A new generation of data: crowdsourcing and language studies

We present a critical account of the use of crowdsourcing technologies in linguistics. Crowdsourcing has received much recent attention as a fast and inexpensive method for annotation, translation and transcription (Callison-Burch, 2009; Callison-Burch & Dredze, 2010; Hsueh, Melville, & Sindhwani, 2009; Marge, Banerjee, & Rudnicky, 2010; de Marneffe, Manning, & Potts, 2010; Snow, O'Connor, Jurafsky, & Ng, 2008), but we hope to provide evidence that some of the greatest gains are in the very nature of the phenomena that we can now study, especially in experimental paradigms that target potentially rare linguistic phenomena.

Reporting on a number of recent projects that utilize crowdsourcing technologies, we find that the quality is comparable to controlled laboratory experiments, and in some cases superior. For each, we discuss methods for evaluating data quality in the absence of direct interaction with participants or a ‘correct’ response for any given data point.

First, we present the results from semantic transparency experiments showing near-perfect interrater reliability and a strong correlation between crowdsourced data and lab results. Specifically, we asked about phrasal verbs—to what extent does “lifting up” entail “lifting”? Correlations between undergrads and crowdsourced workers were as high as 0.9. Kappa scores for interrater reliability of crowdsourced participants were 0.823, which Landis and Koch (1977) would call “almost perfect agreement”.

Second, we move to audio data, giving crowdsourced workers 75 sentences from one of 16 artificially constructed languages and asking them to identify individual words. Crowdsourced participants were statistically indistinguishable from lab results in this segmentation task (Welch two-sample t-test, t(21.21)=−.92, p=.37).

Third, we show that laboratory results from simple Cloze tasks can be reproduced with crowdsourcing, with very strong correlations between participants in controlled lab environment and crowdsourced workers.

Fourth, we demonstrate that crowdsourcing can also replicate limited-population, controlled-condition lab results for grammaticality judgments. We investigated the factors that influence “that” optionality in English complement and relative clauses, finding strong correlations between participants in controlled lab environment and crowdsourced workers.

Fifth, we use crowdsourcing to augment corpus studies, conducting crowdsourced experiments in both the interpretation and production ambiguous phrases like “the investigation of the police”. The results show strong correlations with the WSJ corpus, comparing minimal pairs with a precision not possible even across web-data.

Sixth, we move to the brain itself, demonstrating that ERP brainwave analysis can be enhanced by crowdsourced analysis of experimental stimuli. Crowdsourced participants were asked to categorize instances of words like “attack” as literal or metaphorical. Reanalyzing brainwave data in light of this, we find a significant difference in average ERPs between highly literal and highly metaphorical words (200ms, p=0.01).

Seventh, we demonstrate the use of crowdsourced acceptability judgment to rapidly explore the interaction between various processing factors and syntax. We find significant effects of filler informativity on the acceptability of complex subject island constraint violations extending earlier lab based research.

Finally, we outline simple heuristics for ensuring that microtasking workers are applying the linguistic attentiveness required to undertake more complex tasks.
References


