# Crowdsourcing and language studies: the new generation of linguistic data

**Robert Munro**[a] **Steven Bethard**[b] **Victor Kuperman**[a] **Vicky Tzuyin Lai**[c]
**Robin Melnick**[a] **Christopher Potts**[a] **Tyler Schnoebelen**[a] **Harry Tily**[a]
[a]Department of Linguistics, Stanford University
[b]Department of Computer Science, Stanford University
[c]Department of Linguistics, University of Colorado
{rmunro,bethard,vickup,rmelnick,cgpotts,tylers,hjt}
@stanford.edu
vicky.lai@colorado.edu

## Abstract

We present a compendium of recent and current projects that utilize crowdsourcing technologies for language studies, finding that the quality is comparable to controlled laboratory experiments, and in some cases superior. While crowdsourcing has primarily been used for annotation in recent language studies, the results here demonstrate that far richer data may be generated in a range of linguistic disciplines from semantics to psycholinguistics. For these, we report a number of successful methods for evaluating data quality in the absence of a 'correct' response for any given data point.

## 1 Introduction

Crowdsourcing's greatest contribution to language studies might be the ability to generate new *kinds* of data, especially within experimental paradigms. The speed and cost benefits for annotation are certainly impressive (Snow et al., 2008; Callison-Burch, 2009; Hsueh et al., 2009) but we hope to show that some of the greatest gains are in the very nature of the phenomena that we can now study.

For psycholinguistic experiments in particular, we are not so much utilizing 'artificial artificial' intelligence as the plain intelligence and linguistic intuitions of each crowdsourced worker – the 'voices in the crowd', so to speak. In many experiments we are studying gradient phenomena where there are no right answers. Even when there is binary response we are often interested in the distribution of responses over many speakers rather than specific data points. This differentiates experimentation from more common means of determining the quality of crowdsourced results as there is no gold standard against which to evaluate the quality or 'correctness' of each individual response.

The purpose of this paper is therefore two-fold. We summarize seven current projects that are utilizing crowdsourcing technologies, all of them somewhat novel to the NLP community but with potential for future research in computational linguistics. For each, we also discuss methods for evaluating quality, finding the crowdsourced results to often be indistinguishable from controlled laboratory experiments.

In Section 2 we present the results from semantic transparency experiments showing near-perfect interworker reliability and a strong correlation between crowdsourced data and lab results. Extending to audio data, we show in Section 3 that crowdsourced subjects were statistically indistinguishable from a lab control group in segmentation tasks. Section 4 shows that laboratory results from simple Cloze tasks can be reproduced with crowdsourcing. In Section 5 we offer strong evidence that crowdsourcing can also replicate limited-population, controlled-condition lab results for grammaticality judgments. In Section 6 we use crowdsourcing to support corpus studies with a precision not possible with even very large corpora. Moving to the brain itself, Section 7 demonstrates that ERP brainwave analysis can be enhanced by crowdsourced analysis of experimental stimuli. Finally, in Section 8 we outline simple heuristics for ensuring that microtasking workers are applying the linguistic attentiveness required to undertake more complex tasks.

## 2 Transparency of phrasal verbs

Phrasal verbs are those verbs that spread their meaning out across both a verb and a particle, as in 'lift up'. *Semantic transparency* is a measure of how strongly the phrasal verb entails the component verb. For example, to what extent does 'lifting up' entail 'lifting'? We can see the variation between phrasal verbs when we compare the transparency of 'lift up' to the opacity of 'give up'.

We conducted five experiments around semantic transparency, with results showing that crowdsourced results correlate well with each other and against lab data ($\rho$ up to 0.9). Interrater reliability is also very high: $\kappa = 0.823$, which Landis and Koch (1977) would call 'almost perfect agreement.'

The crowdsourced results reported here represent judgments by 215 people. Two experiments were performed using Stanford University undergraduates. The first involved a questionnaire asking participants to rate the semantic transparency of 96 phrasal verbs. The second experiment consisted of a paper questionnaire with the phrasal verbs in context. That is, the first group of 'StudentLong' participants rated the similarity of 'cool' to 'cool down' on a scale 1-7:

cool cool down _____

The 'StudentContext' participants performed the same basic task but saw each verb/phrasal verb pair with an example of the phrasal verb in context.
With Mechanical Turk, we had three conditions:

*TurkLong*: A replication of the first questionnaire and its 96 questions.
*TurkShort*: The 96-questions were randomized into batches of 6. Thus, some participants ended up giving responses to all phrasal verbs, while others only gave 6, 12, 18, etc responses.
*TurkContext*: A variation of the 'StudentContext' task – participants were given examples of the phrasal verbs, though as with 'TurkShort', they were only asked to rate 6 phrasal verbs at a time.

What we find is a split into relatively high and low correlations, as Figure 1 shows. All Mechanical Turk tests correlate very well with one another (all $\rho > 0.7$), although the tasks and raters are different. The correlation between the student participants who were given sentence contexts and the workers

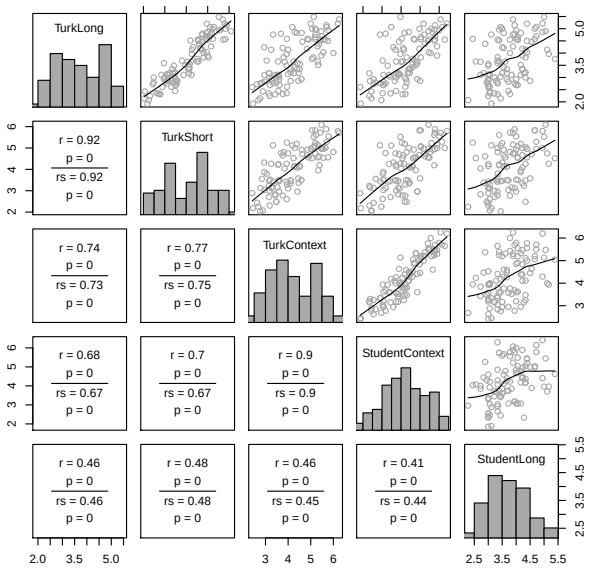


Figure 1: Panels at the diagonal report histograms of distributions of ratings across populations of participants; panels above the diagonal plot the locally weighted scatterplot smoothing Lowess functions for a pair of correlated variables; panels below the diagonal report correlation coefficients (the $r$ value is Pearson's $r$, the $rs$ value is Spearman's $\rho$) and respective $\rho$ values.

who saw context is especially high (0.9). All correlations with StudentLong are relatively low, but this is actually true for StudentLong vs. StudentContext, too ($\rho = 0.44$), even though both groups are Stanford undergraduates.

Intra-class correlation coefficients (ICC) measure the agreement among participants, and these are high for all groups except StudentLong. Just among StudentLong participants, the ICC consistency is only 0.0934 and their ICC agreement is 0.0854. Once we drop StudentLong, we see that all of the remaining tests have high consistency (average of 0.78 for ICC consistency, 0.74 for ICC agreement). For example, if we combine TurkContext and StudentContext, ICC consistency is 0.899 and ICC agreement of 0.900. Cohen's kappa measurement also measures how well raters agree, weeding out chance agreements. Again, StudentLong is an outlier. Together, TurkContext / StudentContext gets a weighted kappa score of 0.823 – the overall average (excepting StudentLong) is $\kappa = 0.700$.

More details about the results in this section can be found in Schnoebelen and Kuperman (submitted).

## 3 Segmentation of an audio speech stream

The ability of browsers to present multimedia resources makes it feasible to use crowdsourcing techniques to generate data using spoken as well as written stimuli. In this section we report an MTurk replication of a classic psycholinguistic result that relies on audio presentation of speech. We developed a web-based interface that allows us to collect data in a statistical word segmentation paradigm. The core is a Flash applet developed using Adobe Flex which presents audio stimuli and collects participant responses (Frank et al., submitted).

Human children possess a remarkable ability to learn the words and structures of languages they are exposed to without explicit instruction. One particularly remarkable aspect is that unlike many written languages, spoken language lacks spaces between words: from spoken input, children learn not only the mapping between meanings and words but also what the words themselves are, with no direct information about where one ends and the next begins. Research in *statistical word segmentation* has shown that both infants and adults use statistical properties of speech in an unknown language to infer a probable vocabulary. In one classic study, Saffran, Newport & Aslin (1996) showed that after a few minutes of exposure to a language made by randomly concatenating copies of invented words, adult participants could discriminate those words from syllable sequences that also occurred in the input but crossed a word boundary. We replicated this study showing that cheap and readily accessible data from crowdsourced workers compares well to data from participants recorded in person in the lab.

Participants heard 75 sentences from one of 16 artificially constructed languages. Each language contained 2 two-syllable, 2 three-syllable, and 2 four syllable words, with syllables drawn from a possible set of 18. Each sentence consisted of four words sampled without replacement from this set and concatenated. Sentences were rendered as audio by the MBROLA synthesizer (Dutoit et al., 1996) at a constant pitch of 100Hz with 25ms consonants and 225ms vowels. Between each sentence, participants were required to click a "next" button to continue, preventing workers from leaving their computer during this training phase. To ensure workers could actually hear the stimuli, they were first asked to enter an English word presented auditorily.

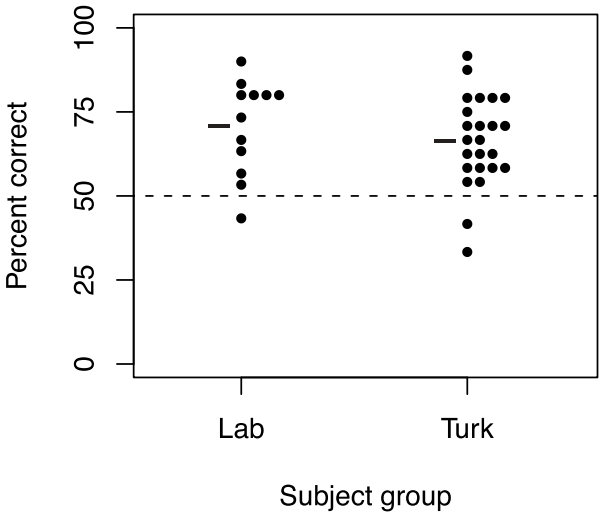


Figure 2: Per-subject correct responses for lab and MTurk participants. Bars show group means, and the dashed line indicates the chance baseline.

Workers then completed ten test trials in which they heard one word from the language and one nonword made by concatenating all but the first syllable of one word with the first syllable of another. If the words "bapu" and "gudi" had been presented adjacently, the string "pugu" would have been heard, despite not being a word of the language. Both were also displayed orthographically, and the worker was instructed to click on the one which had appeared in the previously heard language.

The language materials described above were taken from a Saffran et al. (1996) replication reported as Experiment 2 in Frank, Goldwater, Griffiths & Tenenbaum (under review). We compared the results from lab participants reported in that article to data from MTurk workers using the applet described above. Each response was marked "correct" if the participant chose the word rather than the nonword. 12 lab subjects achieved 71% correct responses, while 24 MTurk workers were only slightly lower at 66%. The MTurk results proved significantly different from a "random clicking" baseline of 50% ($t(23) = 5.92$, $p = 4.95 \times 10^{-06}$) but not significantly different from the lab subjects (Welch two-sample t-test for unequal sample sizes, $t(21.21) = -.92$, $p = .37$). Per-subject means for the lab and MTurk data are plotted in Figure 2.

## 4 Contextual predictability

As psycholinguists build models of sentence processing (e.g., from eye tracking studies), they need to understand the effect of the available sentence context. One way to gauge this is the Cloze task proposed in Taylor (1953): participants are presented with a sentence fragment and asked to provide the upcoming word. Researchers do this for every word in every stimulus and use the percentage of 'correct' guesses as input into their statistical and computational models.

Rather than running such norming studies on undergraduates in lab settings (as is typical), our results suggest that psycholinguists will be able to crowdsource these tasks, saving time and money without sacrificing reliability (Schnoebelen and Kuperman, submitted).

Our results are taken from 488 Americans, ranging from age 16-80 (mean: 34.49, median: 32, mode: 27) with about 25% each from the East and Midwest, 31% from the South, the rest from the West and Alaska. They represent a range of education levels, though the majority had been to college: about 33.8% had bachelor's degrees, another 28.1% had some college but without a degree.

By contrast, the lab data was gathered from 20 participants, all undergraduates at the University of Massachusetts at Amherst in the mid-1990's (Reichle et al., 1998). Both populations provided judgments on 488 words in 48 sentences. In general, crowdsourcing gave more diverse responses, as we would expect from a more diverse population.

The correlation between lab and crowdsourced data by Spearman's rank correlation is 0.823 ($\rho < 0.0001$), but we can be even more conservative by eliminating the 124 words that had predictability scores of 0 across both groups. By and large, the lab participants and the workers are consistent in which words they fail to predict. Even when we eliminate these shared zeros, the correlation is still high between the two data sets: weighted $\kappa = 0.759$ ($\rho < 0.0001$).

## 5 Judgment studies of fine-grained probabilistic grammatical knowledge

Moving to syntax, we demonstrate here that grammaticality judgments from lab studies can also be reproduced through crowdsourcing.

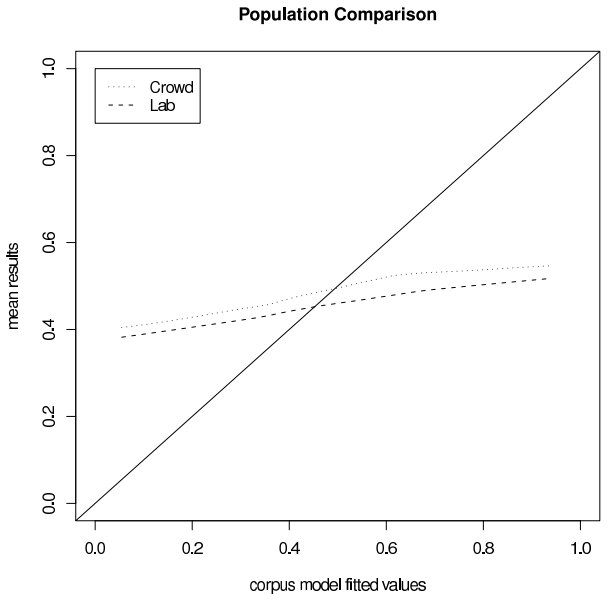


Figure 3: Mean 'that'-inclusion ratings plotted against corresponding corpus-model predictions. The solid line would represent perfect alignment between judgments and corpus model. Non-parametric Lowess smoothers illustrate the significant correlation between lab and crowd population results.

Corpus studies of spontaneous speech suggest that grammaticality is gradient (Wasow, 2008), and models of English complement clause (CC) and relative clause (RC) 'that'-optionality have as their most significant factor the predictability of embedding, given verb (CC) and head noun (RC) lemma (Jaeger, 2006; Jaeger, in press). Establishing that these highly gradient factors are similarly involved in judgments could provide evidence that such fine-grained probabilistic knowledge is part of linguistic competence.

We undertook six such judgment experiments: two baseline studies with lab populations then four additional crowdsourced trials via MTurk.

Experiment 1, a lab trial (26 participants, 30 items), began with the models of RC-reduction developed in Jaeger (2006). Corpus tokens were binned by relative model-predicted probability of 'that'-omission. Six tokens were extracted at random from each of five bins ($0 \leq \rho < 20\%$ likelihood of 'that'-inclusion; $20 \leq \rho < 40\%$; and so on). In a gradient scoring paradigm with 100 points distributed between available options (Bresnan, 2007) partici-

pants rated how likely each choice – with or without 'that' – was as the continuation of a segment of discourse. As hypothesized, mean participant ratings significantly correlate with corpus model predictions ($r = 0.614$, $\rho = 0.0003$).

Experiment 2 (29 participants) replicated Experiment 1 to address concerns that subjects might be 'over-thinking' the process. We used a timed forced-choice paradigm where participants had from 5 to 24 seconds (varied as a linear function of token length) to choose between the reduced/unreduced RC stimuli. These results correlate even more closely with predictions ($r = 0.838$, $\rho < 0.0001$).

Experiments 3 and 4 replicated 1 and 2 on MTurk (1200 tasks each). Results were filtered by volunteered demographics to select the same subject profile as the lab experiments. Response-time outliers were also excluded to avoid fast-click-through and distracted-worker data. Combined, these steps eliminated 384 (32.0%) and 378 (31.5%) tasks, respectively, with 89 and 66 unique participants remaining. While crowdsourced measures might be expected to yield lower correlations due to such unbalanced data sets, the results remain significant in both trials ($r = 0.562$, $\rho = 0.0009$; $r = 0.364$, $\rho = 0.0285$), offering strong evidence that crowdsourcing can replicate limited-population, controlled-condition lab results, and of the robustness of the alignment between production and judgment models. Figure 3 compares lab and crowd population results in the 100-point task (Experiments 1 and 3).

Experiments 5 and 6 (1600 hits each) employed the same paradigms via MTurk to investigate 'that'-mentioning in CCs, where predictability of embedding is an even stronger factor in the corpus model. Filtering reduced the data by 590 (36.9%) and 863 (53.9%) hits. As with the first four experiments, each of these trials produced significant correlations ($r = 0.433$, $\rho = 0.0107$; $r = 0.500$, $\rho = 0.0034$; respectively). Finally, mixed-effect binary logistic regression models – with verb lemma and test subject ID as random effects – were fitted to these judgment data. As in the corpus-derived models, predictability of embedding remains the most significant factor in all experimental models.

The results across both lab and crowdsourced studies suggest that speakers consider the same factors in judgment as in production, offering evidence that competence grammar includes access to probability distributions. Meanwhile, the strong correlations across populations offer encouraging evidence in support of using the latter in psycholinguistic judgment research.

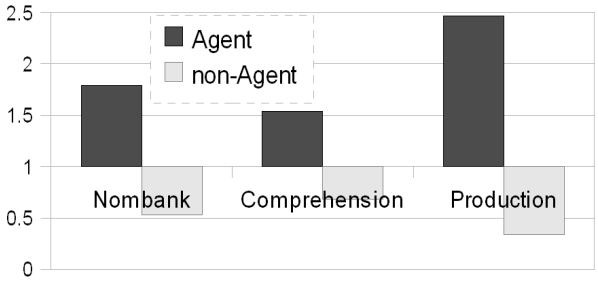


Figure 4: Odds ratio of a Nominal Agent being embedded within a Sentential Agent or non-Agent, relative to random chance. ($\rho < 0.001$ for all)

## 6 Confirming corpus trends

Crowdsourcing can also be used to establish the validity of corpus trends found in otherwise skewed data. The experiments in this section were motivated by the NomBank corpus of nominal predicate/arguments (Meyers et al., 2004) where we found that an Agent semantic role was much more likely to be embedded within a sentential Agent. For example, (1) is more likely than (2) to receive the Agent interpretation for the 'the police', but both have same potential range of meanings:

(1) "The investigation of the police took 3 weeks to complete"
(2) "It took 3 weeks to complete the investigation of the police"

While the trend is significant ($\rho < 0.001$), the corpus is not representative speech.

First, there are no minimal pairs of sentences in NomBank like (1) and (2) that have the same potential range of meanings. Second, the *s-genitive* ("the police's investigation") is inherently more Agentive than the *of-genitive* ("the investigation of the police") and it is also more compact. Sentential subjects tend to be lighter than objects, and more likely to realize Agents, so the resulting correlation could be indirect. Finally, if we sampled only the predicates/arguments in NomBank that are frequent in different sentential positions, we are limited to:

"earning, product, profit, trading, loss, share, rate, sale, price". This purely financial terminology is not representative of a typical acquisition environment – no child should be exposed to only such language – so it is difficult to draw broad conclusions about the cognitive viability of this correlation, even within English. It is because of factors like these that corpus linguistics has been somewhat of a 'poor cousin' to theoretical linguistics.

Therefore, two sets of experiments were undertaken to confirm that the trend is not epiphenomenal, one testing comprehension and one testing production.

The first tested thousands of workers' interpretations of sentences like those in (1) and (2), over a number of predicate/argument pairs ("shooting of the hunters", "destruction of the army" etc). Workers were asked their interpretation of the most likely meaning. For example, does (1) mean: "a: the police were doing the investigation" or "b: the police are being investigated". To control for errors or click-throughs, two plainly incorrect options were included. We estimate the erroneous response rate at about 0.4% – less than many lab studies.

For the second set of experiments, workers were asked to reword an *unambiguous* sentence using a given phrase. For example, rewording the following using "the investigation of the police":

(3) "Following the shooting of a commuter in Oakland last week, a reporter has uncovered new evidence while investigating the police involved."

We then (manually) recorded whether the required phrase was in a sentential Agent or non-Agent position.

Figure 4 gives the results from the corpus analysis and both experiments. The results clearly show a significant trend for all, and that the NomBank trend falls between the comprehension and production tasks, which would be expected for this highly edited register. It therefore supports the validity of the corpus results.

The phenomena likely exists to aid comprehension, as the cognitive realization of just one role needs to be activated at a given moment. Despite the near-ubiquity of 'Agent' in studies of semantic roles, we do not yet have a clear theory of this linguistic entity, or even firm evidence of its existence

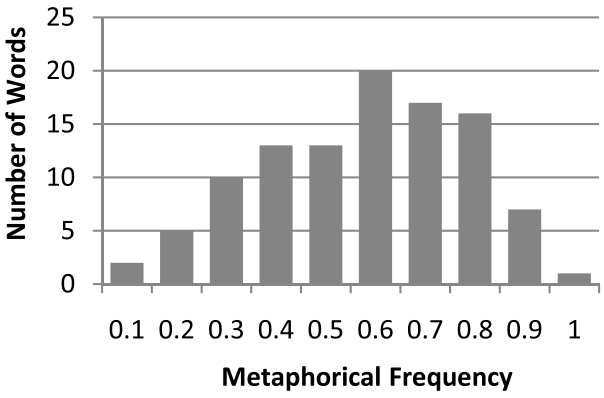


Figure 5: Distribution of metaphorical frequencies.

(Parikh, 2010). This study therefore goes some way towards illuminating this. More broadly, the experiments in this section support the wider use of crowdsourcing as a tool for language cognition research in conjunction with more traditional corpus studies.

## 7 Post-hoc metaphorical frequency analysis of electrophysiological responses

Beyond reproducing laboratory and corpus studies, crowdsourcing also offers the opportunity to newly analyze data drawn from many other experimental stimuli. In this section, we demonstrate that crowdsourced workers can help us better understand ERP brainwave data by looking at how frequently words are used metaphorically.

Recent work in event related potentials (ERP) has suggested that even conventional metaphors, such as "All my ideas were attacked" require additional processing effort in the brain as compared to literal sentences like "All the soldiers were attacked" (Lai et al., 2009). This study in particular observed an N400 effect where negative waves 400 milliseconds after the presentation of the target words (e.g. *attacked*) were larger when the word was used metaphorically than when used literally.

The proposed explanation for this effect is that metaphors really do demand more from the brain than literal sentences. However, N400 effects are also observed when subjects encounter something that is semantically inappropriate or unexpected. While the Lai experiment controlled for overall word frequency, it might be possible to explain away these N400 effects if it turned out that in the real

world the target words were almost always used literally, so that seeing them used metaphorically would be semantically incongruous.

To test this alternative hypothesis, we gathered sense frequency distributions for each of the target words – the hypothesis predicts that these should be skewed towards literal senses. For each of the 104 target words, we selected 50 random sentences from the American National Corpus (ANC), filling in with British National Corpus sentences when there were too few in the ANC. We gave the sentences to crowdsourced workers and asked them to label each target word as being used literally or metaphorically. Each task contained one sentence for each of the 104 target words, with the order of words and the literal/metaphorical buttons randomized. Each sentence was annotated 5 times.

To encourage native speakers of English, we had the MTurk service require that our workers be within the United States, and posted the text "Please accept this HIT only if you are a native speaker of English" in bold at the top of each HIT. We also used Javascript to force workers to spend at least 2 seconds on each sentence and we rejected results from workers that had chance level (50%) agreement with the other workers.

Though our tasks produced words annotated with literal and metaphorical tags, we were less interested in the individual annotations (though agreement was decent at 73%) and more interested in the overall pattern for each target word. Some words, like *fruit*, were almost always used literally (92%), while other words, like *hurdle* were almost always used metaphorically (91%) .

Overall, the target words had a mean metaphorical frequency of 53%, indicating that their literal and metaphorical senses were used in nearly equal proportions. Figure 5 shows that the metaphorical frequencies follow roughly a bell-curved distribution[1], which is especially interesting given that the target words were hand-selected for the Lai experiment and not drawn randomly from a corpus. We did not observe any skew towards literal senses as the alternative hypothesis would have predicted. This suggests that the findings of Lai, Curran, and Menn (2009) cannot be dismissed based on a sense frequency argument.

| Item type | correct | incorrect |
|---|---|---|
| 'easy' | 60 | 2 |
| 'promise' | 59 | 3 |
| stacked genitive | 55 | 7 |

Table 1: Response data for three control items, with the goal of identifying workers who lack the requisite attentiveness. All show high attentiveness. The difference between the 'easy' and 'stacked genitive' is trending but not significant ($\rho = 0.0835$), indicating that any of these may be used.

We also took advantage of the collected sense frequency distributions to re-analyze data from the Lai experiment. We split the target words into a high bin (average 72% metaphorical) and a low bin (average 33% metaphorical), matching the number of items and average word log-frequency per bin. Looking at the average ERPs (brain waves) over time for each bin revealed that when subjects were reading novel metaphors, there was a significant difference ($p = .01$) at about 200ms (P200) between the ERPs for the highly literal words and the ERPs for the highly metaphorical words. Thus, not only does metaphorical frequency influence figurative language processing, but it does so much earlier than semantic effects are usually observed (e.g. N400 effects at 400ms)[2].

## 8 Screening for linguistic attentiveness

For annotation tasks, 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 workers, so here we seek to discover just how attentive the workers really are.

When running psycholinguistics experiments in the lab, the experimenters generally have the chance to interact with participants. It is not uncommon for prospective subjects to be visibly exhausted, distracted, or inebriated, or not fluent in the given language to a requisite level of competence. When these participants turn up as outliers in the experimental data, it is easy enough to see why — they fell asleep, couldn't understand the instructions, etc.

[1] A Shapiro-Wilk test fails to reject the null hypothesis of a normal distribution (p=0.09).

[2] These results are consistent with recent findings that irony frequency may also produce P200 effects (Regel et al., 2010).

With crowdsourcing we lose the chance to have these brief but valuable encounters, and so anomalous response data are harder to interpret.

We present two simple experiments for measuring linguistic attentiveness, which can be used as one component of a language study or to broadly evaluate the linguistic competency of the workers. Taking well-known constructions from the literature, we selected constructions that: (a) exist in most (perhaps all) dialects of English; (b) involve high frequency lexical items; and (c) tend to be acquired relatively late by first-language learners.

We have found two constructions from Carol Chomsky's (1969) work on first-language acquisition to be particularly useful:

(4) John is easy to see.
(5) John is eager to see.

Example (4) is accurately paraphrased as 'It is easy to see John', where John is the object of 'see', whereas (5) is accurately paraphrased as 'John is eager for John to see', where John is the subject of 'see'. A similar shift happens with 'promise':

(6) Bozo told Donald to sing.
(7) Bozo promised Donald to sing.

We presented workers with a multiple-choice question that contained both subject and object paraphrases as options.

In similar experiments, we adapted examples from Roeper (2007), who looked at stacked prenominal possessive constructions:

(8) John's sister's friend's car.

These are cross-linguistically rare and challenging even for native speakers. As above, the workers were asked to choose between paraphrases.

Workers who provide accurate judgments are likely to have a level of English competence and devotion to the task that suffices for many language experiments. The results from one short audio study are given in Table 1. They indicate a high degree of attentiveness; as a group, our subjects performed at the near-perfect levels we expect for fluent adults.

We predict that adding tasks like these to experiments will not only screen for attentiveness, but also prompt for greater attention from an otherwise distracted worker, improving results at both ends.

## 9 Conclusions

While crowdsourcing was first used by linguists for annotation, we hope that the results here demonstrate the potential for far richer studies. In a range of linguistic disciplines from semantics to psycholinguistics it enables systematic, large-scale judgment studies that are more affordable and convenient than expensive, time-consuming lab-based studies. With crowdsourcing technologies, linguists have a reliable new tool for experimentally investigating language processing and linguistic theory.

Here, we have reproduced many 'classic' large-scale lab studies with a relative ease. We can envision many more ways that crowdsourcing might come to shape new methodologies for language studies. The affordability and agility brings experimental linguistics closer to corpus linguistics, allowing the quick generation of targeted corpora. Multiple iterations that were previously possible only over many years and several grants (and therefore never attempted) are now possible in a matter of days. This could launch whole new multi-tiered experimental designs, or at the very least allow 'rapid prototyping' of experiments for later lab-based verification.

Crowdsourcing also brings psycholinguistics much closer to computational linguistics. The two fields have always shared empirical data-driven methodologies and computer-aided methods. We now share a work-space too. Historically, NLP has necessarily drawn corpora from the parts of linguistic theory that have stayed still long enough to support time-consuming annotation projects. The results here have implications for such tasks, including parsing, word-sense disambiguation and semantic role labeling, but the most static parts of a field are rarely the most exciting. We therefore predict that crowdsourcing will also lead to an expanded, more dynamic NLP repertoire.

Finally, for the past half-century theoretical linguistics has relied heavily on 'introspective' corpus generation, as the rare edge cases often tell us the most about the boundaries of a given language. Now that we can quickly and confidently generate empirical results to evaluate hypotheses drawn from intuitions about the most infrequent linguistic phenomena, the need for this particular fallback has diminished – the stimuli are abundant.

## Acknowledgements

We owe thanks to many people, especially within the Department of Linguistics at Stanford, which has quickly become a hive of activitiy for crowdsourced linguistic research. In particular, we thank Tom Wasow for his guidance in Section 5, Chris Manning for his guidance in Section 6, and Florian T. Jaeger for providing the corpus-derived base models in Section 5 (Jaeger, 2006). We also thank Michael C. Frank for providing the design, materials, and lab data used to evaluate the methods in Section 3. Several of the projects reported here were supported by Stanford Graduate Fellowships.

## References

Joan Bresnan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in search of its evidential base*, pages 75–96. Mouton de Gruyter, Berlin.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295.

Carol Chomsky. 1969. *The Acquisition of Syntax in Children from 5 to 10*. MIT Press, Cambridge, MA.

Thierry Dutoit, Vincent Pagel, Nicolas Pierret, Franois Bataille, and Olivier van der Vrecken. 1996. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Fourth International Conference on Spoken Language Processing*, pages 75–96.

Michael Frank, Harry Tily, Inbal Arnon, and Sharon Goldwater. submitted. Beyond transitional probabilities: Human learners impose a parsimony bias in statistical word segmentation.

Michael Frank, Sharon Goldwater, Thomas Griffiths, and Joshua Tenenbaum. under review. Modeling human performance in statistical word segmentation.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35.

Florian Jaeger. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Ph.D. thesis, Stanford University, Stanford, CA.

Florian Jaeger. in press. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*.

Vicky Tzuyin Lai, Tim Curran, and Lise Menn. 2009. Comprehending conventional and novel metaphors: An ERP study. *Brain Research*, 1284:145–155, August.

Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, , and Ralph Grishman. 2004. Annotating noun argument structure for NomBank. In *Proceedings of LREC-2004*.

Prashant Parikh. 2010. *Language and Equilibrium*. MIT Press, Cambridge, MA.

Stefanie Regel, Seana Coulson, and Thomas C. Gunter. 2010. The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research*, 1311:121–135.

Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological Review*, 105:125–157.

Tom Roeper. 2007. *The Prism of Grammar: How Child Language Illuminates Humanism*. MIT Press, Cambridge, MA.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Word segmentation: The role of distributional cues. *Journal of memory and language*, 35:606–621.

Tyler Schnoebelen and Victor Kuperman. submitted. Using Amazon Mechanical Turk for linguistic research: Fast, cheap, easy, and reliable.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew T. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Wilson Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Tom Wasow. 2008. Gradient data and gradient grammars. In *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*, pages 255–271.