A case study in effectively crowdsourcing long tasks with novel categories

It is commonly assumed that crowdsourcing is effective only when the tasks are quick and require no special training. This would seem to make it inappropriate for experiments in which participants must undergo a training period to get acquainted with novel category labels and stimuli, or complete a large number of tasks. We challenge the notion that crowdsourcing is actually limited in this way. We had Mechanical Turk workers annotate 642 event descriptions in sentences using the seven veridicality values of the FactBank corpus. The sentences were presented in blocks of 30 items: four training items designed to acquaint participants with the categories, three 'tests' very similar to the training items, included to ensure that the Turkers were careful, and 23 experimental items. We obtained 10 annotations for each event description. 177 workers participated. Just two workers did the test items incorrectly and had to be thrown out. The overall Fleiss kappa score was 0.50 (this figure is conservative because it is insensitive to the partial ordering of the tags), which is higher than some previous numbers reported for similar tasks done with expert annotators. Our by-category kappas are as high as 0.80. In addition, when we study the distribution of annotations we obtained for each sentence, we find evidence for pragmatic ambiguities that are explicable in terms of theories of utterance meaning. We use these findings to draw general lessons about how crowdsourcing can be done effectively even for long and complex tasks.