The Facilitating Role of Language - Preregistration Report

Study Information

Titel: What makes words special? Words as unmotivated cues

Authors: Nels Acquistapace, Kyra Dunkel, Rosa Großmann, Johanna Linkemeyer

Description: We reconstruct an experiment conducted by Pierce Edmiston and Gary Lupyan (2015) on the facilitating role of language. Participants will hear an auditory cue that is either a spoken label, i.e. ‘dog’ or the natural sound of an object or animal, i.e. a dogs bark. After a short delay, a picture is presented. Participants have to decide whether the previously heard auditory cue matches the category of the shown object or animal.This experiment aims to investigate if environmental sounds activate concepts of the sounds source, while verbal labels and environmental sounds both can be unambiguous representation for a specific category.

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.

Design Plan:

Study Type: Online Experiment

Blinding: The relevant manipulation is within-participants. Participants are not informed about this manipulation. The experiment is conducted via the internet, no direct contact between experimenter and participants will take place.

Study design: The experiment is a within-subjects study with two factors with two levels each. One factor is the cue_type with the levels label (a spoken category label, i.e. the word “dog”) and environmental cue (i.e. a deep dog barking). The second one is sound_type with the two levels congruent (the auditory cue and the visual cue shown display the same category and the same instance) or incongruent (they do display the same category, but not the same instance). For example the picture of a labrador with the sound of a low pitched bark would be congruent, while the same bark with a picture of a small poodle would be incongruent.

The participants task is to indicate via button press if they think that the category of the auditory cue matches the picture of category subgroup instances (e.g. the spoken word “dog” followed by a picture of a labrador would be a matching category).
A full description is given in the attached document “Design for an Experiment on the Facilitating Role of Language”.

Randomization: All participants see half of all experimental items, meaning each participant does 96 trials. This is done in order to avoid fatigue. The decision on which participant sees which of the items and in which order is decided in completely, ad hoc determined random order.

Link to experiment files:
https://github.com/NelsAcquistapace/XPLab_SpecialWords_magpie/tree/master/materials

Sampling Plan:

Existing data: Since this experiment is a replication of the experiment 1A done by Edmiston and Lupyan (2015), data of said previous study was available and guided the specification of statistical models. This data will not be included in the final analysis.
Data from a previous pilot study (N=4) was also available and guided the specification of the statistical models. This data also will not be included in the final analysis.
As of the date of preregistration, the data from the experiment to be preregistered have not yet been collected, created, or realized.

Explanation of existing data: Existing data will not enter into future analysis.

Data collection procedures: Participants will be drafted through social media and direct email contact. Participation is voluntary and will not be compensated. We will wait 6 days after having sent the initial invitations through social media and email before closing data collection.

Sample size: We performed a power analysis based on our plan for the statistical analysis. Since we will use the chi-squared test for the inference of statistical significance - under the section “Analysis Plan” we explain how we are doing this - we used the R function **pwr.chisq.test(w,N,df,sig.level, power)** to find an appropriate number of participants. The function takes 4 of 5 values that are the effect size **w**, the number of participants **N**, the significance level **sig.level**, the degrees of freedom **df** and the **power** value.
To see how the statistical power of the chi-square test changes, we entered a low, medium and high value for the sample size, based on the original number of the participants, which was 43.
For the power analysis we set **w** (effect size) to 0.5 which was calculated from the results of the original paper. We set the degree of freedom equal to 1 and the significance level equal to 0.05 based on the original analysis.

sample size	30	40	50
power	0.78190799873198	0.885379140762351	0.942437543187508

Table: Power of our result according to number of samples

Now we will first set the power we would accept for our statistical results and check how many participants we would need. We chose a power of 80% which is acceptable but rather low, 95% as medium value and 99% as a very strong power value.

power	0.80	0.95	0.99
sample size	31.3954421253175	51.9788369797701	73.4898787237435

Table: Sample size needed if we want to have a certain power

The power analysis is the same for both hypotheses we test. We will try to achieve the highest sample size possible.

Sample size rationale: We will recruit subjects based on voluntary participation and do not have the resources to pay for participation or offer course credits, thus we do not expect reaching the number that is required for a power of 0.99.

Variables

Manipulated variables: As explained above the experiment has a 2x2 factorial design. We manipulate the variable cue_type and the variable sound_type.
The cue_type can be an environmental cue or a category label. The spoken category labels are basic-level category labels for each of the six categories (bird, dog, drum, guitar, motorcycle, and phone). The environmental cues are sounds that originate from the instances of the category subgroups (e.g. a ringing rotary phone). By manipulating the cue_type we can compare if the participants verify the matching of picture and auditory cue faster when they are given a verbal category label instead of a congruent environmental cue. Furthermore we manipulate the congruence of the sound_type, that tells us if the environmental sound fits exactly to the shown picture of the category instance. This means that we have two different natural sound auditory cues that belong to a subgroup of the category (e.g. the sound of an electric guitar and the sound of an acoustic guitar). The two pictures of each category differ in a visual way (e.g. acoustic guitar and electric guitar). Thus the sound_type (the environmental sound cue) can be a congruent environmental sound, that fits exactly to the picture, which shows an instance of the category subgroup (e.g. hearing an acoustic guitar strum and seeing a picture of an acoustic guitar) or an incongruent environmental sound, that fits to the category but not to the instance of the category subgroup (e.g. hearing an acoustic guitar strum and seeing a picture of an electric guitar). To sum up, we manipulate the cue_type (environmental sound or category label) and we manipulate the congruence of the environmental sound_type (congruent, incongruent).

Measured variables: We measure the reaction time from the onset of the picture and button press, that indicates the participants decision about the matching of auditory cue and displayed picture. The variable RT is a metric variable capturing reaction times. We also measure whether the participants response was correct or not (binary choice), making the variable CORRECTNESS a discrete binary variable, registering whether the choice of a trial was correct or not with default reference level.

Analysis Plan

Statistical models:

Only matching trials (category of the auditory cue matches the category of the depicted instance) with correct answer by the participant are included in the following analysis. Like stated in the hypothesis section we would first like to know if congruent environmental sound cues are not as effective as category labels in activating a concept, which means that we expect lower reaction times for trials where the participants were given a category label in comparison to a congruent environmental sound cue.

Further we would like to know if the second factor, the sound_type plays a role in verification performance. This would mean that reaction times for congruent trials will be lower than for incongruent trials.

To examine the two effects explained above we will fit two mixed effects linear regression models that try to explain the data more precisely than two basic models that do not include the factors cue_type and sound_type.

All models include random effects that are modeled based on the original paper. Here, Lupyan and Edminston oriented themselves on the research from the paper “Random effects structure for confirmatory hypothesis testing: Keep it maximal” (Barr et. al., 2013), that accounts for choosing the maximal random effect structure for a model that investigates psycholinguistic effects. Since the experiment is a within-subject design the random intercept and random slope effects of within-subject factors are included. This means that we assume the strength and the intercept of the effect (of the type of sound and the type of cue) varies between the participants. All participants are measured in trials that are unique, so the reaction time of verification might vary by trial item. Therefore we include a random intercept effect for unique trial types (the variable trialID accounts for the unique trial types).

For fitting the regression models we will use the lmer function of the lme4 package and use the lmer() function that computes the mixed effects linear regression model via maximum likelihood estimates.

In order to check if the linear regression models that take the effect into account (e.g. we investigate if the cue_type is relevant for the model to fit the data), we build a second linear regression model that does not include the variable of interest, thus it is an intercept only model. We will mainly keep the random effect structure for the simpler model but exclude the random slope effect that refers to the factor of interest (which is not included in the simpler model). We do this because in case of this simpler model we do not expect cue_type or sound_type to have any effect. Based on the original data set we checked for singularity of the model parameters with the r function isSingular(), that tells us if one or more parameters are very close to zero and thus not needed in our model. In this way we can keep the maximal structure of our model and still exclude unnecessary parameters. Singularity was

not present based on the original dataset, including the mentioned random effects, thus we will keep the structure as described above. In our pilot study analysis the random slope effect was assumed as singular, which might be due to the very few number of participants. To sum up this means that in total we have 4 linear mixed effect models that are all computed using the R function lmer() mentioned above. Two for each effect that we investigate:

1. Sound cues are not as effective as category labels in activating a concept. That is, reaction times for trials with the cue_type environmental cue are higher than for trials with the cue_type verbal label.
RT ~ cue_type + (1+cue_type | participantID) + (1 | trialID)
RT ~ 1 + (1 |participantID) + (1 | trialID) → control model for comparison

2. Environmental sound cues activate specific category exemplars instead of general categories. That is, reaction times for congruent sound_type trials will be lower than for incongruent sound_type trials.
RT ~ sound_type + (1+sound_type | participantID) + (1 | trialID)
RT ~ 1 + (1 | participantID) + (1 | trialID) → control model for comparison

To compare the two models for each hypothesis we use the anova() function in r that calculates the chi-square and the p-value which lets us test the goodness-of-fit of the more complex model in comparison to the simpler model. If the p-value is less than 0.05 it tells us that the parameter of interest plays a role in predicting the verification time so it is not 0 as it is the case in the simpler (nested) model that does not include the key variable (e.g. cue_type or sound_type). Thus the more complex (nesting) model that includes the parameter of interest is most probable better in explaining the observed data, if the p-value is below the significance level.
Coding scheme for categorical variables: We use the simple (contrast) coding scheme to dummy code our factors cue_type and sound_type, which each have two levels. The parameter sound_type is coded as 0.5 for congruent and -0.5. The parameter cue_type is dummy coded as -0.5 for verbal level and 0.5 for environmental cue. The same dummy coding procedure was applied in the original paper.
For hypothesis 1, if the p value is below the set significance level (α=0.05) and the estimated parameter for the factor cue_type (computed by the lmer() function) is positive, we can claim that the reaction time is higher for environmental cue_types than for verbal label cues. For hypothesis 2, if the p value is below the set significance level (α=0.05) and the estimated parameter for the factor sound_type (computed by the lmer() function) is negative, we can claim that the reaction time is lower for congruent sound_types than for incongruent sound_types. Additionally we compute the 95% confidence interval to check in which range and how precise the parameter can estimated.

Transformations: The distribution of reaction times is similar to a normal logarithmic distribution, we perform a log-transformation on the raw reaction times. We do this in order to get a more normalized distribution that helps us to do a more reliable parameter estimation, with the assumption that the data is normal distributed. Refer to the markdown

file containing the statistic analysis of the pilot study to have a look at the reaction time distribution.

Inference criteria: For the evaluation of the chi-squared test outcome we will use a significance level of **$\alpha=0.05$**

Data exclusion: To estimate which data points have to be excluded we plot a box plot of all measured reaction times in the main trails. We take the value of the extreme of the upper whisker as maximum value and the extreme of the lower whisker as minimum value. We will remove data points that are above the max_RT and those that are below the min_RT. We decided to not use the limitation values of the original paper, because the reaction time distribution in the pilot study and the following study with more participants can be different due to varying technical conditions (our experiment is an online experiment where it is not given that every participant uses the same computer with the same internet connection). Using the boxplot whiskers to set the limits of the values allows for a fixed but technically appropriate alternative.

Missing data: Should a data set not be recorded completely, we will not use any data available from that participant.

Exploratory analysis: We would like to investigate the more complex models above by performing a bayesian regression approach, using the “brms” package by Paul Buerkner (2016). It allows for more detailed information output, such as the posterior distribution that can be plotted and compared. We will plot the posterior distribution of the two complex models (from hypothesis 1 and 2) in order to explore and to compare them to the underlying data visually. We will use the compare_groups() function of the “faintr” package to check how likely the effect of our two factors on the reaction time is. We also use the log-transformed reaction times in the exploratory part of the analysis.

As additional exploratory hypothesis we would like to ask if the category of the trials - since we only look at matching trials, the category of sound and pictures are equivalent - influence the verification time. This is the case when people can identify a sound belonging to a certain category faster than if a sound that belongs to another category. Is this effect randomly depending on person, or is it systematically the same for all subjects? In our experiment, we have 6 categories, which can be summarized into animals (dogs and birds), instruments (drums and guitar) and everyday objects (telephone and motorbike). We plot the categories as well as the summarized categories against the reaction time in congruent environmental sound trials. Note that for these two plot we use the original non transformed reaction times. For the following analysis we will only use trials where the sound is congruent to the picture and the participants answer was correct.
We build the model:
RT ~ category_sum + (1+ category_sum | participantID) + (1|trialID)
The model includes the random slope effect of within-subject measurement (because some participants could be more familiar with certain categories than other participants. If the parameter estimation of a category is 0, the category would not have an effect on the dependent variable verification speed (RT).

We will fit the model using brm() and investigate the parameter estimation for the summarized categories “animal”, “guitar” and “utility object”. To investigate how likely it is that one category leads to a lower or higher reaction time than another category, we use compare_groups(). Because we only compare the categories with each other, we can only draw conclusion on the categories that are included in our model and not about all possible categories. Still the result is interesting. If one or more of the categories affect the reaction time decisively we can say that it might be crucial for the comparison between incongruent and congruent sound trials - like we did in hypothesis 1 in the main part of the analysis- to include the categories in our formula. Further analysis of the correlation of category and sound type (congruent and incongruent) is conceivable.

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.

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

Buerkner, P. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80.1, 1-28.