 0.539**
  - Turk/Lab 0.563**
(5) Confirming corpus trends

• The problem:
  – In NomBank, Semantic roles self-embed:
    AGENT (verbal)
    → AGENT (nominal)
  – “The investigation of the police took 3 weeks to complete”

• But the corpus is not representative speech
  – Few minimal pairs (even across the web)
  – Mostly financial terminology
The tasks

What is the closest sentence to this one in meaning:
“The investigation of the police took 3 weeks to complete”

a) the police investigated someone and it took 3 weeks
b) someone investigated the police and it took 3 weeks

Reword the following passage using the phrase “the investigation of the police”:
Following the shooting of a commuter in Oakland last week, a reporter has uncovered new evidence while investigating the police involved.

• Tested different configs/ phrases:
  “It took 3 weeks to complete the investigation of the police”
  “shooting of the hunters”, “destruction of the army” ...
Results

• NomBank falls between comprehension and production

- Confirms a real effect
- Why? Perhaps the ‘role-harmony’ simplifies the cognitive processes required.
(6) Analyzing psycholinguistic data

- N400 effect = semantic anomaly
- N400 effect = metaphor? (Lai et. al. 2009)

The soldiers were attacked
His ideas were attacked
His free time was attacked
Exploring alternative explanations

• Hypothesis:
  – Metaphor processing requires more effort

• Possible counter-hypothesis:
  – Metaphors were semantically unexpected
  – E.g. *spoke* and *fried* are seldom metaphorical

• Task: Collect metaphorical frequencies

Concrete          Abstract

A profile praises Nobel Prize-winning novelist Gabriel Garcia Marquez’s efforts to **heal** his country, as he struggles to overcome lymphatic cancer.
Results and new findings

• Target words were not skewed too literal

• Metaphorical frequency is early!
  – P200 effect for novel metaphor

<table>
<thead>
<tr>
<th>Metaphorical Frequency</th>
<th>Number of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
</tr>
<tr>
<td>0.3</td>
<td>10</td>
</tr>
<tr>
<td>0.4</td>
<td>15</td>
</tr>
<tr>
<td>0.5</td>
<td>20</td>
</tr>
<tr>
<td>0.6</td>
<td>25</td>
</tr>
<tr>
<td>0.7</td>
<td>20</td>
</tr>
<tr>
<td>0.8</td>
<td>15</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Usually-literal (avg. 33%)
Usually-metaphorical (avg. 72%)
(7) Filler informativity and filler-gaps

- Island constraints are syntactic environments that don’t normally permit gaps.
  - If regularities are discrete and inviolable, we want a categorical account.
  - But if they are graded and influenced by non-structural factors, we want a processing account.

SUBJxNULL: [Which perpetrator with a motive]₁ would arresting₁ bother Susan?
SUBJxPOSS: Which perpetrator with a motive would my arresting bother Susan?
SUBJxNAME: Which perpetrator with a motive would Chris arresting bother Susan?

OBJxNULL: [Which perpetrator with a motive]₂ would arresting Susan bother ____?₂
OBJxPOSS: Which perpetrator with a motive would my arresting Susan bother?
OBJxNAME: Which perpetrator with a motive would Chris arresting Susan bother?
Task

• For informativity increase study:
  – 7 point scale for acceptability
  – 12 target sentences and 8 distractor sentences
  – 90 participants, avg of 200 seconds each
Results: Increasing informativity and increase acceptability

Which commissioner would appointing bother Joe?

Reduced informativity of the filler:
Who would deceiving bother Sarah?

Increased informativity of the filler:
Which commissioner would appointing bother Joe?
BONUS! Screening for attentiveness

Challenge: Are our far-flung subjects paying attention (sober, competent, ...)?

Solution: Filler/test items involving subtle but reliable language-specific interpretive contrasts.

Source: Experiments by Carol Chomsky ‘The Acquisition of Syntax in Children from 5 to 10’ (1969)
Screening for attentiveness

1. John is easy to see.
   (John is the understood object of ‘see’)

2. John is eager to see.
   (John is the understood subject of ‘see’)

3. Bozo promised Donald to sing.
   (Donald is the understood object of ‘sing’)

4. Bozo told Donald to sing.
   (Donald is the understood subject of ‘sing’)

5. John’s sister’s friend’s car.
Screening for attentiveness

Results from an auditory study:

<table>
<thead>
<tr>
<th>item type</th>
<th>correct</th>
<th>incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘easy’</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>promise’</td>
<td>59</td>
<td>3</td>
</tr>
<tr>
<td>stacked genitive</td>
<td>55</td>
<td>7</td>
</tr>
</tbody>
</table>

Our subjects performed at the near-perfect level we expect from fluent adults.
Do it yourself

- Turk is actually pretty easy to just pick up
- Keep in mind:
Collect and use demographics
Do it yourself

- Turk is actually pretty easy to just pick up
- Keep in mind:
  - Collect demographic data and restrict AFTER paying, not before (no incentives to lie)
Attract the crowd
Do it yourself

• Turk is actually pretty easy to just pick up
• Keep in mind:
  – Collect demographic data and restrict AFTER paying, not before (no incentives to lie)
  – Make your tasks interesting
  – Keep them brief
  – For a write-up of best practices, see the appendix to Schnoebelen and Kuperman (2010).
• The key to getting reliable results is really going to be your experimental design
• This isn’t just for English!
Most frequent per hour of the day
Conclusions

- Crowdsourcing can quickly reproduce expensive, time-consuming lab studies.
- Linguists have a reliable new tool for experimentation.
- We can quickly and confidently generate empirical results about even infrequent linguistic phenomena.
- This helps confirm, refudiate, and otherwise complicate our introspective data.
Thank you
Some selected references

• A 9-page summary of much of the research in this presentation:
  – Crowdsourcing and language studies: the new generation of linguistic data. Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, Los Angeles, CA. 2010.
    • Rob Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen and Harry Tily
• Longer discussion of reliability (using phrasal verb and Cloze tasks)—including best practices appendix:
  – Tyler Schnoebelen and Victor Kuperman
• Other linguistic tasks and crowdsourcing:
    – Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng
• How attentive are Turkers?
  – "Preventing Satisficing in Online Surveys: A 'Kapcha' to Ensure Higher Quality Data"
    – Adam Kapelner and Dana Chandler
APPENDIX
Demographics

- In early 2010:
  - United States: 46.80%
  - India: 34.00% (amaaaaaazing opportunity for studying Indian languages)
  - Miscellaneous: 19.20%
  - NB: Amazon lets you restrict to any particular country

- From Ipeirotis (2010), see http://bit.ly/c5F1nr
APPENDIX—SEMANTIC SIMILARITY (PHRASAL VERB STUFF)
<table>
<thead>
<tr>
<th></th>
<th>Number of participants under analysis</th>
<th>Number of total ratings</th>
<th>Mode of verbs rated</th>
<th>Median number of verbs rated</th>
<th>Average number of verbs rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>TurkLong</td>
<td>29</td>
<td>2,783</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>TurkShort</td>
<td>66</td>
<td>2,372</td>
<td>6</td>
<td>18</td>
<td>35.94</td>
</tr>
<tr>
<td>TurkContext</td>
<td>81</td>
<td>1,834</td>
<td>6</td>
<td>6</td>
<td>22.64</td>
</tr>
<tr>
<td>StudentContext</td>
<td>27</td>
<td>2,592</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>StudentLong</td>
<td>18</td>
<td>1,728</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>Comparison</td>
<td>ICC consistency</td>
<td>ICC agreement</td>
<td>Weighted kappa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentLong/TurkLong</td>
<td>0.432</td>
<td>0.434</td>
<td>0.384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentLong/StudentContext</td>
<td>0.364</td>
<td>0.320</td>
<td>0.396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentLong/TurkShort</td>
<td>0.439</td>
<td>0.408</td>
<td>0.436</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentLong/TurkContext</td>
<td>0.423</td>
<td>0.365</td>
<td>0.419</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentContext/TurkContext</td>
<td>0.899</td>
<td>0.900</td>
<td>0.823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentContext/TurkShort</td>
<td>0.696</td>
<td>0.692</td>
<td>0.641</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentContext/TurkLong</td>
<td>0.673</td>
<td>0.595</td>
<td>0.596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkContext/TurkLong</td>
<td>0.735</td>
<td>0.642</td>
<td>0.628</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkContext/TurkShort</td>
<td>0.768</td>
<td>0.763</td>
<td>0.746</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkShort/TurkLong</td>
<td>0.913</td>
<td>0.848</td>
<td>0.822</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TurkLong</td>
<td>TurkShort</td>
<td>TurkContext</td>
<td>StudentContext</td>
<td>StudentLong</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>TurkLong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkShort</td>
<td></td>
<td>p=2.797e-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkContext</td>
<td>p=2.606e-09</td>
<td></td>
<td>p=0.0443</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentContext</td>
<td>p=5.025e-08</td>
<td>p=0.101</td>
<td>p=0.869</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentLong</td>
<td>p=0.572</td>
<td>p=0.001</td>
<td>p=6.389e-06</td>
<td>p=4.063e-05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Coefficient of variation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
<td>--------</td>
<td>--------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkLong</td>
<td>3.702</td>
<td>3.649</td>
<td>0.252</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkShort</td>
<td>4.0921</td>
<td>4.173</td>
<td>0.254</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TurkContext</td>
<td>4.224</td>
<td>4.111</td>
<td>0.233</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentContext</td>
<td>4.230</td>
<td>4.204</td>
<td>0.260</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StudentLong</td>
<td>3.740</td>
<td>3.694</td>
<td>0.179</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX—CONTEXTUAL CONSTRAINTS (CLOZE TASKS)
<table>
<thead>
<tr>
<th></th>
<th>Lab</th>
<th>Mechanical Turk</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean predictability</strong></td>
<td>0.332</td>
<td>0.187</td>
<td></td>
</tr>
<tr>
<td><strong>S.D.</strong></td>
<td>0.365</td>
<td>0.246</td>
<td></td>
</tr>
<tr>
<td><strong>Min predictability</strong></td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.15</td>
<td>0.075</td>
<td></td>
</tr>
<tr>
<td><strong>Max predictability</strong></td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>Number of “full misses”</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(not a single match between response word and Schilling target word)</td>
<td>163 / 488</td>
<td>153 / 488</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lab without shared zeros</td>
<td>Turk without shared zeros</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Mean predictability</strong></td>
<td>0.446</td>
<td>0.250</td>
<td></td>
</tr>
<tr>
<td><strong>S.D.</strong></td>
<td>0.358</td>
<td>0.256</td>
<td></td>
</tr>
<tr>
<td><strong>Min predictability</strong></td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.400</td>
<td>0.156</td>
<td></td>
</tr>
<tr>
<td><strong>Max predictability</strong></td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>Number of “full misses”</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>(not a single match between response word and Schilling target word)</strong></td>
<td>39 / 364</td>
<td>29 / 364</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX—FILLER INFORMATIVITY
Extraction from complex subjects

SUBJxNULL: Which commissioner would appointing bother Joe?
SUBJxPOSS: Which commissioner would my appointing bother Joe?
SUBJxNAME: Which commissioner would Steve appointing bother Joe?
OBJxNULL: Which commissioner would appointing Joe bother?
OBJxPOSS: Which commissioner would my appointing Joe bother?
OBJxNAME: Which commissioner would Steve appointing Joe bother?
Not z-scored