A Case for a Range of Acceptable Annotations

Jennimaria Palomaki, Olivia Rhinehart, Michael Tseng
Google
1600 Amphitheatre Parkway
Mountain View, California 94043
{jpalomaki, orhinehart, michaeltseng}@google.com

Abstract

Multi-way annotation is often used to ensure data quality in crowdsourced annotation tasks. Each item is annotated redundantly and the contributors' judgments are converted into a single "ground truth" label or more complex annotation through a resolution technique (e.g., on the basis of majority or plurality). Recent crowdsourcing research has argued against the notion of a single "ground truth" annotation for items in semantically oriented tasks—that is, we should accept the aggregated judgments of a large pool of crowd contributors as "crowd truth." While we agree that many semantically oriented tasks are inherently subjective, we do not go so far as to trust the judgments of the crowd in all cases. We recognize that there may be items for which there is truly only one acceptable response, and that there may be divergent annotations that are truly of unacceptable quality. We propose that there exists a class of annotations between these two categories that exhibit *acceptable variation*, which we define as the range of annotations for a given item that meet the standard of quality for a task. We illustrate acceptable variation within existing annotated data sets, including a labeled sound corpus and a medical relation extraction corpus. Finally, we explore the implications of acceptable variation on annotation task design and annotation quality evaluation.

1 Introduction

With respect to annotation quality, one dichotomy in human-annotated data is the categorization of crowd contributors' annotations or labels into "noise" and the more nebulous notion of "ground truth." The development of a human-annotated "gold" standard has long been seen as essential to the training and evaluation of natural language processing systems in particular.

Recent research suggests that there is no such thing as a single "golden" label or a single ground truth for semantically oriented data (Aroyo and Welty 2013a; 2013b; 2015). Factors such as the ambiguity of the input items and clarity of task guidelines, as well as differences in contributor backgrounds and levels of conservativeness, have been shown to condition disagreement among annotations (Dumitrache 2015; Kairam and Heer 2016; Kapelner et al. 2012).

With respect to noise, whether annotations are provided by "expert" annotators, crowd contributors, or some other source, "spammy" or noisy data is a pervasive and perhaps inevitable feature of human-annotated data.

Regarding ground truth, it may be that for some subset of items paired with a given semantic annotation task, there really is only one label that satisfies the standard of quality for the task. This would likely be the case if for that subset of input items, the task guidelines are clear, the content of the items themselves is less ambiguous, and contributor-specific factors such as background would not condition varying interpretations of the task or items.

So then what lies between noise and the notion of a single "golden" label for each item in a data set for a given task? We propose here that this middle ground is occupied by annotations that exhibit what we call acceptable variation. We define acceptable variation as the range of annotations (labels, answers, etc.) at any annotation stage that meet the standard of quality for a given task, which we conceptualize in terms put forth by Allahbakhsh et al. (2013), who define quality for tasks as "the extent to which the provided outcome fulfills the requirements of the requester."

We will illustrate, using examples from previous semantic annotation tasks that have been used by researchers to argue against the notion of a single ground truth, that there is a kind of disagreement which falls into the range of acceptable variation. We will extend our argument to non-semantic input data to show that the notion of acceptable variation is useful to any annotation task requiring human judgments.

Finally, we will discuss the implications of our proposal. If we accept that for a given annotation task, there is some subset of input items that allow a range of acceptably varying annotations, then we must be explicit about how this affects our orientation to task design, the evaluation of the annotations, and thus the overall quality of the labeled data. We argue that acceptable variation is an important feature of any data set that is meant to be representative of the kind of complex semantic phenomena that require human annotation. We propose that task design and task guidelines should be broad enough to allow for a wealth of answers and thereby facilitate gathering more representative data. We will also discuss ways in which acceptable variation can be identified and leveraged.

Furthermore, we will argue that the existence of acceptable variation has important implications for how we evaluate the performance of contributors. Task designers must be careful not to mistake acceptable variation for noise and unwittingly penalize contributors who are providing us with valuable signals about the complexity of the phenomena we are trying to understand.

2 Conditioning Acceptable Variation

It is worth considering what gives rise to both disagreement and the particular type of disagreement that we characterize as acceptable variation. We follow Dumitrache (2015) and Kairam and Heer (2016) who discuss three main sources of disagreement: (1) differences in contributors, including background knowledge or understanding of the task, which may lead them to annotate more or less conservatively; (2) clarity of the expected annotation label(s); and (3) ambiguity of the content of the task's input items from the data set.

We will show that these factors are important in understanding how acceptable variation is conditioned, and we will demonstrate examples of each type in the annotation tasks we consider below.

3 Acceptable Variation in Published Data Sets

Aroyo and Welty (2013a) introduce "crowd truth" as an alternative to ground truth (the notion that for each task item there is such as a thing as a single correct annotation). They ground their discussion of crowd truth in the disparity in annotator disagreement between entity type annotation tasks and relation annotation tasks. They note that disagreement tends to be much higher for relation annotation tasks because there are multiple ways to express the same relation and, conversely, the same linguistic expression may express many different relations.

Aroyo and Welty (2013a; 2013b) argue that the disagreement seen among contributors in relation annotation tasks is not necessarily noise or a sign of a poorly defined problem, but rather a useful and informative property of human-annotated data that provides a signal about the vagueness or ambiguity of the input items for the annotation task.

In this section, we will adduce examples of acceptable variation in a medical relation corpus and a labeled sound corpus, which are multi-classification and free response tasks, respectively. We acknowledge that acceptable variation may surface in other types of tasks as well, such as ones with binary or scale rating (e.g., sentiment analysis). We may explore those in later work.

Medical Relation Corpus

The medical relation annotation task that Aroyo and Welty (2013b) used to investigate the disagreement in medical relation extraction allowed crowd contributors to select from a given set any relations that they judge as applicable to highlighted terms in a given sentence. An additional step asked contributors to provide a justification for their selection of the relation(s) they chose for each input sentence. The set of relations was selected manually from the United Medical Languages System (UMLS). The set included the relations *treats*, *prevents*, *diagnosed by test or drug*, *causes*, *location*, *symptom*, *manifestation*, *contraindicates*, *associated with*, *side effect*, *is a*, *part of*, *other*, and *none*. Crowd contributors were provided with a definition and example sentence for each relation (and given the design of the task, it is likely that the terms were interpreted in their local linguistic contexts). Each input sentence was annotated by multiple contributors.

The annotations demonstrate the kind of disagreement we characterize as acceptable variation, which we differentiate from noise or spammy annotations.

Consider the example task item in Table 1 with original highlighted terms in capital letters. The task item elicits varying annotations, both within and outside of the acceptable range. Note that since contributors were allowed to select all relations that they judged to be applicable, the number of judgments (16) exceeds the number of contributors (15) who annotated this sentence.

These data suggest that subclinical RIBOFLAVIN DEFICIENCY may occur in adolescents and that deficiency may be related to dietary intake of RIBOFLAVIN.

Relation Annotation	**Count**
associated with	4
symptom	3
causes	3
prevents	1
side effect	1
manifestation	1
part of	1
diagnose by test or drug	1
other	1

Table 1: Counts of labels chosen by crowd contributors for the relations between "RIBOFLAVIN" and "RIBOFLAVIN DEFICIENCY" for this given sentence from the medical relation annotation task.

Not all of the relation annotations for the example above can be characterized as falling within the acceptable range. For example, it is unlikely that (the dietary intake of) "RIBOFLAVIN" is a *manifestation* of "RIBOFLAVIN DEFICIENCY." Yet the under-specification of what kind of "dietary intake" of "RIBOFLAVIN" is related to "RIBOFLAVIN DEFICIENCY" opens the sentence up to different interpretations. The sentence is unclear as to whether "dietary intake of riboflavin" here refers to lack of, insufficient, sufficient, or excessive consumption of riboflavin. While "riboflavin deficiency" can suggest that the relation is conditioned by lack of (or insufficient) riboflavin consumption, the variation of annotations suggests that this condition is not obvious to crowd contributors who are not medical experts. This is unsurprising given that contributor backgrounds have been shown by previous research (Dumitrache 2015; Kairam and Heer 2016) to condition disagreement. It is worth noting that Aroyo and Welty (2013a; 2013b) found that even medical experts had difficulty reaching consensus on what relations held between terms in similar sentences.

Let us assume that riboflavin deficiency is caused by the lack of (or insufficient) consumption of riboflavin. Considering the other factors that Dumitrache (2015) and Kairam and Heer (2016) point to as conditioning disagreement, the relation annotation *associated with* may have been selected by more conservative annotators, as it subsumes a relationship in which any degree of riboflavin consumption may be causally related to riboflavin deficiency. The annotation falls within the range of acceptable annotations for this task.

If we interpret the phrase "dietary intake of riboflavin" as referring to the medically recommended level of riboflavin consumption, then the relation annotation *prevents* also falls within the range of acceptable annotations for this task. The contributor who selected this relation may have had prior knowledge about the relationship between riboflavin and riboflavin deficiency, or may have been more liberal in the interpretation of the relationship suggested by the sentence.

If we interpret "dietary intake of riboflavin" as referring to the lack of (or insufficient) consumption of riboflavin, we can understand why the *causes* and *symptom* and even *side effect* relations were selected so frequently, regardless of the fact that the relation between the individual highlighted terms is not best expressed by these labels. The relations *causes*, *symptom*, and *side effect* may not hold between the specific terms "RIBOFLAVIN" and "RIBOFLAVIN DEFICIENCY," yet a causal relation is indeed suggested by the sentence. These annotations may fall outside of the range of acceptable variation, but they provide a valuable signal of how the input sentence and the terms themselves may lead to ambiguous interpretations by crowd contributors.

The frequency with which contributors selected the *causes* and *symptom* relation annotations places them within the set of relation annotations selected by a plurality of contributors. These judgments are therefore in some sense "true" under the notion of crowd truth, as they depend upon (and provide insight into) the ambiguity of the input sentences as interpreted by the crowd contributors for this task. Acceptable variation is distinct from crowd truth in that it is independent of plurality or majority. For a given input item for a multi-way annotation task, it may be the case that only one contributor selects a label or annotation, yet this choice may fall within the range of acceptable variation, as with the *prevents* relation annotation for the input sentence in Table 1.

Even seemingly straightforward input items can condition acceptable variation. Consider the example in Table 2.

Since the input sentence contains the relation expression "caused," it is not surprising that the majority of contributors selected the relation annotation *causes*. However, the broader relation annotation *associated with* subsumes the *causes* relation, so it also falls within the range of acceptable variation for this task. This is notable because this relation annotation was selected by only one contributor out of 15, suggesting that at least for this particular input item, majority and plurality are not necessarily indicators of the quality or "truth" of annotations.

<table>
<tr><td colspan="2">FUNGAL INFECTIONS may be caused by several FUNGI the most important of these being Candida species including C. albicans, C. glabrata, C. krusei, C. tropicalis, C. parapsilosis, and C. guillermondii.</td></tr>
<tr><td><b>Relation Annotation</b></td><td><b>Count</b></td></tr>
<tr><td>causes</td><td>13</td></tr>
<tr><td>associated with</td><td>1</td></tr>
<tr><td>part of</td><td>2</td></tr>
</table>

Table 2: Counts of labels chosen by crowd contributors for the relations between "FUNGI" and "FUNGAL INFECTIONS" for this given sentence from the medical relation annotation task.

VU Sound Corpus

Crowdsourced annotation tasks can be designed to encourage a variety of acceptable answers. An example of such is the labeling task that generated the VU Sound Corpus (van Miltenburg, Timmermans, and Aroyo 2016). In this task, contributors added keyword descriptions to environmental sound clips from the Freesound database (Font, Roma, and Serra 2013). The authors used quality analysis to root out spammy responses, but otherwise did not consider there to be "right" or "wrong" labels for any given sound clip. Rather than restrict the format of the labels in the task's guidelines or user interface, post-processing was used to normalize and cluster keywords, which helped identify annotation outliers.

We will show that acceptable variation is present in the VU Sound Corpus, and that it is likely conditioned by factors previously discussed, including differences in contributor background and conservativeness. Furthermore, we will discuss how the presence of acceptable variation is a desirable attribute of the labeled data.

Table 3 is an example where contributors' background knowledge may have conditioned the variation seen in the responses. Some contributors (8, 10) added keywords referencing a possible source of the sound, *feedback*, while others (6, 7) opted for more concrete keywords describing the sound itself, such as *whistle*. This difference in emphasis may be due to varying levels of familiarity or world experience with how to generate a sound of this type ("eerie horror film sound").

Table 4 demonstrates differing levels of conservativeness of contributors. The input sound clip of a hand-held electric beater in use is ambiguous enough that contributors who chose to add specific labels for motorized devices split into two groups: those who used labels related to *drill* (2, 3, 5, 7, 8, 9), and those who used labels related to *blender* (4, 6; 9 falls into both groups if *mixer* is understood as a synonym of *blender*). Two contributors (1, 10) opted to back off to a more general keyword, *machine*. Van Miltenberg et al. acknowledge that, while keywords related to *drill* and *blender* both incorrectly identify the source of the sound, the labels still offer useful information about what the recording sounds like, which could be useful for grounded semantic

URL	https://www.freesound.org/people/NoiseCollector/sounds/6212/
Description	Multisamples created with subsynth. Eerie horror film sound in middle and higher registers. Normalized and converted to AIFF in cool edit 96. File name indicates frequency for example: HORROC04.aif= C4, where last 3 characters are C04
Tags	evil, horror, subtractive, synthesis
Keywords	
1	swing, metallic
2	shrill, shriek
3	whine
4	high, tuning, resonance
5	HIGH PITCHED SOUND
6	whistle
7	wistle
8	amplifier feedback, high pitched tone
9	screech
10	feedback

Table 3: Lists of keywords applied by each crowd contributor (identified by number) to the given "eerie horror film sound" audio clip from the VU Sound Corpus labeling task.

URL	https://www.freesound.org/people/terminal/sounds/22795/
Description	A sample of my Sunbeam Beatermix Pro 320 Watt electric beater at low speed setting #2. Recorded on 2 tracks in The Closet using a Rode NT1A and a Rode NT3 mic, mixed to stereo and processed through a multi band limiter.
Tags	appliance, beater, electric, kitchen
Keywords	
1	machine
2	drilling, grating, noisy
3	DRILL
4	blender
5	drill
6	blender
7	drill, drilling
8	drill, rattle, buzz
9	mixer,drill,whirring
10	machine

Table 4: Lists of keywords applied by each crowd contributor (identified by number) to the given "Sunbeam Beatermix Pro 320 Watt electric beater" audio clip from the VU Sound Corpus labeling task.

models. Note that the factually incorrect label *drill* was chosen by a majority of crowd contributors: this is an example of "crowd truth" that may indicate what many in the crowd actually perceive when they hear the audio clip. The more conservative label *machine* would fall into the range of acceptable variation.

The crowd-generated keywords complement the ones generated by the authors who originally uploaded the sounds by providing perspective from contributors who were not involved with the creation of the underlying corpus of sound clips. In reference to their own labels, the authors note, "Well-informed parties commonly overlook things that are obvious to them"—that is, the designers of the corpus suffer from a "curse of knowledge" that constrains the keyword options that they would consider for an ambiguous sound.

4 Task Design Implications

We can extrapolate the application of this insight to anyone who is designing an annotation task. Even if they do not participate in the creation of the underlying data set to be labeled, task designers usually have knowledge and contextual information about the data, the problem space, and the model informing the annotation scheme. If task designers constrain the task definition (guidelines, user interface, annotation options, etc.) and do not provide space for contributors to include new or unexpected annotations, then they risk foregoing the potential insights of the crowd. That is, the crowd contributors' lack of context enables them to serve as a fresh set of eyes (or ears) when evaluating the task input items and making annotation judgments. In addition, task designers should consider including channels for feedback from contributors on the task design and input data, which may take the form of a comment box for text input within the task itself. Especially if leveraged in the earlier stages of annotation, this method can help in shaping the ultimate design of a task.

We recognize that not all task designers would be comfortable with a fully open-ended label set. As a compromise, task designers could implement an iterative process to better define the range of acceptable variation: first, run more open-ended pilots to get a sense of the range of annotations that contributors provide for a given task; next, classify the resulting annotations that diverge from expected options as acceptable or not; and then refine the annotation scheme to include new options corresponding to the acceptably varying annotations.

This approach is similar to the transition along the continuum from "user-driven" to "model-driven" annotation tasks described in Chang et al. (2016) in exploring the domain of conceptual relationships implicitly expressed by noun–noun compounds. The first version of task consisted of an explicitly user-driven design encouraging contributors to write their own paraphrases describing the relationship(s) between the nouns in each given noun–noun compound. After reviewing the variety of paraphrases, the task designers settled on a more consistent format for writing the paraphrases that would allow for contributor creativity while also facilitating more reliable extraction of relationship terms that could be included in training data for a machine learning model. The task designers additionally developed more extensive guidelines and reinforcement training for contributors.

Involving crowd contributors in the iterative design of the task can also increase their understanding of the data and foster a sense of shared purpose in the data labeling effort. We encourage the research community to prioritize such engagement as an explicit component of the task design. For example, the standard practice of injecting items with "golden answers" (that is, task items that have been labeled by the task designers or other experts) into each contributor's task queue is used to evaluate the crowd's individual and aggregate performance against expert performance, but it does not necessarily measure each contributor's conscientiousness or engagement with the task.

Apparent divergence from the expert standard would be particularly exaggerated when the "golden answers" are sampled to over-represent edge cases or otherwise tricky items (this is often done when expert judgments are expensive or difficult to acquire). Furthermore, as we have discussed, for tasks where acceptable variation is expected, there is often no single correct label or annotation for a given task item, and comparing crowd contributors with experts may unfairly penalize contributors who are providing valuable information through their varying judgments. For such tasks, we would like to advance the notion of "golden *questions*"—that is, prompts that are designed to elicit consistent answers or labels. These more objective prompts need not be related to the more subjective task; they could be interspersed into the queue as special task items or appear as a section of the main task template. They would serve to establish a baseline for each crowd contributor's consistency in judgment. Contributors who answer the more objective prompts consistently (with themselves and with others) are likely to be providing useful signal if they exhibit variation in responses to the more subjective prompts. Analysis would still be needed to distinguish the variation due to suboptimal design of the subjective prompts from the variation that would be considered acceptable.

5 Conclusion and Future Work

We have introduced the notion of *acceptable variation*, which we conceptualize as both a natural progression of and complementary to the notion that there is no such thing as a single ground truth or "golden" label for items in semantically oriented tasks. We exhibited the existence of acceptable variation in existing research, showing that the notion is amenable to tasks that range from more inherently semantic (relation annotation) to more inherently perceptual (sound clip labeling).

Identifying and Leveraging Acceptable Variation

If we accept the notion that for a given task that requires human annotation or evaluation, some subset of items can be expected to exhibit acceptable variation, we must determine how we will differentiate acceptable variation from actual noise and how we will extract value from this distinction. This may be particularly challenging for crowdsourcing workflows that depend upon plurality-based resolution, since, as we demonstrated with examples above, it may be the case that an acceptably varying annotation was provided by a single contributor in multi-way annotation for a given input item. However, as Aroyo and Welty (2013b) demonstrate, the sum of different contributor disagreement measures can be used to identify contributors providing annotations of lower quality. These annotations would likely fall outside of the range of acceptable variation.

For the medical relation annotation task described above, Aroyo and Welty found that low-quality annotations were provided by contributors who disagreed consistently with other contributors across tasks, while disagreement across annotations for an individual sentence was used to score each input item for sentence clarity. Contributor annotations were also evaluated on the basis of whether contributors provided original justifications for the relations they chose for a given input sentence (rather than just copying and pasting the sentence itself in part or in whole) and on the basis of the average number of relations a contributor selected for each sentence. The authors determined that contributors attempting to appear more agreeable would consistently select multiple relations for each sentence. These three metrics were used to classify (with 98% accuracy) 12 out of 110 contributors as providers of low-quality annotations.

Aroyo and Welty's approach to identifying contributors providing low-quality annotations suggests that, for the remaining contributors, disagreement is likely to provide signal about the ambiguity of the input sentences or relation labels or other factors that might condition disagreement, such as contributors' backgrounds. We propose that similar approaches to disagreement can also be used to identify contributors whose disagreement is likely to fall within the range of acceptable variation.

Alternatively, agreement may also provide signal about which contributors are likely to provide annotations within the acceptable range. Injecting tasks with input items that are more likely to elicit agreement in multi-way annotation (because they are less ambiguous and less likely to condition disagreement due to worker differences) could serve to establish a baseline for determining trusted contributors whose annotations for items that are likely to elicit disagreement (because they are more ambiguous) would likely fall within the acceptable range.

This type of validation method, in which verifiable annotations are used to validate subjective annotations, was employed by Kittur, Chi, and Suh (2008) in experiments used to assess the utility of the micro-task market on Amazon's Mechanical Turk platform as a way to collect user input.

We argue that acceptable variation is an important feature of human-annotated data and that it is important to distinguish disagreement caused by actual annotation errors and disagreement that falls within the acceptable range. The immediate value of making the distinction is that it facilitates a more nuanced understanding of disagreement for human annotation tasks generally. Furthermore, it prompts consideration of what acceptable variation would look like for specific tasks and how it should be accounted for in task design, annotation quality evaluation, and data quality evaluation.

Acceptable Variation and Annotation Quality

The notion that there is no such thing as a single "golden" label for items in semantically oriented tasks has implications for how we measure and report contributor reliability. If we take majority or plurality as the measure of correctness and penalize contributors whose annotations fall outside of those metrics, we may be punishing contributors for annotations that fall within the range of acceptable variation. As we demonstrated, acceptable variation is not necessarily reflected by plurality or even by a majority of annotations. This suggests that even 90% agreement for an item does not guarantee that the item is unambiguous to the degree that it will not condition some acceptable variation, at least for the medical relation annotation task we examined.

Kittur et al. (2013) propose that the future of crowd work should articulate a fair vision through innovation, which addresses the challenge of creating reputation systems that are not amenable to cheating or gaming while maintaining the benefits of pseudonymity and low-transaction cost hiring. Our findings suggest that to that end, task designers (and everyone else who is using or evaluating the labeled data) must also consider how to differentiate between genuine errors and acceptable variation, particularly if and when they are statistically indistinguishable, to ensure that contributor reputations are fairly and accurately assessed.

We plan to address annotation quality evaluation in future work. In our own labeled data sets, we have been able to identify acceptably varying annotations manually on a relatively small subset of the data, but we will need to consider how to design automatic methods at scale.

We invite the research community to embrace acceptable variation as a useful insight into the complexity of human judgment for ambiguous or otherwise subjective tasks, and to confront the implications explicitly when designing and evaluating crowdsourcing tasks.

6 Acknowledgments

We would like to thank Russell Lee-Goldman, Slav Petrov, and other colleagues in Google Research for their helpful feedback on this work as well as the reviewers for their valuable comments and suggestions. We are grateful for the contributors who annotated the data sets we analyzed.

References

Allahbakhsh, M.; Benatallah, B.; Ignjatovic, A.; Motahari-Nezhad, H. R.; Bertino, E.; and Dustdar, S. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17:76–81.

Aroyo, L., and Welty, C. 2013a. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science 2013*.

Aroyo, L., and Welty, C. 2013b. Measuring crowd truth for medical relation extraction. In *Semantics for Big Data*.

Aroyo, L., and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36(1):15–24.

Chang, N.; Lee-Goldman, R.; and Tseng, M. 2016. Linguistic wisdom from the crowd. In *Crowdsourcing Breakthroughs for Language Technology Applications*.

Dumitrache, A. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *European Semantic Web Conference*, 701–710.

Font, F.; Roma, G.; and Serra, X. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, 411–412.

Kairam, S., and Heer, J. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 1637–1648.

Kapelner, A.; Kaliannan, K.; Schwartz, H. A.; Ungar, L.; and Foster, D. 2012. New insights from coarse word sense disambiguation in the crowd. In *Proceedings of COLING 2012: Posters*, 539–548.

Kittur, A.; Nickerson, J. V.; Bernstein, M. S.; Gerber, E. M.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. J. 2013. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 453–456.

van Miltenburg, E.; Timmermans, B.; and Aroyo, L. 2016. The vu sound corpus: Adding more fine-grained annotations to the freesound database. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.