Crowdsourcing and language studies: the new generation of linguistic data

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzyuin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen and Harry Tily

Seven current projects that reproduce many ‘classic’ lab-based studies with relative ease, showing that researchers have a reliable new tool for experimentally investigating language processing and linguistic theory.

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

Cloze predictability

To build models of sentence processing, linguists need to understand the effect of context. One common way to do this is with a Cloze sentence-completion task, where participants select the most likely ‘missing’ word. We compared 468 crowdsourced participants with 20 participants in controlled conditions. The results show a very strong correlation between the two: Spearman’s rank correlation, rho=0.823 (p<0.001).

Judgment studies of grammatical knowledge

“That”-optionality in English complement and relative clauses varies on a range of probabilistic factors in production (corpus) models. We ran six experiments - two baseline studies with lab populations and four crowdsourced trials via MTurk - finding that the same factors involved in production are accessed in judgment and that lab and MTurk results closely correlate.

Analyzing psycholinguistic data

Brainwaves are known to dip when semantic anomalies are encountered. The leading hypothesis is that this is true of metaphors like ‘attack’ in a non-physical context. However, previous brainwave analysis experiments did not control for sense frequency. We used crowdsourcing to reinterpret existing brainwave analysis data, asking participants to categorize all examples as ‘concrete’ or ‘abstract’. When replaying the data in light of this, we found that while metaphoric frequency influences the processing, it does so much more clearly than semantic effects are usually observed.

Screening for attentiveness

Crowdsourcing is most successful when the tasks are designed to be as simple as possible, but in experimental work we don’t always want to target the shallowest knowledge of the participants. In order to ensure that participants are employing the linguistic attentiveness required for more complicated tasks, we test responses to well-known constructions that: (a) exist in most dialects of English; (b) involve high frequency lexical items; and (c) tend to be acquired relatively late by first-language learners. These include minimal pairs with complementary paraphrases like ‘slub is easy to please’, and ‘John is eager to please’ (similarly for ‘gold/guaranty’), and stacked genitives like ‘John’s friends’ sister’s car’.

The accuracy when participants are asked to interpret such sentences indicates that crowdsourced workers do apply the linguistic attentiveness required for more complex language tasks.

<table>
<thead>
<tr>
<th>Item type</th>
<th>correct</th>
<th>incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>weak</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>

Segmentation of a speech stream

We conducted a crowdsourced replication of a classic psychological result that relies on on-line presentation of speech. We developed a novel web-based interface that allowed us to collect this data. Participants heard 75 sentences from one of 16 artificially constructed languages, and were asked to identify individual words. Results from MTurk were statistically indistinguishable from controlled lab-based studies.

Confirming corpus trends

In the NewBank Wall Street Journal corpus there are many phrases with potentially ambiguous nominal semantic roles like “the investigation of the police” – are the police investigating or being investigated? We found that the former of these – the agentive reading – was more likely to occur when the phrase was embedded within a sentential agent, even if the sentential and phrasal semantic roles were not directly related. The corpus does not contain enough examples to rule out stylistic explanations, so we ran crowdsourcing experiments that (a) asked for the interpretation of ambiguous phrases, and (b) asked the participants to use ambiguous phrases in a sentence. In all cases, the results show a clear correlation between embedded and sentential semantic roles, indicating that this is a widespread cognitive processing strategy.