Crowdsourcing and language studies: the new generation of linguistic data

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelelen and Harry Tily
Introduction

- Much prior linguistic work in annotation and translation
- Within Stanford Linguistics (and friends)
  - Recent hive of activity for crowdsourcing
  - 7 projects summarized here (from many)
- New *kinds* of data
  - Especially in psycholinguistic experiments
  - Rarely a ‘correct’ response
  - *Artificial-artificial intelligence*
Transparency of phrasal verbs

Segmentation of an audio speech stream

Contextual predictability

Judgment studies of fine-grained probabilistic grammatical knowledge

Confirming corpus trends

Post-hoc metaphorical frequency analysis of electrophysiological responses

Screening for linguistic attentiveness
Mary and John argued, hiking in the __ on the day (that) the first rains cooled (down) the tropical air. As they ran for the nearest shelter she attacked his idea that the investigation of the police would locate his friend’s sister’s car.
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

On a scale of 1–7, how similar are ‘cool’ and ‘cool down’?
Correlations are high
- “StudentLong” is way off (even from the other undergrad group).
- All other tests: $\rho > 0.7$

Turkers and undergrads who saw sentences in context have $\rho > 0.9$
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)

<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>
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 HIT
Here: a classic speech segmentation task
Developed using Adobe Flex (freely available, Javascript-like environment)
(after Saffran et al 1996)

Audio presentation of “words” from a fixed vocabulary concatenated without space:

- tibibagukipaladazuzizivatugi
- 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 indistinguishable from lab subjects in this sample
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.

Margie moved into her new ______
Participants:
- 488 on Turk
- 20 undergrads in Reichle et al.’s lab at UMass

48 sentences, 488 words

Margie moved into her new

-----
- “apartment”
- 17/34 Turkers
- 15/20 undergrads
Strong correlation between Turkers and students

- Spearman’s rank correlation, \( \rho = 0.823 \) (\( p < 0.0001 \))
- If shared zeros are removed to be conservative: \( \rho = 0.759 \) (\( p < 0.0001 \))
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’
- Seen in corpus-derived models of speaker production
- Most significant factor is % predictability of embedded clause, given head noun (RC) or verb (CC) lemma

We \[ \text{hope} \] (that)\text{COMPL} you enjoy
\[ \text{the talk} \] (that)\text{REL} we give ____ today
Encourage Turker honesty:
- Pay all
- Post-hoc filter for adult native English speakers living in U.S.

Really quick background first:

1. What is your gender? Was English your first language?
   ○ Male ○ Yes
   ○ Female ○ No


   

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.).*

   
   
   

*Remember: It has to add up to 100!*
Results: Judgment matches Production ... and Turk matches Lab!

- Alignments (Pearson’s R)
  - Lab/Corpus 0.614***
  - Turk/Corpus 0.562**
  - Turk/Lab 0.633***
The problem:

1. In NomBank, Semantic roles self-embed:

   **AGENT** (verbal)

   **AGENT** (nominal)

2. “The investigation of the police took 3 weeks to complete”

But the corpus is not representative speech:

1. Few minimal pairs (even across the web)
2. 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.
7 Analyzing Psycholinguistic Data

- N400 effect = semantic anomaly
  THE PIZZA WAS TOO HOT TO...

- 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

![Graph showing Metaphorical Frequency vs. Number of Words]

- Usually literal (avg. 33%)
- Usually metaphorical (avg. 72%)
7. 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.
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.
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 the most infrequent linguistic phenomena.