TILBURG UNIVERSITY

Tilburg University

The VU Sound Corpus

Miltenburg, Emiel van; Timmermans, Benjamin; Aroyo, Lora

Published in:
Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Miltenburg, E. V., Timmermans, B., & Aroyo, L. (2016). The VU Sound Corpus: Adding More Fine-grained Annotations to the Freesound Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)* European Language Resources Association (ELRA).

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 26. Jan. 2025

The VU Sound Corpus
Adding More Fine-grained Annotations to the Freesound Database

Emiel van Miltenburg, Benjamin Timmermans, Lora Aroyo
The Network Institute
Vrije Universiteit Amsterdam
{emiel.van.miltenburg, b.timmermans, lora.aroyo}@vu.nl

Abstract

This paper presents a collection of annotations (tags or keywords) for a set of 2,133 environmental sounds taken from the Freesound database (`www.freesound.org`). The annotations are acquired through an open-ended crowd-labeling task, in which participants were asked to provide keywords for each of three sounds. The main goal of this study is to find out (i) whether it's feasible to collect keywords for a large collection of sounds through crowdsourcing, and (ii) how people talk about sounds, and what information they can infer from hearing a sound in isolation. Our main finding is that it is not only feasible to perform crowd-labeling for a large collection of sounds, it is also very useful to highlight different aspects of the sounds that authors may fail to mention. Our data is freely available, and can be used to ground semantic models, improve search in audio databases, and to study the language of sound.
Keywords: Sounds, sound terms, Crowdsourcing, Corpus, Onomatopoeia

1. Introduction

Recent years have seen a growing interest in annotated collections of perceptual stimuli. Most of the attention has been directed towards either image labeling (Deng et al., 2009) or image description (Young et al., 2014; Lin et al., 2014). This paper presents a collection of annotations (tags or keywords) for a set of environmental sounds taken from the Freesound database.[1] The annotations are acquired through an open-ended crowd-labeling task. The main goal of this study is to find out how people talk about sounds, and what information they can infer from hearing a sound in isolation. Our data can be used to ground semantic models, improve search in audio databases, and to study the language of sound.[2]

1.1. Freesound and Other Sound Collections

The Freesound database (Font et al., 2013) is an open collaborative database of almost 300 000 sounds released under Creative Commons licenses. Each sound is provided by its author with a description and a set of keywords. Though there are some guidelines for the descriptions,[3] authors are free to add whatever keywords they want. There are also some more standardized datasets, but all of these are much smaller. Hocking et al. (2013) provide a comprehensive overview of environmental sound norming studies, of which Saygin et al. (2005) present the largest dataset with 236 annotated sounds. All studies use sounds with a duration between ten seconds and less than a second.

1.2. Grounded Semantic Models

In distributional semantics, the meaning of a word is a function of the contexts in which that word occurs (Turney et al., 2010). Up until recently, 'context' was simply taken to mean 'surrounding words'. Recent work in *multimodal distributional semantics* challenges this notion because text corpora form "an extremely impoverished basis compared to the rich perceptual sources that ground human semantic knowledge" (Bruni et al., 2014, p. 1). Bruni et al. trained a distributional model using not only text, but also visual features extracted from a large collection of images. This way, part of the referential meaning of the words is also incorporated in their abstract vector representation. Bruni et al.'s results lead others to experiment with other modalities including sound (Lopopolo and van Miltenburg, 2015; Kiela and Clark, 2015) and even smell (Kiela et al., 2015).

Both sound-related studies used the Freesound database because it is the richest resource currently available, but it also has its limitations. The main issue is that it's impossible to tease apart the keywords that are sound-related from the keywords that serve another purpose. (See Strohmaier et al. (2012) for a discussion of user motivations in tagging.) One example is the keyword *field-recording*, which is only indirectly related to acoustic contents of the sounds. The problem with abstract keywords like this is that it is unclear whether we should even try to learn a multimodal representation for them. With concrete terms like *hit, quack, boing, trample*, it is more intuitive that acoustic information might add to the knowledge that we already have about these terms from their corpus distribution. (It tells us *what it's like*.) Kiela and Clark (2015) get around this by annotating a list of words for sound-relatedness. We present an alternative approach: by asking participants to provide keywords that are related to the sounds, we are effectively creating an extensive list of concrete sound-related terms (with examples).

1.3. Crowdsourcing and Disagreement

In order to gather keywords descriptions for the sound, we used an open-ended task so that the crowd worker is not limited by a predefined reference space. Because there is no such thing as a right or wrong answer in this task, it is impossible to use a 'gold standard' to assess whether the crowd-workers are doing a good job or whether they are just providing random responses. Instead, we rely on the CrowdTruth framework which provides several *dis-*

[1] `www.freesound.org`

[2] The code and data is available at: `https://github.com/CrowdTruth/vu-sound-corpus`

[3] See `http://freesound.org/help/faq/`

agreement metrics (Aroyo and Welty, 2014; Dumitrache et al., 2015). In their explanation of these metrics, Aroyo and Welty refer to the *triangle of reference* (Ogden and Richards, 1923, see also figure 1), in which an interpreter perceives a sign and tries to identify the referent (or meaning) of that sign.

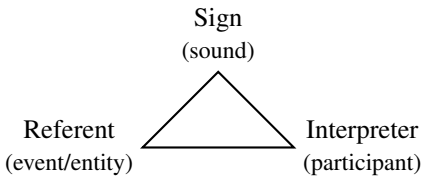

Figure 1: Triangle of reference (Ogden and Richards, 1923), modified to fit our task.

Applied to our crowd-labeling task, we can say that the participants perceive a sound, and try to identify the event or entity producing that sound. In our task, the participants are simply asked to describe the sounds they hear using keywords (which hopefully are associated with the event or entity producing the sound). Because this is an open-ended task, we expect there to be variation in the keywords used by our participants. The amount of variation in the keywords is a measure of disagreement. Based on the triangle of reference, Aroyo and Welty note that this disagreement could be the result of three factors: (i) bad crowd-worker performance, (ii) a vague signal, and (iii) the existence of multiple possible referents. Rather than treating the variation in keywords as noise, the CrowdTruth approach sees disagreement between participants as a signal that (if we filter out the spam) provides us with more information about the interpretability of each sound. We will consider sounds with low disagreement to be *clear*, and sounds with high disagreement to be *unclear*. This will be reflected in a *clarity score* assigned to each sound in our corpus. See section 3. for more details on the CrowdTruth disagreement metrics.

1.4. Goals and Hypotheses

This section provides a list of goals and hypotheses that are the driving factors behind this paper:

Feasibility One of the goals of this work is to see whether it is feasible to collect good quality annotations for a large collection of sounds. This is not trivial, because there is no gold standard, and it is impractical to manually check all annotations. But we assumed that it would be possible through the CrowdTruth framework.

Variation and Range We wanted to get a sense of the range of expressions that people might use in the environmental sound domain, and the variation in keywords used for each recording. To this end, we opted for an open-ended task in which participants were free to enter whatever keywords they thought of, so that they were not limited to a predefined reference space.

Authors versus Crowd We expected that authors of the sounds are more likely to use high-level keywords than our participants, who we expected to use more low-level descriptions. We hypothesized that this should be the case because the authors have full knowledge of the sounds, while our participants were confronted with the sounds in isolation. (And so they only have the sound to base their keywords on.)

Sound Length and Clarity We expected longer sounds to elicit more variation in the keywords used by our participants, yielding a lower clarity score. With longer recordings, our participants might attend to different parts of the recording, leading to different keywords. With shorter recordings, it is more likely that all participants focus on the same aspects of the recording.

Keywords and Search We contacted Frederic Font, developer of Freesound.org, to obtain search queries from users of the database. These queries also reflect how people talk about sounds. We hoped to be able to show an overlap between our crowd-sourced keywords and the search queries, beyond the keywords already used by the authors of the sounds. This would open up the possibility of using crowdsourcing to improve search results in the Freesound database. At the same time (and this was our main concern), a large overlap between the crowd-sourced keyword and the search queries supports the idea that our crowd-sourced keywords actually correspond to how people generally talk about sounds.

2. Setup

As mentioned in the introduction, we carried out an open-ended crowd-labeling task. This section provides more details on the data and the task design.

2.1. Data

For our task we used 2,133 mp3 sounds from the Freesound database.[4] All these recordings and their metadata are freely accessible through the Freesound API.[5] Font et al. (2014) manually classified a part of the database into five categories: *SoundFX, soundscapes, samples, music* and *speech*. We focused on sounds from the SoundFX category.

2.2. Task Design

We used the Crowdflower platform[6] to collect annotations for the sounds. Figure 2 shows a trial from our annotation task. In this task, participants were asked to listen to a sound and provide (comma-separated) keywords in an empty text field to describe that sound. The separate keywords are previewed live below the input to provide feedback.

Participants were allowed to enter an infinite number of keywords, where each could consist of multiple words. They were also asked not to use phonetic words as these have no semantic meaning, and not to write full sentences. Lastly, they were asked to check their spelling, to reduce the chance of false spelling corrections during post-processing. We distributed our task to crowd workers from the US, UK, Australia and Canada in order to maximize the English vocabulary of the annotators. After the pilot study, each task was set up to contain three sounds and ordered randomly

[4] Mp3 is good enough for our goal of crowdsourcing labels. For analyses requiring uncompressed sounds, it is possible to get the original recordings in .WAV format from the database as well.

[5] https://www.freesound.org/docs/api/

[6] http://www.crowdflower.com/

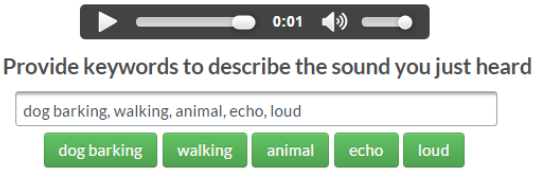

Figure 2: Example item from the annotation task. By giving a preview of the entered keywords, the green 'buttons' encourage the use of commas to separate the words.

upon display to prevent biases. The crowd workers were rewarded $0.02 for each completed task containing three sounds.

2.3. Batches

We uploaded our task in three batches (summarized in table 1) in order to determine the best setup. First, a pilot experiment was carried out for which we downloaded a sample of 99 sounds with a maximum duration of 30 seconds. We then ran several tests ran to determine the optimal number of judgments per sound, number of sounds per task and payment per task. We settled on 10 judgments per sound, resulting in an average of 17 keywords for each sound. We also found that having three sounds annotated in one task proved more optimal, because it allows the crowd worker to revise their answers and it contributes to reducing the overall cost.

During the pilot phase, we added live feedback in the form of keyword labels appearing below the input field (see also figure 2). These labels preview the keywords as they would be submitted with the current input. We added this functionality for two reasons: (i) it stimulates the crowd workers to verify their input, and (ii) seeing the labels appear might encourage workers to add more keywords.

After the pilot phase, we created another set of 900 randomly selected recordings, evenly divided into classes of 300 short (< 1 second), medium ($5 - 6$ seconds) and long ($17 - 21$ seconds) recordings. We used this data set to study the influence of duration on the homogeneity of the keywords. During the experiments, we found that the recordings with a longer duration were more likely to contain multiple sound-events (see also section 4.4.). This created noise in the results: on an individual level, it is impossible to tell which keywords correspond to which sound event in the recording, and on a global level it means that keywords become associated with each other even though they describe different events. In order to reduce this noise, we created a third set with 1134 short sounds. This increases the chance of only having 'atomic' sound-recordings that correspond to single sound-events.

For the third batch with short sounds, we gathered 10 judgments per sound with three sounds per task. These settings increased the average number of keywords annotated by a crowd worker from 1.21 to 1.65 and reduced the time workers took to annotate the sound from 30 seconds to 18 seconds per sound. The improvements of the preliminary experiments reduced the cost to gather the annotations per sound from $0.20 to $0.09.

Batch	# Sounds	Duration (sec)
1. Random	99	< 30
2. Short	300	< 1
Medium	300	5 to 6
Long	300	17 to 21
3. Short only	1134	< 1

Table 1: Batches of sounds that were used in our task.

3. Measuring Disagreement

In order to evaluate the quality of the results and remove low quality annotations, the CrowdTruth[7] framework was used. The framework uses the disagreement based metrics on vector space representations of the annotations, based on the triangle of reference as explained in Section 1.3.. The vector of each sound contains the frequency of each unique keyword annotated for that sound. In order to optimize the effectiveness of the metrics, the outliers were removed and similar keywords were clustered. The next paragraphs will explain these post-processing steps in detail.

3.1. Outlier removal

First, we used several functions to detect and filter out workers of which the annotations were obvious outliers and considered spam. This was done before the clustering of similar keywords, in order to prevent keywords from being clustered into outlying keywords. All annotations of a worker were removed if more than two of its annotations did not match any of five criteria: (i) An annotation for a given trial cannot contain more than two duplicate keywords. (ii) an annotation cannot contain more than eight keywords. (iii) A keyword cannot contain more than 5 words. (iv) Keywords have to be shorter than 50 characters. (v) The average keyword length should be less than 20 characters. In addition, we removed all annotations of workers with more than two trials failing the criteria. Manual evaluation showed that these criteria removed the workers of which the annotations were considered spam, but further evaluation was needed to identify and remove low-quality annotations.

3.2. Keyword Clustering

In order to further evaluate the results, we normalized the keywords by clustering them together. This increases the contrast of agreement between participants and the *clarity* of the sound. A recording is clear when there is little dispute among the participants as to what was actually recorded (e.g. everybody agrees that a sound is the result of glass shattering). By contrast, a recording is unclear if it is open to different interpretations or if it contains multiple events. One difficulty with automatically determining sound clarity on the basis of keywords is that different keywords may be used to describe the same event. Therefore we clustered all the keywords on a sound-by-sound basis (i.e. each time we compare the keywords associated with a particular sound to each other) to reduce variation as much

[7] http://crowdtruth.org

as possible. We defined several functions to normalize the data, including a lemmatizer (e.g. *walks* → *walk*), a spell-checker (e.g. *synthesiser* → *synthesizer*), and a function to standardize compound spellings (e.g. *gunshot*/*gun-shot* → *gun shot*).

CheckDashes Replace all dashes inside keywords with spaces.

CheckSpaces Check for all keywords containing a space whether a space-less variant is also included in the list. Then replace the space-less variant with the one containing a space. E.g. if we find a sound tagged with both *gunshot* and *gun shot*, we replace the former with the latter.

CheckSpelling Check for all keywords whether they are in a vocabulary list. If not, then compute their Levenshtein distance to all keywords that do appear in the list. If the distance is equal to 1, replace the misspelled word with the correct one. E.g. *synthesiser* → *synthesizer*.

CheckOrder Check for all keywords containing a space whether there is another keyword containing a space with the same words, but ordered differently. If so, order them alphabetically. E.g. if we find both *gun shooting* and *shooting gun*, we replace the latter with the former.

CheckMorphology Check for all keywords whether there is another keyword for the same sound that has the same stem but a different ending (including null endings). If so, replace them with the stem. E.g. *walk, walks, walking* all get replaced with *walk*.

CheckInclusion Check whether a keyword without spaces constitutes one of the words of a keyword with spaces. If so, replace the former with the latter. E.g. replace *water* with *water dripping* if both keywords occur with a particular sound.

CheckSemantics Check for all pairs of keywords whether they are semantically similar or related, using the pre-trained GoogleNews `word2vec` model (Mikolov et al., 2013). If so, replace the least common keyword with the most common one.

CheckSubstring Check for all pairs of keywords whether one is a substring of the other. If so, replace the substring with the full string.

The order of application is important. One function may make the list of keywords more suitable for another function, so that it can reduce it further, or it may actually change the data so that another function cannot reduce the list anymore.[8] We have ordered the functions such that the more basic operations are applied first. We did not experiment with any other orderings.

3.3. Spam Filtering

Following the clustering procedure, we used the CrowdTruth disagreement-based metrics to filter out low-quality annotations. For each sound, a vector was constructed of each clustered keyword in that sound. Then, for each worker annotating that sound a worker vector was created with the annotated frequency of each keyword represented in that vector. If a worker's annotations are extremely dissimilar to the other annotations (average cosine distance of less than 0.05 to other workers' annotation vectors) or extremely similar (average cosine distance of 0.95 or higher to other workers' annotation vectors), they were regarded as too low quality and removed from the data set. This resulted in removal of 5% of the annotations, increasing the overall quality of the results.

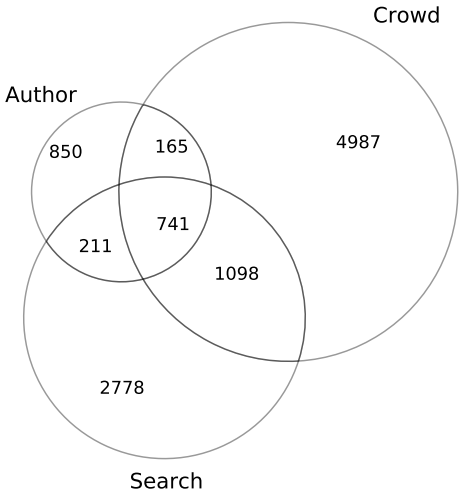

Figure 3: Type overlap between author tags, crowd tags and the search terms from the Freesound search logs.

3.4. Clarity Score

After filtering out the low-quality workers, the variation in the annotations for each sound were used to compute a *clarity score*. For this, all worker vectors for one sound were aggregated to form a unit vector. Then, for each keyword in the aggregated vector, the cosine distance between that keyword and all other keywords annotated in that sound was computed. The clarity score corresponds to the inverse of the maximum cosine distance between the keywords. This is a number between 0 and 1, that is high if there is agreement between the annotators on what can be heard in the sound, and low if the annotators disagree on what can be heard.

4. Results and Discussion

In total 573 crowd workers annotated 2,133 sounds with 30,289 keywords (6,057 unique terms), amounting to an average of 15 keywords per sound. We exported the data in XML format (see the appendix for the specifics), and also made a browser-based viewing tool that makes it easier to inspect the data.[9]

4.1. Variation and Range

Because we obtained search logs from the Freesound.org website, we now have three collections of sound terms: the keywords provided by the authors, those provided by the crowd, and the search logs. To understand how these sets

[8] Kiparsky (1968) uses the terms 'feeding' and 'bleeding' to describe such relations between functions applying one after another (in the context of phonological rules).

[9] The tool is available at: `https://github.com/evanmiltenburg/SoundBrowser`

Set	% Overlap
Author tags	57
Search terms	72
Both	55
Non-overlapping	26

Table 2: Token overlap between the crowd-labels and other sets. 'Both' refers to the intersection of the author tags and the search terms.

differ, we first generated a Venn-diagram showing the overlap in terms of types between the author-tags, search terms, and the raw (unclustered) crowd tags (figure 3). Here we see that 71.3% of the crowd-labels does not occur in the search terms or the author-provided tags. This characterization is a bit misleading, though, because those labels might just occur once or twice (e.g. typos or creative variations), while other keywords that do overlap with the other sets might be more frequent. When we look at the overlap in terms of tokens, we indeed see that the situation is reversed: table 2 shows that only 26% of the crowd-labels does not occur in the set of author-provided tags or in the set of search terms while 55% of the crowd-labels occurs in both. Moreover, 72% of the crowd-labels occurs in the set of search terms. From this, we can conclude that the keywords used by the crowd are relatively similar to the search terms entered by users of the Freesound database. They differ more with respect to the tags used by the authors. And as we will see next, there are also differences between the authors and the crowd in terms of the *distribution* of the keywords.

4.2. Authors versus Crowd

Our hypothesis was that the crowd would provide us with more low-level descriptions of the sounds. That is: terms that are less connected with the *production* of the sound, and more related to the *experience* of the sound. To find out whether this is indeed the case, we computed the log-likelihood for all the words in the intersection between the Freesound database and the crowd-annotated tags (Rayson and Garside, 2000). This measure corresponds to the 'surprisal' of finding a particular word n times in a given corpus, given its combined frequency in both corpora and the overall size of the corpora. Using the log-likelihood, we compiled two lists of words that are typical for each of the sources ($LL > 3.84$).[10] Our impression is that the crowd-annotations are indeed more closely related to the experience of the sounds themselves. Here is a sample (the full lists are available online):

Crowd airplane, whip, whir, pluck, bird, horn, ping, ching, fade, drop, whistle, cymbals, squeaking, swoosh, steps, punch, splash, flutter, grind, zipper, blender, tinkle, jingle

Freesound recording, industrial, freaky, glitch, synth, explosions, impact, concrete, melody, percussion, cat, mechanical, pad, record, voice, cinematic, retro, raw, slash

[10] 3.84 is the cutoff value 0.05 significance level, see: `http://ucrel.lancs.ac.uk/llwizard.html`

A good example of the low/high level contrast is sound #158802, which has been put on Freesound.org by a commercial party (with a higher-quality version on their own website). This sound is tagged by 'sound-experts' with the following tags:

> Film; Radio; Future; Alien; Futuristic; effects; Broadcasting; Recording; fx; Music-Production; Video; Screen; media; TV; space; Remixing; alien-sound-effects; Sound-Effects; pod-Cast; DVD; Home-Videos

Here are the crowd-tags for the same sound:

> ultro sound; video pings; mysterious; computerized; chaotic; robotic; robot; scales; buttons; computer; bleeping; synthetic; chimes; random; descending; tones; discordant; electronic; chaos; science; beeps; technology; playing music backward; high; pitchy; mix

Note that the author-tags is more related to the context (aliens) and possible applications for the sounds (music production), while the crowd-tags focus more on the structure (tones, descending, chaotic) and the experience of the sounds (mysterious).

4.3. Sound interpretation

Because the authors have full knowledge of how the sound was produced, we can use the author tags and descriptions as a reference to see where the crowd 'gets it wrong.' More research is needed to be able to do this automatically, but we can already make some observations. For example: while the crowd is good at identifying 'typical' sounds (*velcro, toilet flushing, musical instruments, dogs, cats, chickens*), they have more difficulty with 'generic' mechanical sounds and loud bangs. And when our crowd workers weren't able to identify the source of the sound, they did either of two things: (i) use more concrete terms, or (ii) guess or associate the sound with something familiar. An example: one of the sounds (number 151837) is described by its author as "rubbing a knife around on some cabbage to create some squeaking, creaking sounds." This is very difficult to guess, so here our crowd-workers used the concrete terms *squeak, squeaking, creak, buzz* or *quacking*. Guesses/associations for this sound are: *balloon, rubbing, goat, animal* and *baby crying*.

While the former strategy is very useful for us to get more concrete keywords, what should we think of the latter? (Besides telling us something about the clarity of the sound.) Strictly speaking, the guesses provided by the crowd are wrong. On the other hand, these associations do tell us what the recordings *sound like*. This information might be very valuable for people who don't care what produced the sound, but only about the impression that the sound makes. Grounded semantic models, for example, should be fine with this kind of data.

4.4. Sound Length and Clarity

In order to investigate the noise in the data, we compared the duration of the sounds with the measured clarity of the sounds in the second batch. Using a one-way ANOVA, we found no significant difference in the clarity of the short, medium or long sounds ($F(2, 897) = 2.87, p = 0.056$). We did find a significant difference in the amount of keywords annotated by the crowd for each set ($F(2, 897) =$

$227.84, p = 9.64 * 10^{-81}$), with an average amount of 14, 18, and 21 keywords for the short, medium, and long sounds. This indicates that there was a difference as to *how much* can be heard, but also that that there was not a significant difference as to *what* our crowd-workers heard.

4.5. Keywords and Search

As figure 3 shows, there is a significant overlap between the crowd-provided keywords and the search terms, including over a thousand keywords that do not even occur in the set of author-provided tags. We did not expect there to be a full overlap, because the search terms are meant to search the *entire* database, while we only study a small part of the database.

Regarding quantity, table 2 shows that 72% of the keyword tokens provided by the crowd is in the set of search terms. With these numbers, crowdsourcing keywords seems a promising strategy to improve search recall in the Freesound database. More research is needed to assess the precision of the keywords.

4.6. Feasibility

Quality-wise, our overall impression is that the keywords are relevant for the sounds they are associated with. This is strengthened by the fact that there's a significant overlap between the crowd-tags on the one hand and the author-tags/search logs on the other. Furthermore, as we've seen above the crowd-labels complement the author-provided keywords really well in that they highlight different aspects of the sounds.

In total $190.46 was spent to gather 34,960 keywords for 2133 sounds. With an optimal cost of $0.09 per sound, it would cost at most $27,000 to crowdsource annotations for the full Freesound database containing almost 300,000 sounds. The total runtime to gather the data was 446 hours. By linear extrapolation, annotating the entire database would take 60,000 hours. It must be noted however, that this time can be significantly reduced by running multiple tasks at the same time. Also, the tasks will likely be completed faster because the crowd workers will be more experienced and motivated because of the large number of available crowdsourcing tasks.

Instead of annotating the entire database using our labeling task, we can also imagine a gamified sound-labeling solution similar to the ESP-game (Von Ahn and Dabbish, 2004). In the ESP-game, two players have to agree on appropriate labels for 15 images within 2.5 minutes, without communicating. They can only enter keywords, and they are notified when there is a match. With our corpus of annotated sounds, we now have a means to bootstrap an audio-ESP-game with 'gold' data, and a good basis to compare and evaluate the results. The big advantage of this approach is that it's a way to collect labels for free, and you have more control over the kind of labels that are used. The downside might me that it could take longer to collect the labels, and the labels might be qualitatively different because the game encourages users to keep variation as low as possible. (Free association is a bad strategy if you need to score points through agreement with someone else.)

5. Conclusion and Future Work

In sum, we have presented a large corpus of sounds annotated with low-level keywords from a listener's perspective. These annotations are complementary to the high-level keywords that were already in the Freesound database. Our main finding is that it is not only feasible to perform crowd-labeling for a large collection of sounds, it is also very useful to highlight different aspects of the sounds that authors may fail to mention. In short: *annotator perspective matters*. There is a large amount of tags that are never used by the authors that are still important for sound retrieval. Moreover, uninformed annotators are more likely to add tags that provide lower-level descriptions of the sounds. This might have to do with the different goals of the annotators. Strohmaier et al. (2012) make the distinction between *categorization* and *description*. If one's goal is to categorize a large set of sounds, one might be more tempted to use high-level descriptions. In our task, we explicitly asked for keywords to describe the sounds and this is exactly what was gathered. In addition, we might explain the difference with an appeal to the 'curse of knowledge' (Camerer et al., 1989): well-informed parties commonly overlook things that are obvious to them. And vice versa: uninformed parties can only provide superficial descriptions of the sounds, because they have no knowledge regarding the actual production.

Our resource is useful for people working on the relation between language and sound; in particular we hope that this corpus will spur the development of perceptually grounded distributional models, and models that can predict labels for any given sound. For now, the direct benefits lie in information retrieval, where we have shown that the crowd-annotations help to find the sounds that you are looking for.

6. Acknowledgments

Thanks to three anonymous reviewers for their comments. The first author is supported by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019), which is gratefully acknowledged. We also thank Frederic Font for providing the search logs.

Appendix: XML-format

Figure 4 shows the XML structure of our resource. We represent our data as a collection of sounds. Sounds have the following attributes: *id, batch, name, type, samplerate, duration, channels, bitrate* and *bitdepth* (the *id* and *name* attributes correspond to the ID and name in the Freesound.org database, and the *batch* attribute corresponds to the task batch in the crowdsourcing process, for full transparency about the data collection).

Sounds also have a number of elements: *file, uri, descriptions, webrating* and *author-tags* correspond to the Freesound.org metadata (with *file*-elements linking to high-quality MP3 and OGG files). The *crowd-tags* element contains the normalized tags as *tag*-elements, which in turn contain the raw tags that they subsume. The *ratings*-element provides information about the quality of the sound: *webrating* contains the user-rating from

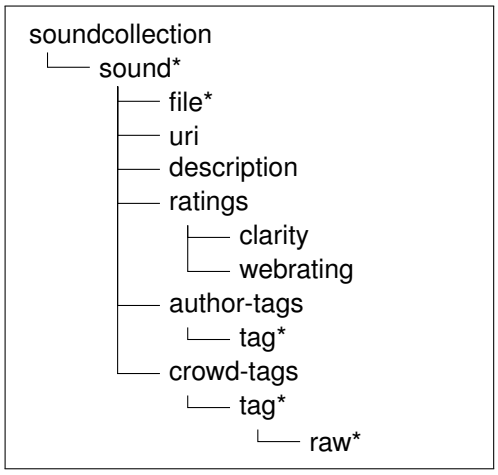

Figure 4: The XML structure of our resource. Elements that may occur multiple times are marked with an asterisk.

Freesound.org, and *clarity* contains the automatically generated clarity rating (based on the clustered tags).

7. Bibliographical References

Aroyo, L. and Welty, C. (2014). The three sides of crowdtruth. *Journal of Human Computation*, 1:31–34.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 1:1–47.

Camerer, C., Loewenstein, G., and Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *The Journal of Political Economy*, pages 1232–1254.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Dumitrache, A., Inel, O., Timmermans, B., Aroyo, L., and Sips, R.-J. (2015). Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In Artem Grotov, et al., editors, *Proceedings of the 14th Dutch-Belgian Information Retrieval Workshop (DIR)*, page 15.

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

Font, F., Serrà, J., and Serra, X. (2014). Audio clip classification using social tags and the effect of tag expansion. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society.

Hocking, J., Dzafic, I., Kazovsky, M., and Copland, D. A. (2013). Nessti: norms for environmental sound stimuli. *PloS one*, 8(9):e73382.

Kiela, D. and Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, September. Association for Computational Linguistics.

Kiela, D., Rimell, L., Vulić, I., and Clark, S. (2015). Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 119–124, Beijing, China, July. Association for Computational Linguistics.

Kiparsky, P. (1968). Linguistic universals and linguistic change. In Emmon Bach et al., editors, *Universals in Linguistic Theory*, pages 170–202. New York: Holt, Reinhart, and Winston.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Lopopolo, A. and van Miltenburg, E. (2015). Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 70–75, London, UK, April. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ogden, C. K. and Richards, I. A. (1923). *The meaning of meaning*. London: Kegan Paul, Trench, Trubner & Co.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC '00, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saygin, A. P., Dick, F., and Bates, E. (2005). An online task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods*, 37(1):99–110.

Strohmaier, M., Körner, C., and Kern, R. (2012). Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:1–11.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.