The Facilitating Role of Language

A Replication Study

Nels Acquistapace
Kyra Dunkel
Rosa Großmann
Johanna Linkemeyer

Department of Cognitive Science,
University of Osnabrück

Experimental Psychology Lab SS2020

Supervisor:
Prof. Dr. Michael Franke

Osnabrück, July 2020

ABSTRACT

Verbal labels, such as "dog" or "motorcycle" are more effective in activating conceptual knowledge than environmental sounds. Following Edmiston and Lupyan (2015), we hypothesize that this label advantage is due to environmental sounds, unlike words, being motivated cues that vary according to their source in a lawful way. In contrast to the original study, we found no significant evidence in favor of the hypothesis that the label advantages happens because of sound cues being motivated cues.

***K*eywords** Concepts · Language · Categories · Labels · Environmental Sounds · Replication

Contents

1 Introduction

Imagine you hear a very low-pitched bark while on a walk. If you find out that the sound came from a tiny Chihuahua, you will probably find yourself surprised. The sound of a bark not only tells you that the source of the noise is a dog, which can come in different shapes and sizes. It also allows for inferences being made about this specific instance of a dog based on the pitch of the bark, i.e. it being smaller or larger (Edmiston & Lupyan, 2015). The relationship between the dog and the pitch of its bark is predictable, because the sound varies according to the properties of what it is caused by in a lawful way (Kockelman, 2005). The type of predictive relationship between aspects of an object and the object itself is called "motivation" (Kockelmann, 2005), which led to cues that allow for inferences being made about the object referenced being called "motivated sounds" (Edmiston & Lupyan, 2015). Environmental sounds are typically motivated, i.e. the sound of rain hitting concrete in comparison to it hitting a roof sounds different, even though we know it is both the sound of rain. Or, as another example, the sound of an old telephone ringing in comparison to a new one. Humans have a lot of knowledge about such relationships. For example, it is possible to determine the shape and material of a hidden object based on the sound they make when dropped or hit (Carello, Anderson, & Kunkler-Peck, 1998).

Words on the other hand are unmotivated. The utterance "dog" can be used to describe any and all dogs (see figure 1). Hearing the word may allow for inferences about the speaker, but not about the dog being referred to (Edmiston & Lupyan, 2015). Therefore, the acoustic form of the word does not vary lawfully according to the properties of the object or event it refers to. So, any and all dogs can be referred to with the instances of the word "dog" in an abstract and categorical way by transcending the within-category variability common in motivated cues, i.e. barks (Edmiston & Lupyan, 2015). This renders words uniquely suited for activating mental states corresponding to categories (Boutonnet & Lupyan, 2015).

Edmiston and Lupyan hypothesized and tested in their 2015 paper "What makes words special? Words as unmotivated cues" whether this could be an explanation of the "label-advantage" found in previous studies (Boutonnet & Lupyan; 2015; Chen & Anderson, 2011; Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). In those previous studies, nonverbal cues resulted in slower recognition of subsequently presented pictures that represented the same category, in comparison to verbal cues. They further explain this could not be due to different levels of familiarity with certain cues, because both are arbitrary cues (Edmiston & Lupyan, 2015). That dogs bark has to be learned just as much as an English-speaker has to learn that the word "dog" refers to dogs. Further, the label-advantage can even be observed for novel categories, such as "alien-instruments", which come with new names and sound and subsequently equal familiarity for the participant (Lupyan & Thompson-Schill, 2012). Edmiston and Lupyan (2015) found that there is in fact evidence in favor of the motivation of environmental sounds being the reason behind the label-advantage and argue that it is due to a different and less direct pathway in the brain, where they activate specific category exemplars instead of general categories. Words on the other hand activate the general category directly (Edmiston & Lupyan, 2015).

Here we will reconstruct experiment 1A from Edmiston and Lupyan (2015). The experiment tests the hypothesis that environmental sounds fail to activate concepts as effectively as spoken category labels, because they are motivated cues, which activate specific category examples instead of general categories using a picture-verification task. If true, participants will demonstrate higher reaction times for trials with environmental cues compared to trials with verbal labels. Further, the reaction times in trials where the environmental sound matches the exact subcategory (congruent trials) will be lower than for trials where the environmental sound matches the picture in category but not in the specific example (incongruent). A congruent trial is a trial, where the sound, i.e. the strum of an acoustic guitar, matches the picture in the subcategory, so the picture of an acoustic guitar. In contrast, an incongruent trial is one where the sound and the picture belong to the same category but different subcategories. This would be the case if the sound is a strum of an acoustic guitar, but the picture is of an electric guitar. By demonstrating that environmental sounds activate conceptual knowledge differently than words because of their ties to their particular source, we hope to better understand the unique way language influences perception and its abilities to manipulate mental content.

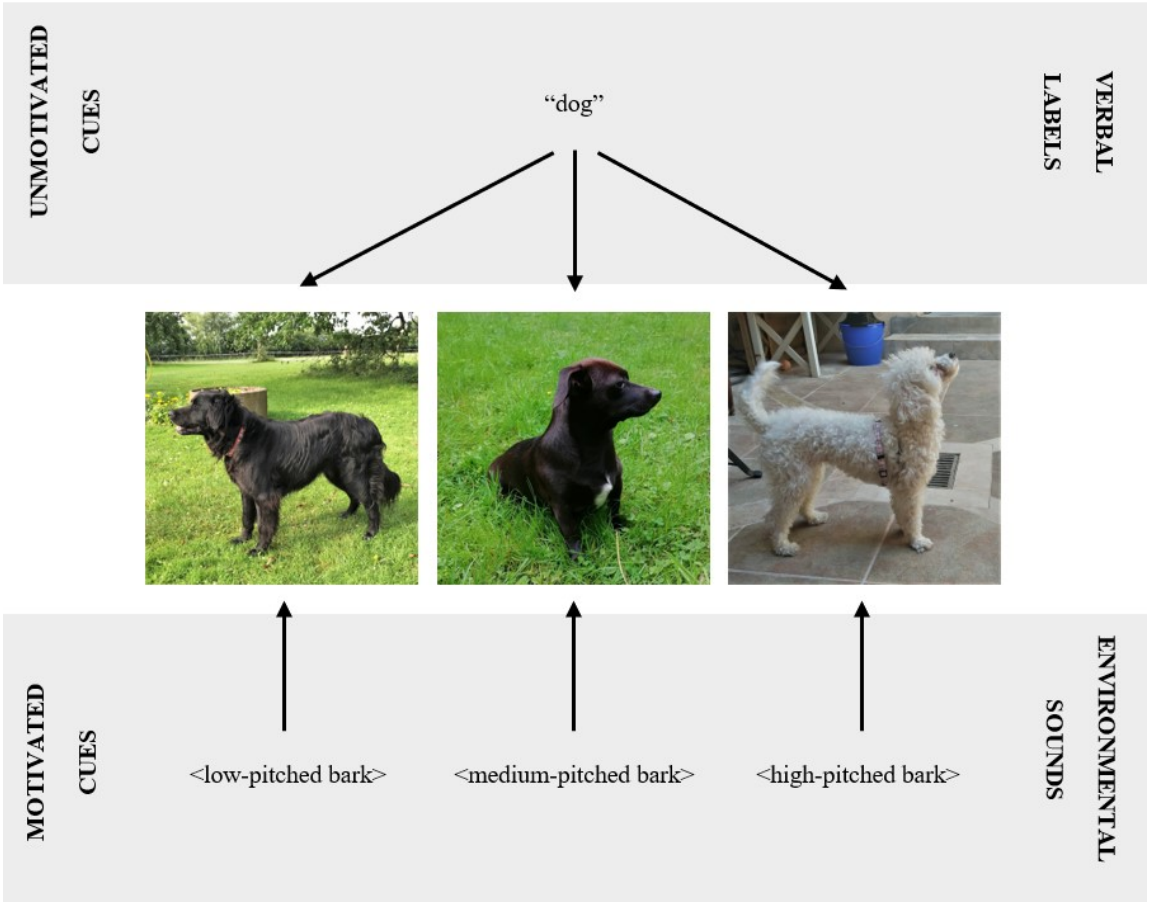


Figure 1: Examples of unmotivated and motivated cues for the category dog. The unmotivated cue (verbal label) "dog" refers to all three breeds (Labrador mix, Dachshund mix and Miniature Poodle). There is no specific pronunciation of the word "dog" in order to refer to a specific breed. The motivated cues (environmental sounds), the differently pitched barking, vary for each breed represented.

2 Method

This experiment is a picture-verification task modeled after Edmiston and Lupyan (2015), in which the participants had to recognize basic-level categories of pictures as fast and accurately as possible. We hypothesized that environmental sounds would lead to slower recognition compared to labels because they cue a more specific category exemplar, making them not as effective as labels in cueing for an entire category. It was broken into two hypotheses:

1. Sound cues are not as effective as category labels in activating a concept. That is, reaction times for trials with environmental cues are higher than for trials with verbal labels.

2. Sound cues activate specific category exemplars instead of general categories. That is, reaction times for congruent trials will be lower than for incongruent trials.

These hypotheses were tested by manipulating the within-category specificity for the mapping between the cues and picture targets.

2.1 Participants

In total, 40 participants took part in this study. They were recruited through email and personal contact and did not receive compensation of any kind in return for their participation. The experiment was conducted as an online experiment, wherefore participants took part on their own devices. The average age of the participants was 25.21.

2.2 Materials

Materials are four pictures and four auditory cues for each of the six categories *bird, dog, drum, guitar, motorcycle*, and *phone*. Auditory cues consist of two verbal labels and two environmental sounds. Of the verbal labels, one is spoken by a male and one by a female. The environmental sounds are of two distinct subcategories of the respective category. For example, environmental sounds for the category dog are <low-pitched bark> for the subcategory Labrador mix and <high-pitched bark> for the subcategory Miniature Poodle. All auditory cues are normalized in volume and equated in duration (600ms). Pictures are color photographs in square format (see figure 6 in the appendix). For each category, there are two pictures for each subcategory. For example, the subcategories of the category *bird* are *owl* and *sparrow.* So, for the category bird there are two pictures of owls, and two of sparrows. Materials are available for viewing-only purposes here.

2.3 Procedure

Each trial starts with an auditory cue,followed by a picture. Participants had to answer as quickly and accurately as possible whether the two stimuli were of matching categories. The experiment was implemented as an online experiment, wherefore participants were tested at home on their laptop or computer. Picture size was automatically adapted to screen size and participants were asked to sit away from their screen such that the screen was fully visible in their visual field. Participants were instructed to wear headphones if possible and to adjust their volume during the practice trials and not to change it during the main trial phase.

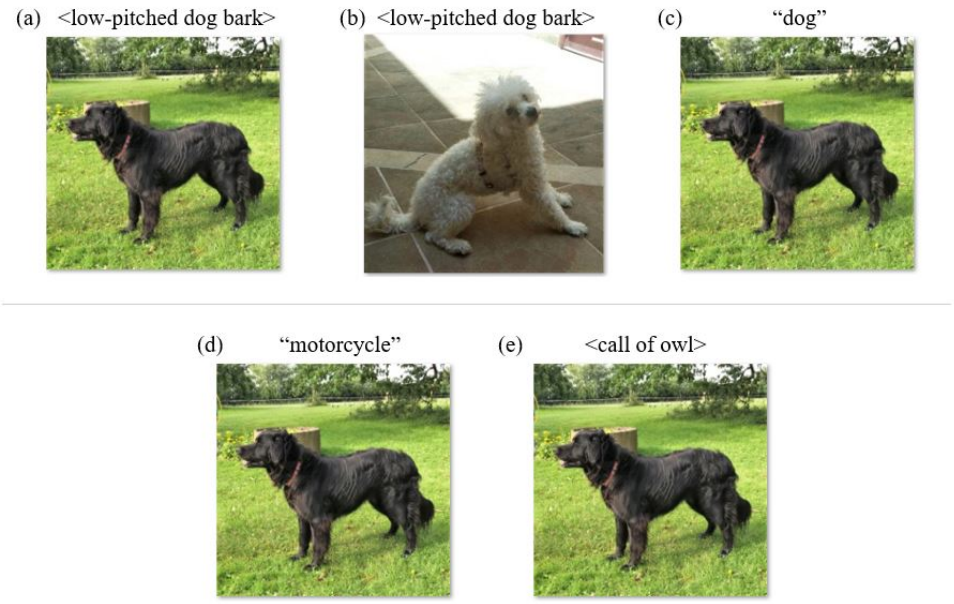


Figure 2: Different trial types for the picture category dog. Do these auditory cues match the category shown in the picture? Verbal labels are written in quotation marks, environmental sounds in angle brackets. The upper row shows trials where auditory cue and presented picture have matching categories. The row below shows trials where auditory cue and presented picture do not represent matching categories. (a) is an example for a congruent combination of stimuli. (b) represents an incongruent combination. (c) is an example for a matching label-picture combination. (d) shows an example of a non-matching label-picture combination and (e) a non-matching sound-picture combination.

Each trial started with a 500ms break, followed by a 250ms fixation cross. Participants then heard an auditory cue. There was a 1s pause before a picture appeared centrally in square format on the screen. The picture disappeared only after a response was made by the participant. Responses were made by pressing the keys "y" or "n", representing "yes" and "no" as answers to the question, whether the two stimuli of one trial had matching basic-level categories. For example, participants had to press "y" when hearing <low-pitched bark> and seeing the picture of a dog. In contrast, participants had to press "n" when hearing, as an example, a <cell phone ring> and seeing the picture of a bird. There were 6 practice and 192 main trials, which both had a 50:50 division to matching and non-matching trials. The trials with the varying factors cue type and sound type were presented in random order to each participant. In each trial, participants received auditory feedback (buzz or bleep) depending on the correctness of their answer. During practice trials, the auditory feedback was accompanied by visual feedback in the form of the words "correct" and "incorrect". This served to familiarize the participants with the auditory feedback.

3 Results

Forty participants took part in the experiment which was online for six days. During the first two days of data collection, some participants reported technical problems via the comment section or personally. Pictures and sounds

were presented simultaneously, or the sound failed to appear. We were able to detect a link of these problems to the Safari browser. Therefore, the instruction not to participate in the experiment using Safari was added. We decided to exclude participants that left a comment about technical issues in the post-test survey. This held for one participant. Because it is not possible to link the personally reported technical issues to specific data sets, we decided to remove all data from participants that showed an accuracy below the accuracy of the participant who experienced the described technical problems (92%). In total, data from 8 participants was excluded. Therefore, data from 32 participants remained for the statistical analysis.

3.1 Main Analysis

The following analysis is based on the original paper by Edmiston and Lupyan (2015) as well as their code of the statistical analysis provided on GitHub. As explained in the preregistration report, our experiment is a 2x2 factorial design with the factor cue type (with the levels environmental sound and verbal category label) and the factor sound type (with the levels congruent and incongruent environmental sound trial). With our two hypotheses the effect of these two factors on the reaction time, which is the dependent measured variable, is investigated. Figure 3 shows that the reaction time of participants differs dependent on the different conditions (sound congruent trials, sound incongruent trials and verbal label trials) which lets suggest a correlation between the two factors *cue type* and *sound type* and the verification speed.

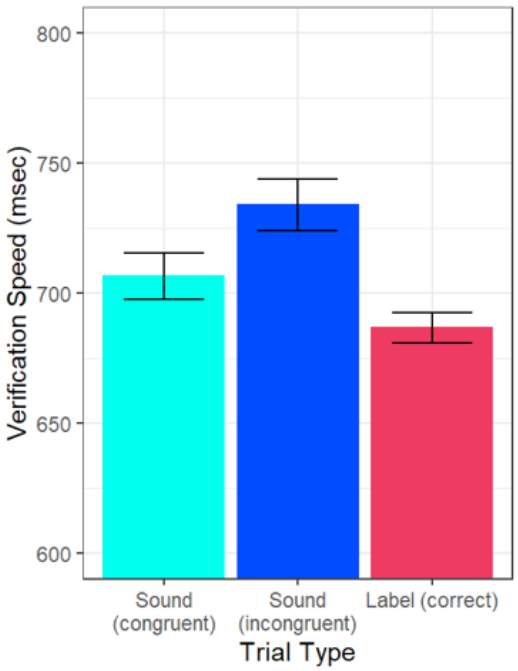


Figure 3: Means of the verification speed (ms) grouped by the trial type.

In order to test the two hypotheses stated above, we used linear mixed effects regression models that include our factors of interest as fixed effects (type of cue and type of sound). The two models also account for random effects,

based on recommendations from the paper "Random effects structure for confirmatory hypothesis testing: Keep it maximal" (Barr et. al., 2013). This means that we want to keep the maximal complex model structure, without including unnecessary variables. The random effects account for random intercept and random slope effects of within-subject factors, as well as random intercept effects for unique trial types. Thus, the random effects allow the predictors - our two separated factors in our two regression models - to vary by subject and by item (since our design includes unique trials). For a more detailed explanation of the model structure, please refer to the analysis files in the GitHub repository. We use the lme4 package (Bates D, Mächler M, Bolker B, Walker S, 2015) to fit the linear mixed effects regression models. Additionally, to our two complex models that include the factors sound type and cue type, we build two simpler models that don't include the two factors as predictors. We use the chi-square test for comparing the simpler model with the complex model and use the *p*-value to check if the complex model (that accounts for an effect of the factor cue type or sound type) fits the underlying data significantly better than the simpler model that accounts for a zero-correlation of the factor. We set the significance level $\alpha = 0.05$, thus a sufficient evidence for a non-zero correlation between the factors cue type and sound type would require a *p*-value that is below the significance level ($p < 0.05$). The model fitting is done with log-transformed reaction times as dependent variable.

To test the first hypothesis, we compare the linear mixed effects regression model that includes the factor cue type with a model that differs from the more complex model only in that it does not include the factor cue type. The type of cue can be either an environmental sound or a word label describing a (basic-level) category. We test if the type of the cue has an effect on the verification speed of the participant. The reaction time is expected to be higher for trials with environmental cues than for trials with verbal labels. The results show that the effect is negative as expected - the level "verbal label" is dummy coded as -0.5, thus the effect is negative when multiplying it with a positive parameter b – (b= 0.02, Std. Error: 0.01, 95% CI [-0.00, 0.05], $\chi^2(1) = 2.99$, $p = 0.083$), but the *p*-value is slightly higher than the significance level. Therefore, we can conclude that the type of cue does not significantly influence the verification speed of the participants and that verbal labels are as effective as (congruent) environmental sound cues. Still the *p*-value is very close to be significant which indicates a probableeffect of the type of cue on the verification speed. It is important to remark that the small number of participants leads to less powerful results. If we would have collected more participants, the *p*-value might have been below the significance level and reveal the significant effect of the type of cue on the reaction time. It is also possible that the *p*-value stays above the significance threshold and that there is not enough evidence against a zero correlation of the factor and the dependent variable.

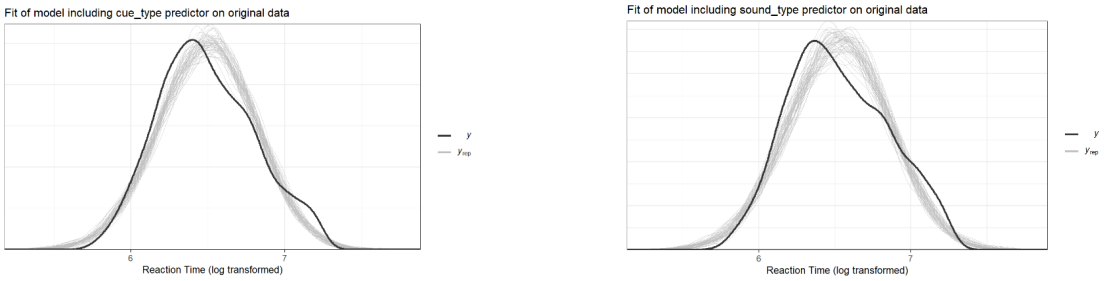
The second hypothesis, stating that sound cues activate specific category exemplars instead of general categories, is tested in a similar way to the first. We compare the linear mixed effects regression model that includes the factor *sound type* with a model that differs from the more complex model only in that it does not include the factor *sound type*. The type of sound can be either a congruent sound or an incongruent sound. The reaction time is expected to be lower for congruent sound trials than for incongruent sound trials. The resulting parameter estimation seems to support this assumption (b = -0.04, Std. Error: 0.02, 95% CI [-0.09, -0.00], $\chi^2(1) = 5.43$, $p = 0.142$). The congruent sound is coded as 0.5, meaning that there is a negative effect on reaction times when the sound type is congruent. Though, the results show that there is not enough evidence for a non-zero correlation between the type of the sound and the reaction time (*p*

= 0.142). The more complex model does not fit the data significantly better than the simpler model that accounts for a zero correlation between the factor *sound type* and reaction time.

To be critical with our statistical results and their interpretation we refer to our power analysis that we did before publishing the online experiment. Since we had to exclude participants from our data, the sample size is small and therefore the power of the results is insufficient to claim that there is no effect of the factor *cue type* and *sound type* on the reaction time. A larger number of participants could lead to a different p-value and reveal an effect of one or both factors on the dependent variable. Referring to the power analysis we computed before the analysis was done, more than 50 participants would be required to achieve a power of 95%. For comparison: the original study had a number of 43 participants (power of 0.88), whereas we only had 32 valid participants, thus there is only 78% chance that the results of our hypotheses tests are meaningful.

3.2 Exploratory Analysis

This part of the analysis follows the preregistration report but is not based on the procedure of the original paper. In order to explore the model performance more visually we use the brms package by Paul Buerkner (2016) to fit a Bayesian regression model. Additionally, it enables us to check how likely the effect of our two factors is. Figure 4a and 4b show the plotted posterior distribution of the two Bayesian regression models including the factors *cue type* and *sound type*.



(a) Posterior distribution of the model using cue type as predictor (b) Posterior distribution of the model using sound type as predictor

Figure 4: Posterior distributions of the regression models including our two factors cue type and sound type

The model that includes the factor sound type has greater deviations from the data than the model that includes the factor cue type. This suggests that the effect of *cue type* on the reaction time of the participants is more present in the underlying data than the effect of the factor *sound type*. The results of the comparison between the two levels of the factors show, that if there is a non-zero correlation between the type of cue and verification speed, it is likely (95% chance) that environmental sounds lead to higher reaction times that verbal labels. If there is a non-zero correlation between the sound congruency and verification speed, it is very likely (98% chance) that congruent sound trials lead to lower reaction times than incongruent sound trials. There are less deviations of the posterior distributions from the underlying data in comparison to the results of the pilot study, but it is visible that the fit of the two models is not

perfectly representing the data. Thus we can not be sure that the effect of the two factors on the verification speed is actually present in the underlying data.

Furthermore, we tested for a correlation between the categories and the reaction time in congruent sound trials. Our assumption is that participants tend to find it easier recognizing the sound of a certain category faster in comparison to other categories. Figure 5 shows the reaction time means grouped by category for congruent sound trials.

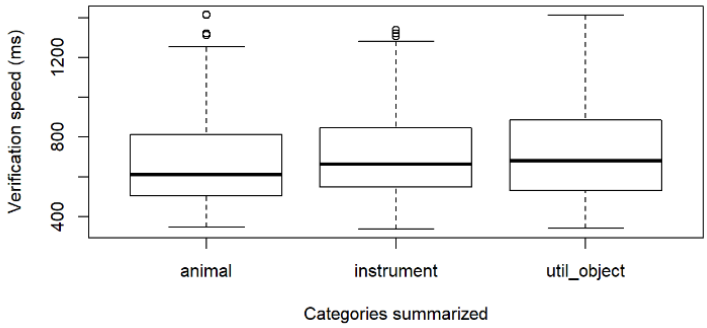


Figure 5: Means of the verification speed (ms) grouped by summarized categories.

One can observe that the mean of the animal categories is slightly lower than for the other categories. Therefore, we would explicitly like to know if animal sounds lead to faster reaction times than sounds belonging to instruments or utility objects. The results show that it is likely (88% chance) that the sound of an animal is faster to recognize than the sound of an instrument and very likely (94% chance) that an animal is faster to recognize than the sound of an utility object such as a phone or a motor cycle. The chance that utility objects are harder to identify than instrument categories is lower at 72% chance. Investigating the parameter estimation for the factor category one can observe that there is not enough evidence for a non-zero correlation between categories and reaction time (*intercept* (*animal*) = 6.48, CI = [6.39, 6.56], *instrument* = 0.04, CI = [-0.024, 0.1], *util_object* = 0.06, CI = [-0.01, 0.125]). If we would add more categories to our experimental design this could of course change the results of this hypothesis test, since we make between category caparisons. In order to be critical with our model, we observe its posterior distribution and compare it to the underlying data. The plot shows that the model roughly fits but, in some points, deviates from the data. Because of the insufficient goodness of fit we cannot make statistically reliable conclusions about the effect of different categories on the verification speed.

4 Conclusion

In this replication study, the hypotheses could not be proven to be true. We were not able to show that the label-advantage persists when environmental cues are matched to the visual targets at a subordinate level. This, however, does not change the fact that environmental sounds are motivated cues and verbal labels are not. This means that, according to our results, the fact that sound cues are motivated cues is not the reason why category labels are more effective in

activating a concept. However, we could also not prove to find evidence in favor of this label advantage. On our account, no label advantage can be obtained at all, which would make the second hypothesis irrelevant as it builds on the first.

While these results might in fact be accurate, they may as well be the result of problems that aroused during the execution of the experiment. Eight datasets have been excluded completely because of major technical issues. There still is a chance, that for the data sets of the remaining 32 participants, reaction times might have been influenced by slow internet connections and slowness of the connection between netlify and github. As this experiment was conducted online, there was no possibility for us to guarantee uniformity in those external factors influencing reaction time measurements.

Further, as we only had 32 remaining data sets, the power of the results is insufficient to claim that there is no effect of the cue type and the sound type on the reaction time. A larger number of participants could lead to a different p-value and reveal an effect of one or both factors on the dependent variable. Referring to the power analysis that we computed before the analysis was done, we would need more than 50 participants to achieve a power of 95%. We only had 32 valid participants, thus there is only a 78% chance that the results of our hypotheses test is meaningful.

One major weakness of our study was that it was done as an online experiment. During such, parameters, such as volume of the sounds, size of the pictures, distance of the participant to the screen, etc. can not be controlled like in a lab study, as was the case with the original study done by Edmiston and Lupyan (2015). Additionally we did not use a controller to indicate the response but a normal keyboard. We also shortened the experiment, so that each participant only saw each possible trial once, whereas in the original experiment, every trial was shown twice. This was necessary for an online study, because the original length of the experiment (about 30 minutes) left the participants of our pilot study fatigued and we feared participants would not finish the experiment.

Even though our study could not prove that a label advantage exists, there are several previous studies by other researchers that found evidence in favor of it (Boutonnet & Lupyan; 2015; Chen & Anderson, 2011; Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). This does include the study by Edmiston and Lupyan (2015), whose experiment 1A we replicated. Contrary to our results, they also found evidence that the label advantage exists because sound cues are motivated and labels are not. However their original study had a number of 43 participants (power of 0.88). While this makes their results more likely than ours; our power analysis revealed that more than 50 participants are needed to achieve a power of 95%. So, there is the possibility that the results found in the original study might not be correct either.

Further research needs to be done in this field in order to confirm the label effect does in fact exist and whether it occurs due to different pathways taken in the brain while processing. Since most previous studies involved a similar experimental design, we would like to suggest that future experiments involve the use of a different approach than what was used in this study. Using just a picture-verification task might not be enough to give definitive answers to the question in which way language influences perception and its abilities to manipulate mental content and if it is different from the way environmental sounds do.

References

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013, 01). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*, 255-278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Boutonnet, B., & Lupyan, G. (2015, 06). Words jump-start vision: A label advantage in object recognition. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *35*, 9329-35. doi: 10.1523/JNEUROSCI.5111-14.2015

Carello, C., Anderson, K., & Peck, A. (1998, 05). Perception of object length by sound. *Psychological Science*, *9*, 211-214. doi: 10.1111/1467-9280.00040

Chen, Y.-C., & Spence, C. (2011, 06). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of experimental psychology. Human perception and performance*, *37*, 1554-68. doi: 10.1037/a0024329

Edmiston, P., & Lupyan, G. (2015, 06). What makes words special? words as unmotivated cues. *Cognition*, *143*, 93-100. doi: 10.1016/j.cognition.2015.06.008

Franke, M., Ilieva, S., Ji, X., & Rautenstrauch, J. (2019, Jul). *_magpie - minimal architecture for the generation of portable interactive experiments.* Retrieved July, 2020, from https://magpie-ea.github.io/magpie-site/

Ivanova, A., & Hofer, M. (2020, 05). Linguistic overhypotheses in category learning: Explaining the label advantage effect. *preprint*. doi: 10.31234/osf.io/x9e4z

Kockelman, P. (2005, 01). The semiotic stance. *Semiotica*, *2005*, 233-304. doi: 10.1515/semi.2005.2005.157.1-4.233

Lupyan, G., & Thompson-Schill, S. (2012, 09). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of experimental psychology. General*, *141*, 170-86. doi: 10.1037/a0024904

Toon, J., & Kukona, A. (2020, 01). Activating semantic knowledge during spoken words and environmental sounds: Evidence from the visual world paradigm. *Cognitive Science*, *44*. doi: 10.1111/cogs.12810

Waxman, S., & Gelman, S. (2009, 06). Early word-learning entails reference, not merely associations. *Trends in cognitive sciences*, *13*, 258-63. doi: 10.1016/j.tics.2009.03.006

A Appendix

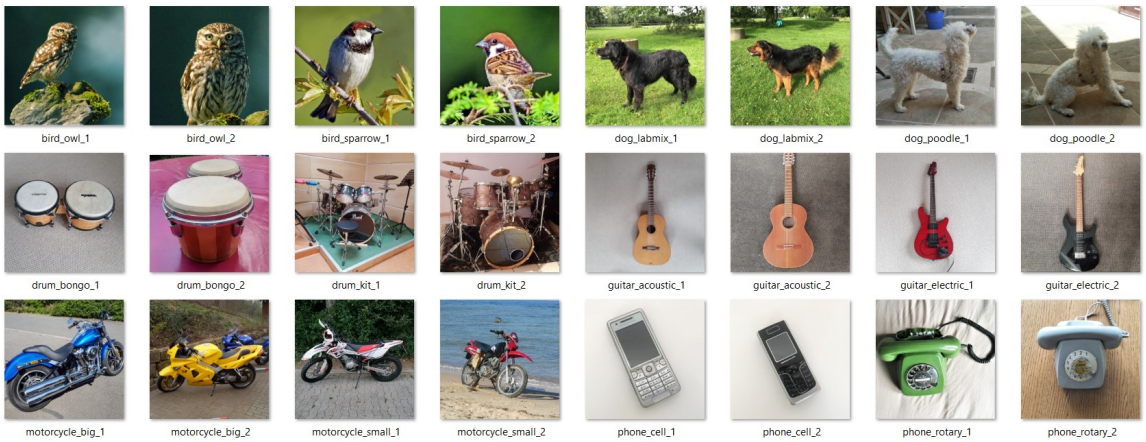


Figure 6: Collection of all pictures used in this experiment.