Superordinate referring expressions in abstraction: Introducing the concept-level reference game

Kristina Kobrock (kristina.kobrock@uni-osnabrueck.de)

Charlotte Uhlemann (chuhlemann@uni-osnabrueck.de)

Nicole Gotzner (nicole.gotzner@uni-osnabrueck.de)
Institute of Cognitive Science, Wachsbleiche 27
49090 Osnabrück, Germany

Abstract

We study referential communication about concepts at different levels of abstraction in an interactive concept-level reference game. To better understand processes of abstraction, we investigate superordinate referring expressions (*animal*). Previous work identified two main factors that influence speakers' choice of referring expressions for concepts: the immediate context and the basic-level effect, i.e. a preference for basic-level terms such as *dog*. Here we introduce a new concept-level reference game that allows us to study differences in the basic-level effect between comprehension and production and to elicit superordinate referring expressions experimentally. We find that superordinate referring expressions become relevant for groups of objects. Further, we reproduce the basic-level effect in production but not in comprehension. In conclusion, even though basic-level terms are most readily accessible, speakers tailor their expressions to the context, allowing the listener to identify the target concept.

Keywords: reference game; concepts; categorization; superordinate level; abstraction

Introduction

Concepts allow us to make sense of the world. They help us to structure and organize knowledge, and to generalize from one instance to a class of objects that share similar properties through a process that is commonly called "abstraction" (Yee, 2019; Rosch, 1978). We use referring expressions at different levels of abstraction, ranging from subordinate terms like *dalmatian* to superordinate ones like *animal*, to communicate about concepts at different levels of abstraction.

Previous work suggests two main factors that influence the choice of referring expressions (REs) people use to refer to concepts at different levels of abstraction. On the one hand, Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) famously found that basic categories are special because these are "the most inclusive categories for which a concrete image of the category as a whole can be formed" (Rosch et al., 1976). It has been shown that children acquire basic-level terms like *dog* first (Clark & Johnson, 1994; Mervis & Crisafi, 1982) and that objects can be categorized faster at the basic level than at the sub- or superordinate levels (Murphy & Smith, 1982). On the other hand, the Gricean maxim of quantity predicts that speakers provide as much information as required for the listener to identify a target in a given context and not more (Grice, 1975). This means that speakers should tailor their utterances to the communicative situation at hand, considering both the concept they would like to communicate and the context of their utterance. It has also been shown empirically that context plays a role in the selection of REs in referential situations (see for example Hawkins, Franke, Smith, & Goodman, 2018; Hawkins, Frank, & Goodman, 2020; Konopka & Brown-Schmidt, 2014; Sedivy, 2005). Our goal is to pit the two factors directly against each other. We use a reference game similar to Graf, Degen, Hawkins, and Goodman (2016), where a speaker describes an object and a listener needs to identify this object from a set of distractors. Graf et al. (2016) showed that while speakers tailor their utterances to the context, they also prefer basic-level expressions (e.g, *dog*) overall.

Going beyond this study, we test the tradeoff between the basic-level advantage and informativity considerations in production and comprehension. We ask whether there are differences between the production and comprehension of basic-level terms. The tasks that have mainly been used to study the basic-level effect have been instance-naming, i.e. production tasks (Clark & Johnson, 1994; Murphy & Smith, 1982; Rosch et al., 1976). More recently, the basic-level advantage has been challenged by studies on visual categorization, which involve comprehension (Macé, Joubert, Nespoulous, & Fabre-Thorpe, 2009; Poncet & Fabre-Thorpe, 2014). In these studies, participants reacted faster in response to a superordinate rather than a basic-level referent. One way to make sense of these conflicting findings is to suggest that the basic-level effect only holds in language production but not in comprehension.

While the basic-level advantage has been studied extensively, little is known about the process of abstraction from a basic-level category to a superordinate category, which is crucial to understanding abstraction itself. One hypothesis brought forward by Murphy and Wisniewski (1989) and Wisniewski and Murphy (1989) is that superordinate terms are used to refer to groups or classes of objects rather than to specific examples. Indeed, Wisniewski and Murphy (1989) found that superordinate terms are used more frequently in corpora to refer to groups and classes of objects, whereas basic-level terms are more frequently used to refer to single objects. Thus, superordinate terms might be produced more frequently in reference to concepts that include multiple objects. This hypothesis has not yet been tested experimentally.

Our goal is to study the process of abstraction jointly in comprehension and production. Specifically, we investigate

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

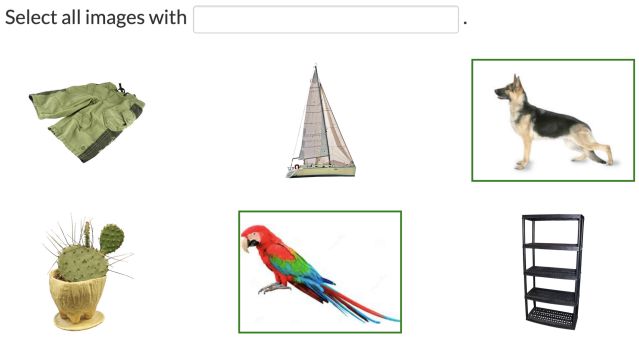

Figure 1: The speaker display in a concept-level reference game: Two images form the superordinate level target concept ANIMAL.

which referential expressions speakers produce in different contexts and how listeners categorize objects in an interactive setting. This will allow us to understand both pragmatic factors in referential communication as well as cognitive mechanisms involved in the production and comprehension of conceptual abstraction.

Current study

Reference games are the classic paradigm for studying referential expressions in dyadic communication. The goal of the game is the successful communication about a target within a set of distractors. Typically, one participant is assigned the speaker role and produces utterances that help the participant with the listener role to identify the intended target. Here, we introduce a new concept-level reference game to study how interlocutors communicate about concepts that comprise more than one object. In our task, the speaker needs to find a label that describes two targets (see Figure 1) and the listener needs to identify the targets based on the label. Increasing the number of targets in a reference game ensures that speakers and listeners communicate about concepts rather than about single objects (see Mu & Goodman, 2021, for a similar game with artificial agents). One of the main goals of our study is to elicit the production of the superordinate level so that we can study the process of abstraction, i.e. the differences between the basic and the superordinate level. The interactive setting allows us to look at production and comprehension at the same time, and to answer questions such as whether speakers tailor their utterances to the context and to the listeners' needs. The alternative is that speakers prefer the basic level regardless of context or listener needs because the basic level is more accessible and hence less costly to produce.

Our main manipulation is the conceptual context, which is defined by the combination of targets and distractors. We compare three critical conditions, as shown in Figure 2: In a fine conceptual context (2A), the targets belong to the same subordinate category and the closest distractor belongs to the same basic-level category (dalmatian vs. other kinds of dogs).

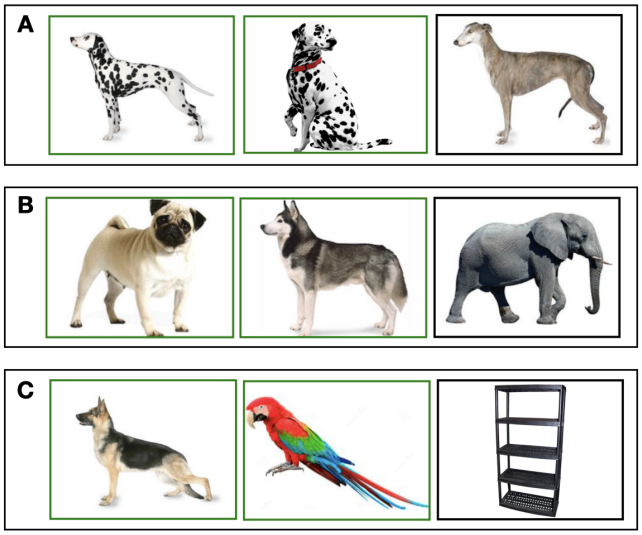

Figure 2: Examples for a fine (A), basic (B) and coarse (C) conceptual context: two targets and their closest distractor.

In the basic conceptual context (2B), the targets belong to the same basic-level category and the closest distractor belongs to the same superordinate category (dog vs. other animals). In the coarse conceptual context (2C), the targets belong to the same superordinate category and the closest distractor is unrelated (animal vs. non-animals).

We develop three hypotheses based on the existing literature. First, we expect speakers to tailor their utterances to the conceptual context, i.e. to the combination of target and distractor objects, as predicted by the Gricean maxim of quantity (Grice, 1975). That means they will choose the optimal expression that allows the listener to identify the targets in a given context. Accordingly, we expect speakers to produce utterances on a subordinate level in fine conceptual contexts (e.g. *dalmatian*), to produce utterances on a basic level only in basic conceptual contexts (e.g. *dog*), and to produce utterances on a superordinate level in coarse conceptual contexts (e.g. *animal*). Second, we expect to reproduce the basic-level advantage (Rosch et al., 1976) in the production data. This means that we expect speakers to respond more quickly when they produce a basic-level term than when they produce a sub- or superordinate term. Third, we will look at how quickly listeners comprehend the REs produced by the speakers. Here, we have two mutually exclusive hypotheses: Either, we will also see a basic-level advantage in the listener data, or, we expect superordinate terms to be processed faster than basic-level terms as suggested by the visual categorization studies (Macé et al., 2009; Poncet & Fabre-Thorpe, 2014).

Methods

The study design, hypotheses and analyses have been preregistered and the preregistration, data and analysis scripts are publicly available here: `https://osf.io/y7eqw/`.

Participants

We recruited 120 participants via Prolific. We only included data from pairs that completed the full experiment or at least 80% of trials in the analysis. This led to a final sample of 116 participants, i.e. 58 pairs. We had 61 female, 48 male and 7 participants of diverse gender. Their age ranged from 18 to 30 with a median age of 26. All were English native speakers currently living in the United States. All had normal or corrected-to-normal vision and no problems with perceiving color. Participants were also screened to have a high approval rate of 90-100% on Prolific to ensure high-quality data. Participants were paid around £2.35 for the approximately 10 minutes that the experiment lasted (£14.15/hour).

Design and procedure

We implemented a concept-level reference game in the tradition of the classical reference game paradigm, in which a speaker describes a target to a listener and the listener has to pick out the correct target object from a set of distractors.[1] To bring the game to a concept-level, we introduced one main alteration to the classic setting: Instead of one target, we used two targets that together form a target concept. Target concepts were sampled from three levels of reference: the subordinate, basic or superordinate level of a taxonomy. For example, if the subordinate concept is DALMATIAN, two pictures of dalmatians form the target concept.

The main manipulation in our experiment is the conceptual context which is manipulated within-subjects. It is defined by the combination of two target and four distractor objects and has three levels: "fine", "basic" and "coarse". In the fine condition, the two target objects share the same subordinate category, e.g. DALMATIAN (two dalmatians), one distractor shares the same basic category (i.e. another dog), one distractor shares the same superordinate category (i.e. another animal) and the two other distractors are unrelated. In the basic condition, the targets share the same basic-level category, e.g. DOG (two dogs; a pug and a dalmatian), one distractor shares the same superordinate category and three distractors are unrelated. And in the coarse condition, the targets share the same superordinate category, e.g. ANIMAL (two animals; a pug and a parrot) and the four distractors are unrelated. We defined distractor objects that are closely related to the targets to make the production of a RE on an appropriate level, i.e. on a level that matches the concept level, necessary for disambiguation.

Participants played 18 rounds of the concept-level reference game together in dyads, where one participant was assigned the speaker role and the other participant was assigned the listener role. Both roles stayed constant throughout the experiment. One round of the game followed this procedure: The speaker sees a display of target and distractor objects, in which the target objects are marked by a green frame. The speaker sends a message to the listener in the form of a cloze task: They are asked to fill in the gap in the following sentence with a noun or a compound: "Select all images with [gap]" (see Figure 1). The listener receives the message and selects two images from the visual display. We collected the utterances produced by the speaker, the objects selected by the listener, as well as response times of both speaker and listener. Speaker response times were logged once a speaker had typed in a label and pressed Enter on the keyboard. Listener response times were logged once a listener had clicked on two images. We randomized the trial order and the position of the images in the display. Participants were paired on a random basis and were randomly assigned the speaker or the listener role.

Stimuli and typicality norming

The image stimuli were reused from a related study on nominal REs with a reference game (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Graf et al., 2016). We used the same superordinate domains as in the original study: ANIMAL, CLOTHING, FOOD, FURNITURE, PLANT and VEHICLE. In the superordinate domain FURNITURE, we replaced the basic category TABLE with CHAIR. In the superordinate domain ANIMAL we only kept the DOG basic category and excluded BIRD, BEAR and FISH such that we had the same number of basic-level categories for each superordinate domain. Because we increased the number of targets and distractors, we needed more stimuli than used in the original study for each concept. We chose these stimuli from the BOSS image database (Brodeur, Guérard, & Bouras, 2014) with coherence to the superordinate categories. For these new stimuli, we collected typicality ratings from 75 participants in a separate experiment with the same typicality rating task as in Graf et al.'s original study.[2] All unrelated distractors were rated for their (un)typicality for the superordinate category for which they function as unrelated distractors in the experiment. For example, the unrelated distractor BOOKSHELF in Figure 2C would be rated for its typicality in the superordinate category ANIMAL. On a scale from 0 (very atypical) to 1 (very typical), all unrelated distractors received a mean score below 0.1. Targets, on the other hand, received a mean score of above 0.5 for the superordinate category they belong to.

Results

The analyses were run in R version 4.2.2 (R Core Team, 2022) with the R package brms version 2.20.4 (Bürkner, 2017) and bayestestR version 0.13.0 (Makowski, Ben-Shachar, & Lüdecke, 2019). The models have run for 4,000 iterations with a warm-up period of 1,000 iterations and with treatment-coded predictors if not specified otherwise. All reported models have converged with an R-hat value of 1.0 and effective sample sizes of over 1,000. We use bootstrapped 95% Confidence Intervals (CIs) for plotting error bars in the data. We report posterior estimate means and 95% Credible

[1] The experiment was programmed in Labvanced (Finger, Goeke, Diekamp, Standvoß, & König, 2017).

[2] This experiment was programmed in pcIBEX (Zehr & Schwarz, 2018).

Intervals (CrIs) to show the size and direction of an effect. Whether the CrI includes zero is used to indicate whether the predictor is needed to explain the data. Bayes Factors (BFs) are used as a hypothesis test. As an additional basis for understanding the probabilities of the investigated effects, we also report the results of Bayesian tests for the existence and significance of effects, probability of direction (pd), and ROPE (Region Of Practical Equivalence) (Kruschke, 2018; Makowski, Ben-Shachar, Chen, & Lüdecke, 2019).[3]

Data cleaning

The produced utterances have been cleaned and sorted into the levels "sub", "basic", and "super" depending on the taxonomic level of reference they contain. Following the procedure described in Degen et al. (2020); Graf et al. (2016), typographical errors and meaning-equivalent alternatives have been included as well. Utterances that did not contain a nominal RE and could not be assigned to one of the three categories are coded as NA in the data. For example, if participants communicate an attribute of a category rather than the category itself, these trials are coded as NA.

Communicative success

First, we look at whether the production of a RE at the appropriate level of reference increases communicative success, i.e. that the listener selects the correct target objects. Appropriate levels of reference are defined by the conceptual context as described above. All other mentions have either been references on inappropriate levels, e.g. *spotted dog* for DALMATIAN or *dog and bird* for ANIMAL or attributes or associations with a category that could not be assigned to one of the three levels. We ran a Bayesian model with uninformative priors that predicts communicative success by the appropriateness of the reference level (coded as true/false) with a Bernoulli link function.[4] The model predicts that while the success rate is quite high (83.62%) even if speakers choose a term on an inappropriate level of reference, it is substantially higher (98.37%) if speakers choose a term on the appropriate level of reference (M=2.47, CrI=[1.92, 3.03], pd=100%, ROPE=[-0.18, 0.18], 0% in ROPE).

Choice of reference level

Our first hypothesis was that the conceptual context determines the level of the referential expression speakers choose. We predicted that speakers produce subordinate terms in the fine conceptual context, basic-level terms in the basic conceptual context and superordinate terms in the coarse conceptual context. This predicted pattern is visible in the data in Figure 3. As preregistered, we excluded trials in which the communication was unsuccessful, i.e. the listener did not select both correct target objects. This led to an exclusion of 4.34% of the data. We also excluded trials in which the produced utterance could not be sorted into the three reference

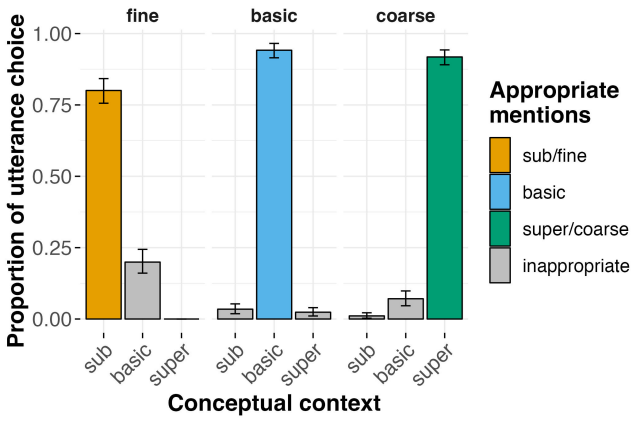

Figure 3: Proportion of choosing a RE on a particular level of reference in each conceptual context. Error bars represent bootstrapped 95% CIs. Appropriate mentions are color-coded.

levels. These were mostly utterances that referred to properties of an object rather than to the object itself, or utterances that listed both objects in the display with a conjunction, e.g. *chair and shelf*, rather than referring to the concept. This led to the exclusion of another 5.59% of the data with a total exclusion of 9.93%. We ran a Hierarchical Bayesian model with a Bernoulli link function that predicted reference level (coded as true/false) by conceptual context with group-level effects for participant group and item category. We used weakly informative but conservative priors[5]. The model confirms our predictions: Mentions on the appropriate level are substantially more frequent than mentions on an inappropriate level. This is supported by the estimates for the basic conceptual context (M=3.62, CrI=[2.45, 4.83]), coarse conceptual context (M=4.53, CrI=[2.96, 6.25]) and fine conceptual context (M=2.23, CrI=[0.97, 3.56]) which translate into the probabilities 97.39%, 98.93% and 90.29%, respectively. We find substantial evidence for the differences between both the basic and fine conceptual context (M=1.39, CrI=[0.34, 2.34], pd=99.08%, ROPE=[-0.18, 0.18, 0% in ROPE]) and between the coarse and fine conceptual context (M=2.3, CrI=[0.33, 4.35], pd=99.11%, ROPE=[-0.18, 0.18], 0% in ROPE).[6] The effect of conceptual context on the reference level is further supported by a Bayes Factor of 13.5 in favor of the model that includes conceptual context as a predictor against a null model, providing strong evidence for the effect of conceptual

[3]The ROPE range was calculated with the rope_range function from bayestestR (Makowski, Ben-Shachar, & Lüdecke, 2019).

[4]This model was not preregistered.

[5]The priors were specified as follows: intercept prior: student-t(3, 0, 2.5), population-effects slope prior: student-t(3, 0, 1), group-level effects standard deviation prior: student-t(3, 0, 1), group-level effects correlation prior: lkj(2).

[6]At the suggestion of our reviewers, we also ran a model on the data without excluding unsuccessful trials. This model provides similar evidence to the original model: basic-fine: M=1.63, CrI=[0.35, 2.81], pd=98.65%, ROPE=[-0.18, 0.18], 0% in ROPE; coarse-fine: M=2.58, CrI=[0.44, 4.79], pd=99.16%, ROPE=[-0.18, 0.18], 0% in ROPE.

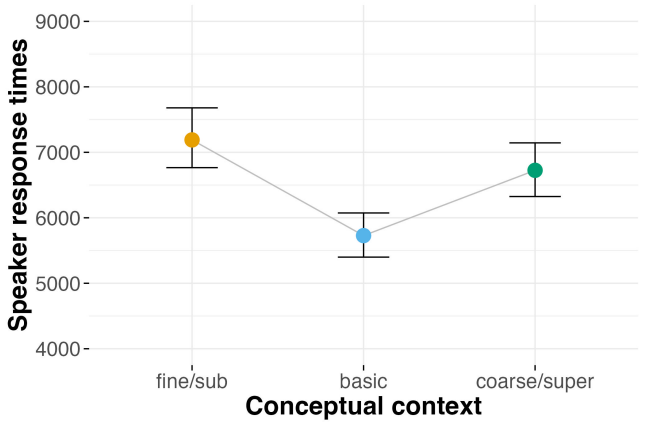

Figure 4: Speaker response times with bootstrapped 95% CIs.

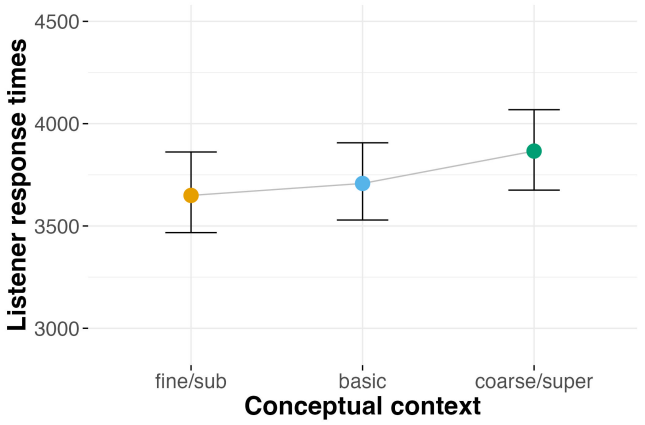

Figure 5: Listener response times with bootstrapped 95% CIs.

context on the choice of the reference level.[7]

Response times

For the response time analyses, we were interested in response times of trials in which the listener selected the correct target objects, i.e. communication was successful, and in which the speaker chose an utterance on the appropriate level of abstraction. The exclusion of data on inappropriate levels led to a data loss of 11.12%.

Speaker response times Our second hypothesis was that speakers choose utterances on the basic level more quickly than utterances on sub- or superordinate levels. This hypothesis was motivated by the basic-level effect found in the literature (Rosch et al., 1976; Tanaka & Taylor, 1991). We thus predicted that speaker response times would be shorter when the produced utterance was on the basic level compared to the other two levels. As preregistered, trials with response times over 2.5 standard deviations above the mean (cut-off: 22,360 ms) were excluded, leading to the exclusion of 2.61% of the remaining data. Figure 4 shows the data means and bootstrapped confidence intervals of the cleaned data. Indeed, responses on the basic level are shorter than response times on the other two levels. We ran a Bayesian model with a lognormal link function, predicting speaker response times by conceptual context and including group-level effects for the participant pair and item category. We specified weakly informative priors to enhance model fit. We had to deviate from one prior specified in the preregistration: Prior predictive checks showed that we overestimated the effect size, and the prior we preregistered for the population-level effect was too wide. We thus changed the prior's standard deviation from 2 to 0.5 to enhance model convergence and fit.[8]

The model results confirm that speakers respond substantially faster on the basic level than on the other two levels of reference (M=-0.22, CrI=[-0.36, -0.10], pd=99.86%, ROPE=[-0.01, 0.01], 0% in ROPE).[9][10]

Listener response times Our third hypothesis was that we would also find differences between levels in the listener response times, showing that either a) listeners select the correct targets more quickly when a basic-level term was produced or b) listeners select the correct targets more quickly when a superordinate term was produced. These hypotheses were based on the basic-level effects literature (Rosch et al., 1976) and more recent studies debating the basic-level effects in certain tasks (Macé et al., 2009). The data in Figure 5 shows that we do not find either of the expected patterns. We ran a Bayesian model with the same model specifications and priors as the speaker response time model.[11] We employed the same data exclusion criteria as for the speaker response time model with one exception: We decided to exclude the highest response time data point and defined the cut-off of 2.5 standard deviations above the mean after this exclusion to get a more sensible cut-off (17,076 ms). The model predictions do not offer enough reason to believe that either of the two proposed patterns is at play: The difference between the basic level and the other two levels was estimated at M=-0.04 with a CrI of [-0.12, 0.04] (pd=84.96%, ROPE=[-0.01, 0.01], 12.28% in ROPE). And the difference between the coarse level and the other two levels was estimated at M=0.04 with a CrI of [-0.07,

[7]The BF models ran for 20,000 iterations with a warm-up period of 2,000 as recommended for BF estimation (Nicenboim, Schad, & Vasishth, 2023).

[8]The priors were specified as follows: intercept prior: normal(8.65, 0.5), population-level effects slop prior: normal(0, 0.5), group-level effects standard deviation prior: normal(0, 0.1), group-level effects correlation prior: lkj(2), sigma prior: normal(0, 0.5).

[9]This posterior difference was contrast-coded as preregistered: basic vs. (fine + coarse)/2.

[10]At the suggestion of our reviewers, we also ran a model including length as a predictor. This model provides smaller evidence for the difference of the basic level compared to the other two: M=-0.10, CrI=[-0.19, 0.00], pd=98.07%, ROPE=[-0.01, 0.01], 0.62% in ROPE. This is mostly driven by the difference between basic and coarse being not as pronounced as the difference between basic and fine.

[11]The only exception was the prior on the intercept that was preregistered to depend on the data distribution: normal(8.15, 0.38).

0.14] (pd=79.07%, ROPE=[-0.01, 0.01], 11.81% in ROPE).[12]

Discussion and Conclusion

We studied the expressions speakers use to refer to concepts at different levels of abstraction in a new interactive concept-level reference game. We have shown that the level of abstraction of the utterances speakers choose to communicate a certain concept mainly depends on the concept and context in question. Speakers tailor their utterances to the context, producing subordinate terms in fine conceptual contexts, basic-level terms in basic conceptual contexts and superordinate terms in coarse conceptual contexts. In fact, we found that if speakers produce a RE on the expected level of reference, this reduces errors in the target selection by the listeners, or, in other words, it increases communicative success by about 14%. We also reproduced the basic-level effect on the production side, i.e. speakers are faster in producing a basic-level term than in producing a term on the other two levels. Interestingly, even though speakers show a basic-level advantage, they still tailor the utterances to fit the context and make it easier for the listener to identify the correct targets. We did not find evidence for an advantage of basic-level processing on the comprehension side, i.e. listeners are equally quick in selecting the targets regardless of the level of abstraction of the utterance they receive. A possible reason for this is that the basic-level advantage is mostly driven by accessibility in production. In comprehension, on the other hand, listeners might be as quick in categorizing objects at other levels than at the basic level.

Despite the basic-level effect, speakers use sub- and superordinate terms frequently in natural conversation. In the case of subordinate terms, this is usually determined by the context. When, for example, a dalmatian is the target and a greyhound is the distractor, the basic-level term *dog* does not sufficiently discriminate the target from the distractor. It has been shown that context warrants the use of a more costly, i.e. usually longer and less frequent, subordinate term (Graf et al., 2016). In the case of superordinate terms, however, the context does not sufficiently explain why these terms might be used because participants could always use the basic-level term to refer to single objects even in a coarse conceptual context. Our concept-level reference game shows that superordinate terms become relevant when dealing with multiple target objects, or when communicating the idea of a more generic class. The longer response times we see when speakers use a superordinate compared to a basic-level term might be an indicator of a process of abstraction that speakers undergo when trying to find which superordinate class the two targets have in common and retrieving the respective superordinate RE. This could be a good starting point for further research on abstraction.

One limitation of our current study setup is that so far, we only investigate REs in a context that is very close to the target concept, i.e. that includes a distractor from the same basic category in the fine conceptual context or from the same superordinate category in the basic conceptual context. This means that we cannot account for overinformative REs (see for example Degen et al., 2020) because the context makes a certain level of reference necessary for disambiguation. Future studies can extend our setup and include context conditions that make a certain level of reference only sufficient for discrimination, by using wider contexts that, for example, only include unrelated distractors. Such a manipulation would allow the investigation of over- and underspecification in the concept-level reference game. However, even in our current set-up, we do see some utterance choices that are at odds with our predicted level of reference for each conceptual context. For example, in the fine conceptual context, basic-level expressions are produced almost 20% of the time. A closer look at these productions reveals that speakers either underspecify, i.e. produce *dog* for DALMATIAN, or they use a modified basic-level expression, i.e. produce *spotted dog*. The high proportion of these mentions provides further evidence for a strong basic-level effect on the speaker side.

The response time results in which we find a basic-level effect only for production, but not for comprehension, lead to an interesting observation: Speakers tailor their utterances to the conceptual context even when it results in higher processing costs for them. On the comprehension side, however, we do not find higher processing costs for sub- or superordinate terms. This could suggest that speakers are willing to bear a higher cost because they know that it would make identification of the target objects easier for the listener. This phenomenon has been discussed in the literature as audience design (see for example Gann & Barr, 2014; Horton & Gerrig, 2002). We should note, however, that the comprehension response times were logged when listeners had clicked on both targets. Thus, our measure is rather offline, and we cannot completely rule out that there are more immediate differences across levels in comprehension that might be revealed by more sensitive measures. If however the differences we observed for production and comprehension are not just due to such methodological aspects, this could indicate that the basic-level advantage is related to lexical accessibility and not categorization itself. On the listener's side, basic level categories may not have a privileged representation.

In conclusion, the concept-level reference game allows us to test hypotheses on the use of superordinate REs and abstraction. While we see differences in response times between the basic and superordinate levels on the production side, we do not see the same differences on the comprehension side. This opens up exciting possibilities for future research on audience design and costs associated with abstraction in production and comprehension. Here, we showed that superordinate REs become relevant when a speaker needs to describe more than one target object.

[12]These posterior differences were contrast-coded as preregistered: basic vs. (fine + coarse)/2 and coarse vs. (basic + fine)/2.

Acknowledgments

We would like to thank Ilva Hovemann for help with programming the experiment in Labvanced, Berit Reise for help with programming the norming study in pcIBEX, and Elli Tourtouri for helpful discussions on the experiment design. We also thank three anonymous reviewers for their helpful comments and feedback.

Kristina Kobrock is supported by the DFG-funded Research Training Group "Computational Cognition" (DFG-GRK 2340).

Author Contributions:

Kristina Kobrock: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization. **Charlotte Uhlemann:** Conceptualization, Methodology, Software, Writing - Review & Editing. **Nicole Gotzner:** Conceptualization, Methodology, Writing - Review & Editing, Supervision.

References

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) Phase II: 930 New Normative Photos. *PLOS ONE*, *9*(9), e106953. doi: 10.1371/journal.pone.0106953

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Clark, J. M., & Johnson, C. J. (1994). Retrieval Mechanisms in the Development of Instance and Superordinate Naming of Pictures. *Journal of Experimental Child Psychology*, *57*(3), 295–326. doi: 10.1006/jecp.1994.1015

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591–621. doi: 10.1037/rev0000186

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). *LabVanced: a unified JavaScript framework for online studies.* Cologne, Germany. Retrieved from https://www.labvanced.com/publication.html

Gann, T. M., & Barr, D. J. (2014, July). Speaking from experience: audience design as expert performance. *Language, Cognition and Neuroscience*, *29*(6), 744–760. doi: 10.1080/01690965.2011.641388

Graf, C., Degen, J., Hawkins, R. D., & Goodman, N. D. (2016). Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2261–2266.

Grice, P. H. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vols. 3, speech acts, pp. 41 – 58). New York: NY: Academic Press.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020, June). Characterizing the Dynamics of Learning in Repeated Reference Games. *Cognitive Science*, *44*(6). doi: 10.1111/cogs.12845

Hawkins, R. D., Franke, M., Smith, K., & Goodman, N. D. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 463–468).

Horton, W. S., & Gerrig, R. J. (2002, November). Speakers' experiences and audience design: knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589–606. doi: 10.1016/S0749-596X(02)00019-0

Konopka, A. E., & Brown-Schmidt, S. (2014). Message encoding. In M. Goldrick, V. Ferreira, & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 3–20). Oxford University Press.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi: 10.1177/2515245918771304

Macé, M. J.-M., Joubert, O. R., Nespoulous, J.-L., & Fabre-Thorpe, M. (2009). The Time-Course of Visual Categorizations: You Spot the Animal Faster than the Bird. *PLoS ONE*, *4*(6), e5927. doi: 10.1371/journal.pone.0005927

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.02767

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. doi: 10.21105/joss.01541

Mervis, C. B., & Crisafi, M. A. (1982). Order of Acquisition of Subordinate-, Basic-, and Superordinate-Level Categories. *Child Development*, *53*(1), 258. doi: 10.2307/1129660

Mu, J., & Goodman, N. (2021). Emergent Communication of Generalizations. *Advances in Neural Information Processing Systems*, *34*, 17994–18007.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, *21*(1), 1–20. doi: 10.1016/S0022-5371(82)90412-1

Murphy, G. L., & Wisniewski, E. J. (1989). Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 572–586. doi: 10.1037/0278-7393.15.4.572

Nicenboim, B., Schad, D., & Vasishth, S. (2023). *An Introduction to Bayesian Data Analysis for Cognitive Science*. Retrieved 2023-05-31, from https://vasishth.github.io/bayescogsci/book/

Poncet, M., & Fabre-Thorpe, M. (2014). Stimulus duration and diversity do not reverse the advantage for superordinate-level representations: the animal is seen before the bird. *European Journal of Neuroscience*, *39*(9),

1508–1516. doi: 10.1111/ejn.12513

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rosch, E. (1978). Principles of Categorization. In *Cognition and Categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. doi: 10.1016/0010-0285(76)90013-X

Sedivy, J. C. (2005). Evaluating Explanations for Referential Context Effects: Evidence for Gricean Mechanisms in Online Language Interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions* (pp. 345–364). Cambridge, Massachusetts; London, England: The MIT Press.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. doi: 10.1016/0010-0285(91)90016-H

Wisniewski, E. J., & Murphy, G. L. (1989). Superordinate and basic category names in discourse: A textual analysis. *Discourse Processes*, *12*(2), 245–261. doi: 10.1080/01638538909544728

Yee, E. (2019). Abstraction and concepts: when, how, where, what and why? *Language, Cognition and Neuroscience*, *34*(10), 1257–1265. doi: 10.1080/23273798.2019.1660797

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX).* doi: https://doi.org/10.17605/OSF.IO/MD832