Experimenting with subject alternatives for analysing generic sentences[1]

Arnold KOCHARI — *Institute of Logic, Language and Computation, University of Amsterdam*
Robert VAN ROOIJ — *Institute of Logic, Language and Computation University of Amsterdam*
Katrin SCHULZ — *Institute of Logic, Language and Computation, University of Amsterdam*

Abstract. In this paper we argue that for the (probabilistic) interpretation of generic sentences of the form '*G*s are *f*' alternative groups, or kinds, of *G* play an important role. We describe the results of some experiments that empirically test this use of alternatives.

Keywords: Generic sentences, Experimental Semantics, Subject-term alternatives, Probability.

1. Introduction

Generic sentences are sentences of the form '*G*s are *f*' that, by their very nature, express useful generalisations. Thereby, the question of their truth, or acceptability, can be translated into the question of when we think that the expressed generalisation holds. A very natural and often explored approach to this question is the **majority rule** for the interpretation of generics (cf. Cohen, 1999). According to the majority rule a generic is true or acceptable in case the probability of a member of group G having feature f is high, (much) higher than $\frac{1}{2}$ ($P(f|G) > \frac{1}{2}$).[2]

Definition 1 *A simple majority rule for generics.*
A generic sentence 'Gs are f' is true in case $P(f|G) > \frac{1}{2}$.

Thus, taking a generic like (1), we say that this sentence is true in case the majority of the birds fly.

(1) Birds fly.

This natural approach to the meaning of BP generics nicely accounts for the fact that not all birds need to fly in order for the generic to be true and still plays an important role in the literature on generic expressions.

Much ink has been spilled on the following 'Port-Royal' type of generics:

(2) a. Dutchmen are good sailors;
b. Bulgarians are good weightlifters.

Intuitively, the above sentences are appropriate, although only a small percentage of Dutchmen are good sailors and only few of all Bulgarians are good weightlifters.

Cohen (1999) proposed that generics like (2a)-(2b) are true, because they should be interpreted differently than standard generics, namely in a *relative* way: (2a) is true iff compared to relevant alternative prople in the 17^{th} century (Frenchmen, Spaniards, Englishmen, and people from the Germanic countries), *relatively many* Dutchmen are good sailors. Similarly for (2b). In probabilistic terms this means that $P(f|G) > P(f)$.

[1]We would like to thank the audience of SUB 24 for their useful questions after our presentation of this material.
[2]To be sure, Cohen (1999) also makes use of alternatives for feature f, $Alt(f)$, to determine the 'domain' of the probability function: $P(f|G \cap \bigcup Alt(f)) > \frac{1}{2}$. We ignore those alternatives in this paper, however.

Proceedings of Sinn und Bedeutung 24, vol. 1, pp.411–428. Osnabrück University.

There exists an alternative formulation of the same idea. First, it is a fact of probability theory that $P(f|G) > P(f)$ iff $P(f|G) > P(f|\neg G)$. Second, it is natural to think of $\neg G$ as being an abbreviation of the set of individuals that are members of a group that can be thought of as an alternative to group G. If we refer to $Alt(G)$ as the set of groups alternative to G (all incompatible with G), we can think of $\neg G$ as an abbreviation of $\bigcup Alt(G)$. Thus, we end up with the following definitions of the truth-conditions of generic sentences.

Definition 2 *Truth conditions for generics with G-alternatives.*
A generic sentence 'Gs are f' is ambiguous between an **absolute** *and a* **relative reading**. *In its absolute reading the truth-conditions of Definition 1 apply. In its relative reading the generic is true in context c in case for a contextually salient set $Alt(G)$ of alternatives to G it holds that:*

$$P(f|G) > P(f|\bigcup Alt(G)).$$

There is an important justification for assuming that generic sentences (also) have a relevant reading, and thus that the alternatives $Alt(G)$ matter for the interpretation of a generic sentence. Above, we have stated that generic sentences express, by their very nature, useful generalisations. This suggests that there is a close relation between our acceptance of generic sentences, on the one hand, and the way we *learn* generalisations, on the other (cf. Leslie, 2008). Much psychological research on learning was done before the cognitive revolution in psychology, in classical conditioning.

For animal learning, Rescorla (1968) observed that rats learn a tone (cue/cause)-shock (outcome) association if the frequency of shocks immediately after the tone is higher than the frequency of shocks undergone otherwise. This holds, even if in the minority of cases a shock actually follows the tone. Gluck & Bower (1988) and others show that humans learn associations between the representations of certain cues (properties or features) and outcome (typically another property or a category prediction) in a very similar way. Thus, we associate outcome o with cue c, not so much if $P(o|c)$ is high, but rather if $\Delta P_c^o = P(o|c) - P(o|\neg c)$ is high, where ΔP_c^o is known as the *contingency* of o on c. As noted above, $\Delta P_c^o = P(o|c) - P(o|\neg c) > 0$ if and only if $P(o|c) > P(o)$, i.e., the condition Cohen (1999) demands to be satisfied for relative readings of generics. In Tessler & Goodman (2019) and in van Rooij & Schulz (2019) it is hypothesised that (a strengthened version of) Cohen's relative reading is the basic reading of generics. So, in contrast to Cohen (1999), we don't think that the absolute reading is the default reading, but only a special case of the (strengthened version of the) relative reading. It is this that we want to test.

2. Empirical results on the role of *G*-alternatives

In the previous section we have argued in favour of the claim that subject-alternatives are relevant for the interpretation of a generic sentence of the form '*G*s are *f*'. Moreover, we have argued that alternatives to the subject term *G* are important in any case to learn the (inductive) generalisation. We provided independent evidence coming from the field of psychology of learning. In this section we will present the results of three empirical studies on the relevance of *G*- alternatives for the interpretation of generics.

2.1. The hypotheses that we will test

The central goal of this part of our research was to empirically test whether alternatives to the subject term G do indeed affect the acceptability of a generic sentence. Specifically, we hypothesize that the probability with which the alternatives carry the relevant feature f affects the acceptability of the generic.

Hypothesis 1 *The acceptability of a generic sentence 'Gs are f' depends on the conditional probability of the feature f given salient alternatives G′ of G.*

To test this hypothesis, we manipulate $P(f|G')$ and see whether we can observe an effect on the acceptability of the generic. Depending on whether or not this hypothesis is supported by the data, we can then test different approaches to the meaning of generic sentences that explain the result. For instance, if the observed acceptability is in line with Hypothesis 1, then we can test whether contingency is a good predictor for the acceptability of generic sentences.

Hypothesis 2 *The acceptability of a generic sentence 'Gs are f' is given by the formula*

$$acceptability\ of\ 'Gs\ are\ f' = P(f|G) - P(f|\bigcup Alt(G)).$$

In the following, we will present the results of two experiments testing the hypotheses formulated above. We were looking for a setup that allowed us to probe the intuitions of people concerning generics about a group of objects for which they do not have any prior knowledge. This will allow us to ensure that participants do not have prior beliefs about features typical for the objects they will see. A second objective was to control the G alternatives that the interpreters were considering. This is the factor that we will manipulate in order to see whether it influences the acceptability of the generic sentence.

We presented participants with a picture-sentence verification task similar to that used in Bordalo et al. (2016). The participants saw pictures with samples of fictive insect species from two Galapagos islands, Genovesa and Marchena (see Figure 1).[3] Their task was to assess whether animals from one of the islands, Genovesa, could be described with a given sentence. All sentences were generics stating that the species from Genovesa – our target group G – has a particular feature having to do with their colouring – our target feature f. We controlled the conditional probabilities $P(f|G)$ that the participants of the studies assigned by manipulating how many of the animals G in the sample form Genovesa showed the particular colouring pattern f. The second sample from Marchena served as contextually salient alternative. By manipulating the frequency of insects with the relevant feature in this group we controlled $P(f|\bigcup Alt(G))$, which we will denote from now on by $P(f|Alt(G))$.

We presented pictures in two conditions. In the non-contrastive condition an equal number of insects (80%) in both samples had the relevant feature f (see Figure 1). Thus, in this case $P(f|G) = P(f|Alt(G))$. In the contrastive condition, none of the insects in the sample from Marchena (the salient alternative) had the feature, while 80% of the insects from Geneva (the target G) had the feature f (see Figure 2). In other words, in this condition $P(f|G) = 0.8$ and

[3]The names of the islands are real. The participants were also shown a map of the Galapagos islands with the location of the islands. We chose animals instead of, for instance, manipulating the clothing of people, because the colouring of animals would not be perceived as an accidental feature of the observed individuals.

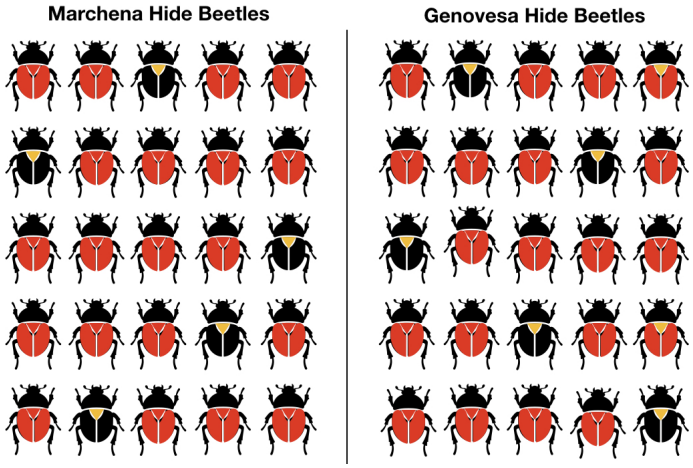

Figure 1: Sample picture in the non-contrastive condition with beetles.

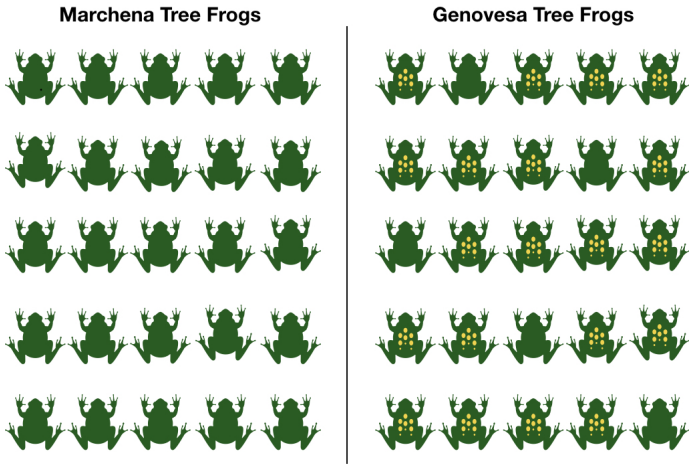

Figure 2: Sample picture in the contrastive condition with frogs.

$P(f|Alt(G)) = 0$. Based on Hypothesis 1, we expect that the strong difference of $P(f|Alt(G))$ between both conditions should have a significant effect on the acceptability of the generic sentences. Hypothesis 2 predicts that the judgments of acceptability people give for the generics should correspond to the contingency or the relative difference of feature f given group G.

2.2. Study 1

In the first study we used a within-subjects design. All participants gave an acceptability score to one sentence in the contrastive condition, one in the non-contrastive condition and one filler sentence. Each question was presented with a different animal species (spiders, frogs or bugs). Below the two samples, a generic sentence was given that always described the species from Genovesa. The participants were asked to judge on a scale from 0 to 5 whether the generic sentence was acceptable given the provided data (e.g., *"Can you say the following to describe Tree Frogs from Genovesa?"*, see also Figure 3). They gave a response by dragging a slider as depicted in Figure 3. They could adjust their response with a accuracy of two decimals, so they

Figure 3: Example question from the study.

experienced the scale as continuous.

Based on Hypothesis 1, we expected a significant difference in the judgments of acceptability for both conditions. Hypothesis 2 claims that the judgments of acceptability people give for the generics should correspond to the contingency of feature f given group G. In terms of proportions this measure predicts that the acceptability of a generic should increase if feature f becomes more distinctive for the group G. Applied to the two conditions distinguished here we would expect that the generic is significantly more acceptable in case of the contrastive condition than in the non-contrastive condition. The measure of contingency also makes precise numerical predictions for the acceptability of generics. However, these predictions need to be translated into the scale presented to the participants in the study, because the range of the contingency function does not match the scale presented to the participants of the study: the contingency function ranges between -1 and 1, whereas the scale the participants saw let them grade the acceptability of the sentences between 0 and 5. We used a linear transformation to map their responses directly onto the range $[-1,1]$ of the contingency function. Thus, 0 on the scale corresponds to a contingency of -1, 2.5 to a contingency of 0, and 5 to a contingency of 1. If we apply this linear transformation to the conditions that the participants of our study saw, Hypothesis 2 predicts that in the non-contrastive condition the contingency of the generic is 0, thus the participants should move the slide to around 2.5 on the given scale. In the contrastive case the contingency is $P(f|G) - P(f|Alt(P)) = 0.8 - 0 = 0.8$. This corresponds to the value 4.5 on the scale the participants saw. Given that there will be variation in how participants interpret the scale, we did not expect exactly the values predicted by the measure of contingency. However, the general proportional prediction described above should be visible in the data.

2.2.1. Method

Materials & procedure We used pictures of three different animal species (Tree Frogs, Hide Beetles, Jumping Spiders). For each species we designed a picture in the contrastive and in the non-contrastive condition. All the pictures contained two samples, one with 25 animals of the species from Marchena, one with 25 animals from the species from Genovesa. For each species we had one corresponding generic sentence: ”Hide Beetles from Genovesa have red wings”, ”Tree Frogs from Genovesa have yellow dots”, ”Jumping Spiders from Genovesa have green backs”.

The participants saw each animal species once, one in the contrastive condition, one in the non-contrastive condition and a third species as a filler. This resulted in 3 experimental trials per participant. In the filler condition, participants saw a generic that claimed the group to have a feature that none of the animals had (for instance, it could be the picture on Figure 1 with the generic "Hide Beetles from Genovesa have green wings") and, therefore, this sentence was clearly unacceptable. The filler condition was used to control whether participants completed the study in good faith: we excluded participants who gave a score above 1.5 in the filler condition as they likely did not pay attention in the other conditions either. The order in which the contrastive and the non-contrastive condition where shown was randomised. The filler always occurred last.

The study was implemented in Qualtrics. Participants started by reading the informed consent text and agreeing to taking part. They then read the instructions. Average time spent on the task was 143 seconds.

Participants Participants were recruited via Prolific.ac, an online platform aimed at connecting researchers and participants willing to fill in surveys and questionnaires in exchange for compensation for their time (Palan & Schitter 2018). We recruited native English speakers (British and American English) who reported no vision impairments.[4] Eighty-two participants completed the task. Three participants were excluded: two because they did not give a response in one of the experimental items, one because they gave a score of 1.5 or above on the filler item. Thus, 79 responses were included in the analyses reported below.

Due to a mistake in the set up of the experiment, the participants were not forced to answer the filler questions. We therefore ended up with 27 participants who gave no response to the filler conditions. However, the slider was always at 0 by default, so these participants most likely simply agreed with the score 0 and therefore pressed "respond" without moving the slider. For this reason, we still included these participants in the analyses.[5]

2.2.2. Results

The mean score given by the included participants in the filler condition was 0.04 (SD 0.16); the mean score in the contrastive condition was 3.51 (SD 1.06); and, finally, the mean score in the non-contrastive condition was 2.88 (SD 1.50). We performed a Bayesian paired samples t-test to test for the strength of evidence in favour of the null hypothesis (no difference between conditions) as opposed to the hypothesis that the score given by participants should be higher for contrastive than for non-contrastive condition using JASP software (JASP team 2018) with default priors. This analysis resulted in $BF_{10} = 104$, meaning that the data was 104 times more likely under our hypothesis than under the null hypothesis. Thus, the first study does lend support to Hypothesis 1 claiming that alternatives to G do affect the acceptability of a generic sentences and the general prediction of Hypothesis 2 about the tendency of this dependency: comparing situations in which a feature is distinctive vs. ones where it is not distinctive for a group, the generic has a higher acceptability in the situation in which the feature is distinctive.

[4]Since the material involved colours, the participants were required to have had normal vision of colours.
[5]Excluding these participants did not make a difference to the results reported here.

In order to evaluate compatibility of the data with the actual given scores based on the Hypothesis 2, we investigated the 95% confidence interval (CI) around the mean in each condition (assuming normal distribution). We expected a mean score 4.5 in the contrastive condition, but observed 3.51 with 95% CI [3.27 3.74] which does not include the expected score. For the non-contrastive condition, we expected a mean score 2.5, but observed 2.88 with 95% CI [2.54 3.21] which again does not include the expected score, but does come close. Overall, while the scores come close to the expected ones, we cannot conclusively say that the observed values support our second hypothesis (but see the issues raised below in the *Interim Discussion* regarding the potential caveats of our approach).

Figure 4 depicts the difference between given scores in the contrastive and non-contrastive conditions for each participant (specifically, displayed is score in contrastive condition minus score in non-contrastive condition). We can see that not all participants uniformly gave higher scores to the contrastive as compared to the non-contrastive condition. In fact, there was a sizable proportion of participants who gave approximately the same score in two conditions, and even a small group that gave the non-contrastive condition a higher score than the contrastive condition. Thus, we seem to be observing different behaviour patterns by different participants. We will come back to this in Section 2.4.

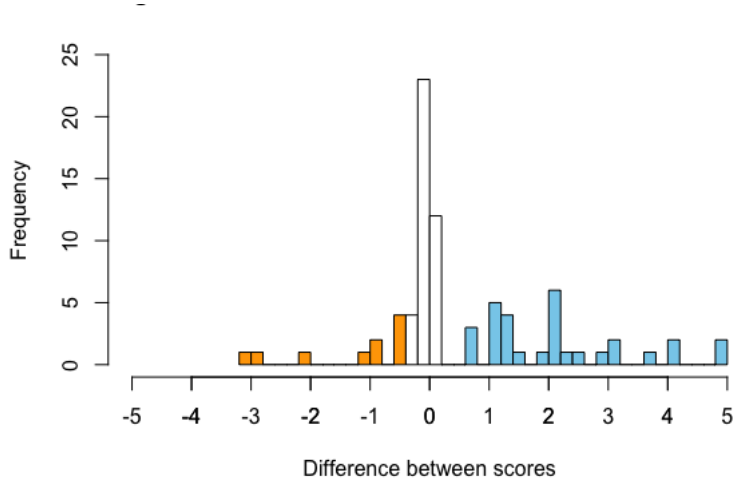

Figure 4: Histogram of differences in scores between conditions: contrastive condition minus non-contrastive condition. Differences below -0.5 are marked in orange color, differences above 0.5 are marked in blue color. Orange bars thus indicate participants who gave a higher score in the non-contrastive condition, non-coloured bars indicate participants who gave a similar score in both conditions, and blue bars indicate participants who gave a higher score in the contrastive condition.

2.3. Study 2

The results of Study 1 supported the hypothesis that the score given by participants to acceptability of a generic sentence will differ for the case with an alternative present and the case with no alternative present. The generic *'Gs are f'* becomes in general more acceptable in case the discussed feature f is distinctive for the group G. The results also partially confirm Hypothe-

sis 2: in the non-contrastive condition the generic was judged to be in between acceptable and unacceptable. In the contrastive condition the generic was on average rated to be acceptable, though not to the degree predicted by the contingency measure. In order to replicate the original finding, we administered the same task in a between-participant set-up: each participant saw only one of the two conditions (contrastive or non-contrastive) plus a filler item.

2.3.1. Method

Materials and procedure The materials used in this study were the same as in Study 1 except this time the participants saw only either contrastive or non-contrastive condition and a filler trial (2 trials in total). Average time spent on the task was 128 seconds.

Participants Participants were recruited via Prolific.ac with the same eligibility criteria. One hundred eighty-two participants completed the task. Three participants were excluded from the analysis because of a missing response to one of the items. Further 7 participants were excluded because of giving a score above 1.5 in the filler question. That left 172 participants for further analyses.

2.3.2. Results

The mean score given by the included participants in the filler condition was 0.07 (SD 0.23), in the contrastive condition 3.49 (SD 1.29; 95% CI [3.29, 3.68]), and in the non-contrastive condition 3.06 (SD 1.37; 95% CI [2.85-3.26]). We performed a one-sided Bayesian independent samples t-test to test for the strength of evidence in favour of the null hypothesis (no difference between conditions) as opposed to the hypothesis that the score in the contrastive condition is higher than the score in the non-contrastive condition using JASP software with default priors. We obtained BF_{10} = 2.5, meaning that the data was 2.5 times more likely under the alternative hypothesis than under the null hypothesis. While this is not particularly strong evidence in favour of the alternative hypothesis, the data does show the same pattern as observed in Study 1. The diminished difference between conditions is likely due to that in Study 1, having two cases to compare, the participants noticed that the second set of objects changed (i.e., animals from Marchena), and this in turn strengthened the perceived contrast.

2.4. Interim discussion

The results of both studies were in line with our Hypothesis 1: the probability of the feature f given a contextual salient alternative did affect the acceptability of a generic sentence *'Gs are f'*. We also saw the direction of the dependence predicted by our theory confirmed: if $P(f|G)$ is substantially larger than $P(f|Alt(G))$ then the acceptability of the generic sentence is higher than in case there is no difference between both probabilities. We did not see the exact acceptability scores that the theory predicts (Hypothesis 2). In the non-contrastive condition, the theory predicts an acceptability of 2.5, while in Study 1 the average acceptability in this condition was 2.88 and in the Study 2 3.06 with 95% confidence intervals around mean not including the expected value in either case. In the contrastive condition, we predicted an acceptability

of 4.5 and observed an average of 3.51 in Study 1 and 3,49 in Study 2, again with the 95% confidence intervals around the mean not including the expected value.

Contrary to our expectation, the participants were not uniform in the scores they were giving - we observed large differences between participants' behaviour, so it in fact it does not make much sense to look at the overall means as we set out when we started this project.[6] However, this observation does not necessary contradict the theory tested here. The predictions made by contingency as measure of the acceptability of generic sentences depends on which alternatives to G the interpreter considers. We assumed that the setup of the study would lead the participants to consider the sample from Marchena as alternative to the sample from Genovesa that the generic talked about. The theory predictions outlined above are only valid if the participants took the alternative into account. However, we cannot be sure that the participants really did take the sample from Marchena to be a relevant G alternative. If they did not take any alternatives to the target group into account, the theory predicts the acceptability of the generic sentence to be equal to the conditional probability $P(f|G)$. Consequently, the acceptability value assigned by the participants would be 4.

To explore this possible interpretation of the data, we separated the participants of the Study 1[7] into three groups: those that assigned the same acceptability rating to the generics in both conditions (difference between scores in two conditions less than 0.5[8]), those that judged the generic in the contrastive condition to be at least 0.5 points more acceptable and those who considered the generic at least 0.5 points less acceptable. 51% (N=40) of the participants in the first study did not give a substantially different score in two conditions, while 38% (N=30) considered the generic in the contrastive condition more acceptable than in the non-contrastive condition and 11% (N=9) of the participants took the generic to be less acceptable. We then looked at the scores given by participants in the first two groups[9]. If Hypothesis 2 is correct but only participants in the group that gave a higher score to the contrastive condition took the sample from Marchena as an alternative to the sample from Genovesa, these participants should have given the scores predicted by Hypothesis 2 whereas the participants in the group that did not take into account the sample from Marchena should have given score 4 in both conditions (as discussed above). This was not the case. In the group of participants that gave a higher score in the contrastive condition than to the non-contrastive position, the average acceptability in the contrastive condition was 3.86 (SD 0.79; 95% CI [3.57, 4.14]) whereas the average acceptability in the non-contrastive condition was 1.72 (SD 1.22; 95% CI [1.28, 2.15]). Thus, even in this subgroup of participants, the scores come close to the ones predicted by theory, but we do not observe the exact values predicted by Hypothesis 2. The group that did not see a difference gave a mean score 3.35 (SD 1.18; 95% CI[2.98, 3.71]) in the contrastive and a mean score 3.4 (SD 1.22; 95% CI [3.02, 3.77]) in the non-contrastive condition.

[6]Note that we report the mean values and statistics with the whole group despite this since we committed to an analysis plan before we collected data.

[7]This was not possible for the second study since we used a between-participants setup in that case.

[8]This is an arbitrary threshold that we chose. We assumed that a difference of 0.5 could arise from the participants trying to drag the slider to the same point on the scale, whereas larger differences would necessarily arise from intentional positioning of the slider at different points of the slider.

[9]We will not discuss the participants in the third group which gave the non-contrastive condition a higher score than the contrastive condition further as we do not know why they behaved like that. They could have not understood the instructions or they could have changed their interpretation of the target sentence halfway through the experiment.

There are a couple of remarks we want to add about the discrepancies between the acceptability values predicted by the theory and the data obtained in the study. First of all, it is difficult to say how exactly the participants interpreted the scale that we asked them to use to indicate the acceptability of the generic sentences they saw. We tried to avoid the ambiguity by labeling the extremes of the scale verbally as 'not at all' and 'certainly', but cannot be sure what the participants did in case they were not sure about acceptability of the sentence (when it is neither acceptable nor unacceptable).

Depending on how the participants interpreted the scale, there might be also an issue with the way we interpreted the numerical values that our theory predicts. The range of the contingency function is the interval $[-1,1]$. We took this to mean that -1 corresponds to a completely unacceptable sentence, 1 to a sentence that is completely acceptable and 0 describing the turning point from not acceptable to acceptable. This is how we translated the values of the contingency function to the scale that we presented to the participants of both studies. To some extend this is also confirmed by the data. The obviously wrong filler items got average acceptability judgments that were very close to 0. However, there is no guarantee that even if the acceptability of generic sentences can be described in terms of contingency, as we proposed, the values are interpreted in the linear manner that we assumed. Maybe a 0 for contingency already means that we wouldn't accept the sentence. To avoid such issues, we could show the participants a scale with numerical values from -1 to 1 instead 0 to 5 as we did here and see whether this affects their acceptability judgments for the same set of test data. This will need to be taken up in the follow-up research.

To sum up, in general the results support the theory proposed here, though we did not see exact scores that we expected. As discussed above, this could be because we did not transform the values from the theory to the scale seen by participants correctly. For this, more research in the future is necessary. What we can assess is in how far the theory explains the general tendencies in the data that we gathered, and in this respect the results are encouraging.

2.5. Study 3

The main goal of this final study was to test a different aspect of the theory developed in Section 1. We repeat here for reasons of convenience Hypothesis 2, which contains the heart of the proposal.

Hypothesis 2 *The acceptability of a generic sentence 'Gs are f' is given by the formula*

$$\textit{Acceptability of } {}'\textit{Gs are } f' = P(f|G) - P(f|\bigcup \textit{Alt}(G)).$$

So far, we have focussed on testing whether we can observe the predicted effects of manipulating the second argument of the measure of acceptability. We saw that indeed $P(f|AltG)$ does affect the acceptability of generic sentences and also that the kind of influence predicted (acceptability goes up if $P(f|AltG)$ goes down) can be observed. In this study, we focused on the first part of the measure: $P(f|G)$. Manipulation of this factor should, according to our theory, also have an effect on the acceptability of a generic. Roughly put, increasing this variable

should have a positive effect on the acceptability ratings.

As a side question, we also wanted to test with this study whether another new aspect of our proposal can be confirmed by the data. Note that the approach introduced in Hypothesis 2 also differs from the one described in Definition 2 in measuring the acceptability of generics in degrees instead of proposing cut-off points that define the limit between being or not being acceptable. For instance, if alternatives do not play a role, then Hypothesis 2 predicts a steady linear increase in the acceptability of the generic with growing $P(f|G)$. In some sense, the data of the first two studies already speak against a clear cut-off point of 0.5, given that even though $P(f|G)$ was $0,8$ the acceptability ratings were not close to ceiling.[10] Given that in this final study we consider different conditional probabilities $P(f|G)$, the results should provide us with a clearer picture of whether the cut-off approach or the gradual change approach defended here come closer to reality.

In this last study, we used the same set-up as in the first two studies. The participants judged the acceptability of generic sentences with respect to the two conditions, the non-contrastive condition in which $P(f|G) = P(f|Alt(G)$ and the contrastive condition in which $P(f|Alt(G)) = 0$. The only difference is that now we varied $P(f|G)$ between participants.

As Study 3 was a follow-up to the first two studies, this time we assumed from the start that there will be two groups of participants. Participants that do not take alternatives into account when evaluating the generic sentence (we will refer to this group as *noCon*) are predicted to use the conditional probability of the feature f given the group G as measure of the acceptability of the generic sentence. In this case, our theory predicts that in both conditions the acceptability of the generic should increase linearly with a growing conditional probability $P(f|G)$. For participants that *do* take the presented alternative into account (group *Con*) the acceptability score should depend on $P(f|G)$ and $P(f|Alt(G)$. In the contrastive condition, $P(f|Alt(G))$ is 0 while $P(f|G)$ is not, so again the acceptability of the generic sentence should grow linearly with the increase in $P(f|G)$. Furthermore, we predict that the acceptability ratings for this condition should overall be slightly higher (approximately 0.5 points) for the *Con* group that for the *noCon* group.[11] In the non-contrastive condition, both $P(f|G)$ and $P(f|Alt(G))$ are identical so the contingency of the sentence is 0. In this case, for the *Con* group there should be no effect of proportion on the acceptability of the generic sentence - the acceptability score should be the same independent of $P(f|G)$.

[10] Cohen could argue that this is because some or all of the participants applied the relative reading of generics. However, notice that even after we split participants into groups according to whether they saw a difference between the two conditions, those that did not see a difference still did not give a ceiling acceptability score to the generic sentence. Furthermore, in the relative reading, Cohen would predict that still the generic should be completely acceptable in the contrastive condition and completely unacceptable in the non-contrastive condition, which is again not what we found.

[11] The reason for this is a difference in how $P(f|G)$ counts for acceptability for participants that take alternatives into account and those that don't. The acceptability rating of a participant that doesn't take alternatives into account in the condition where 80% of the animals carries the relevant feature, for instance, should be $P(f|G) * 5 = 4,0$. But a participant that takes alternatives into account should give in the contrastive condition a rating of $\frac{P(f|G)+1}{2} * 5 = 4,5$.

2.5.1. Method

Materials This study had the same design as Study 1, but now we collected data for different proportions with which the animals possessed the relevant color feature. We used four proportions: 54%, 68%, 80%, and 92%.[12] Furthermore, we also varied the distribution of the feature among the 25 animals that were shown to the participants: for each condition we used 3 pictures with different, randomly selected distributions of the feature over the presented animals.

Each participant had to make three judgments: she saw one picture in the contrast condition, one picture with the no contrast condition and one filler, all using the same frequency for the distribution of the feature. Each animal species was shown once. The order of the contrast/no contrast question was randomised, the filler was always shown as the third and last question.[13]

Participants Participants were again recruited via Prolific.ac with the same criteria. 401 participants completed the task. Twenty participants were excluded because they gave inadequate responses to the filler items (score above 1.5). Six further participants were excluded because they gave all three conditions a score 0. 375 participants were thus included in the analyses reported below: 97 for frequency 54%, 89 for frequency 68%, 94 for frequency 80%, and 95 for frequency 92%.

2.5.2. Results

Because the condition in this study where $P(f|G) = 0,8$ is exactly the same as what we presented in Study 1, we start by inspecting the results for participants that saw this condition (N=94) to check for the robustness of the results we obtained there. For this group, the mean score in contrastive condition was 3.50 (SD 1.25), whereas the mean score in non-contrastive condition was 2.88 (SD 1.47). When split into groups, there were 32 participants (34%) who gave the contrastive condition a higher score (difference more than 0.5) than the non-contrastive condition and 58 participants (61%) who gave them the same score (difference less than 0.5). Both the averages and the proportions of participants in each group are close to what we observed in Study 1. Hence, these findings are robust.

As stated above, in this study we distinguish two groups of participants: group *Con* contains participants that found the generic more acceptable in the contrastive condition than in the non-contrastive condition; participants in group *noCon* did not give a different score in the two conditions. We split the participants into these two groups using the same criteria as we used in Study 1. There were 135 participants (36%) who gave a higher score in the contrastive condition (*group Con*). When collapsing across different proportions, this group gave a mean score 3.69 (SD 0.97) in the contrastive condition and a mean score 2.0 (SD 1.21) in the non-contrastive condition. There were 209 participants (55%) who gave the same score in two

[12] All sample-pictures contained 25 animals of one species, see Figure 3. Thus, for example, in the contrastive condition a proportion of 54% means that 14 out of 25 animals in the sample from Genovesa have the property and none of the animals in the sample from Marchena. In the non-contrastive condition in both samples 14 out of 25 animals would have the property.

[13] As a consequence, the trials using 80% were a complete replication of the first study. We will come back to this in the discussion of the results.

conditions (*group noCon*). This group gave a mean score 3.2 (SD 1.26) in the contrastive and a mean score 3.18 (SD 1.25) in the non-contrastive condition. Finally, there were 18 participants (9%) who gave a higher score in the non-contrastive condition. The table in Figure 5 shows the results for the different probabilities split up according to the two groups that we distinguish.

		group	**Con**		**group**	**noCon**	
condition	$P(f\|G)$	Mean	SD	N	Mean	SD	N
contrast,	54%	3.34	1.04	37	2.80	1.01	49
$P(f\|Alt(G) = 0$	68%	3.67	0.69	35	2.81	1.32	43
	80%	3.79	1.20	32	3.43	1.22	58
	92%	4.06	0.81	31	3.59	1.30	59
no contrast,	54%	1.71	1.15	37	2.78	1.00	49
$P(f\|Alt(G) = P(f\|G)$	68%	2.20	0.91	35	2.77	1.27	43
	80%	1.89	1.40	32	3.41	1.21	58
	92%	2.27	1.38	31	3.58	1.32	59

Figure 5: Results of study 3

To test our predictions, we conducted a Bayesian ANOVA with condition (contrastive vs. non-contrastive) and proportion (as an ordinal variable) as independent variables for each group separately. To evaluate whether a certain variable has an effect on the given scores, we compared a model including this effect with a model excluding this effect. For the group that gave the same score to both conditions (*group noCon*), we predicted an effect of proportion - the scores should linearly increase with increasing proportions. In the ANOVA analysis, we observed modest evidence against the effect of condition ($BF_{Inclusion} = 0.2$, given by the definition of the group), strong evidence for the effect proportion ($BF_{Inclusion} = 13$), and strong evidence against the interaction of condition and proportion ($BF_{Inclusion} = 0.02$). Thus, we do observe an effect of proportion. However, while the participants did give a higher score with increasing proportions, this increase does not seem to be equally present for all proportion steps. A post-hoc test comparing each proportion to the other ones showed that scores given for proportion 54% were not different from scores given for proportion 68% ($BF_{10,U} = 0.16$), and scores given for proportion 80% were not different from scores given for proportion 92% ($BF_{10,U} = 0.22$); for the other proportion pairs, we had evidence for the difference in scores. Thus, participants here did not seem to care about the difference between the lowest two proportions and the highest two proportions, exhibiting rather behaviour that would correspond to there being some sort of threshold between $P(f|G) = 68\%$ and $P(f|G) = 80\%$.

For the group that gave the contrastive condition a higher score than the no contrast condition (*group Con*), we predicted an interaction between condition and proportion: the scores given by participants should linearly increase with increasing proportions in the contrastive condition, but they should be the same across proportions in the no contrast condition. In the ANOVA analysis, we observed extreme evidence for the effect of condition ($BF_{Inclusion} = \infty$), inconclusive evidence for presence or absence of the effect of proportion ($BF_{Inclusion} = 0.8$) and modest evidence against the interaction of condition and proportion ($BF_{Inclusion} = 0.2$). Hence, based on our analysis, here the predictions were not borne out - the effects of condition and proportion did not clearly interact. When inspecting averages for each proportion in the two conditions, there *does* indeed seem to be a gradual increase of the scores in the contrast condition in this

group, whereas in the no contrast condition there seems to only be a random fluctuation of the scores. But even if we focus only on the judgments for the contrastive condition, there is no evidence for an effect of proportion. It seems like the increase in scores was not consistently present for all participants (see Figure 6 for a depiction of the individual scores).[14]

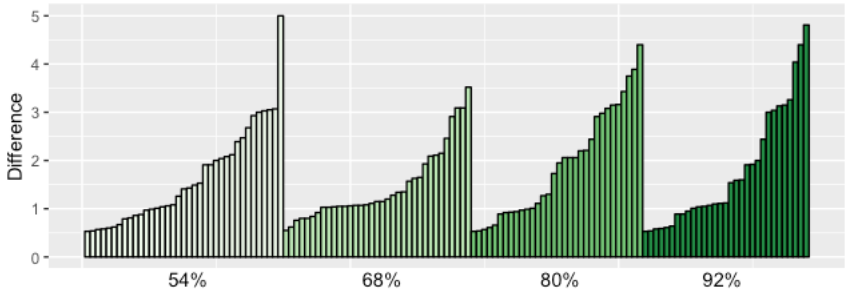

Figure 6: This plot depicts the difference between contrastive and non-contrastive condition (on the Y axis) for each of the 135 participants of the Con group (on the X axis). We grouped the participants by the proportion that they saw. We can see that it is not the case that there are mostly higher scores for higher proportions. NB: each participant saw only one proportion.

2.6. General discussion

All three studies that we reported on seem to confirm Hypothesis 1: for many participants the acceptability of a generic sentence *'Gs are f'* depends on the conditional probability of the feature f given salient alternatives G' of G. We also found evidence for the type of dependency predicted by our proposal: if the feature f is much more frequent given G than given the alternative G', then the acceptability of the generic improves. Study 1 and study 2 did not confirm the exact acceptability scores predicted by the theory, but as discussed in Section 2.4, this might have to do with the particular methodology we used. In particular, our proposal for transformation of the scores in our task to those predicted by the theory might not be accurate.

With study 3, we wanted to investigate whether the predicted dependency on the absolute probability of f given G is also supported by empirical results. Based on the discussion in Section 2.4, we now immediately distinguished two groups within the participants: group *Con* consisted of participants that judged the generic more acceptable in the contrastive condition, while in group *noCon* were those participants that gave the same scores in the two conditions.

For the group *noCon*, the results of study 3 supported *a* dependence of acceptability on proportion: the acceptability increased with the probability $P(f|G)$, independent of condition. But, as discussed above, we could not confirm the predicted linear increase in acceptability that Hypothesis 2 predicts. Instead, there was some evidence for an acceptability threshold between the second and third condition of proportion. This provides some evidence for threshold theories like the one of Cohen (1999), though the value of the threshold clearly seems to differ from the 50% threshold that Cohen proposes. Also the values below the threshold are not what one

[14]The reader might notice that in the 54% condition only one participant had a very large difference - 5, and there is no other participant in other proportions with such a large difference between the contrastive and the non-contrastive condition. One might think that maybe this participant is the reason why we do not observe an effect of proportion. But excluding this participant does not affect the results.

normally would expect. Even in the conditions with $P(f|G) = 54\%$ and $P(f|G) = 68\%$, the generics still where not clearly rejected, but on average still marginally acceptable. We need more empirical data, also for different conditions of proportions to be able to say whether we should prefer a threshold account and what form exactly it should take.

For group *Con* we could not confirm an interaction between condition and proportion. Note that the mean acceptability score given to the generic in the contrastive condition did steadily increase with growing conditional probability of the feature f given the group G, and in a rate that comes close to what is predicted by the theory. However, statistically the result was not significant. Here, either the theory is wrong or perhaps our experiment was not tapping into the interpretation/significance of alternatives clearly enough to reliably detect the difference. One reason for this could be that this effect (i.e., the increase in scores due to increasing $P(f|G)$) is rather small, so our sample size of approximately 30-35 participants in each group is not large enough to detect it. In this connection, notice also the surprising low acceptability ratings of group *Con* for the non-contrastive condition. The theory predicts an acceptability value of 0 in this case, independent of $P(f|G)$, which should correspond to a score 2.5 on the scale the participants saw in our study (with our transformation). However, in study 1 and for all four proportions in study 3, the given acceptability score was lower than that and varied quite a lot. We already discussed in Section 2.4 that a possible explanation might lie in the way people interpreted the scale on which they gave their judgments.

Let us turn to the relevance of the data from the group *Con* for the cut-off point hypothesis built into theories like the one proposed in Definition 2 in contrast to the gradual increase in acceptability that Hypothesis 2 predicts. As discussed above, for the group *noCon* there was some evidence for a cut-off point between $P(f|G) = 68\%$ and $P(f|G) = 80\%$. In contrast, for the group *Con* we do not see the same 'jump' in acceptability ratings between proportions. Instead, as discussed above, at least in terms of just the means there appears to be a linear increase of acceptability in the contrastive condition. From a theoretical point of view this observation is rather difficult to make sense of. Why should there be a cut-off point in case no alternatives are taken into account, while acceptability increases linearly in case alternatives do matter? Of course, we could easily propose an ambiguity with two possible readings of generics; one with threshold, one without. But that seems to be an awfully arbitrary difference between two readings of the same sentences. Before we take such a theoretical step we need more evidence that this difference is real. To conclude, our results do not support a clear threshold account, as, for instance, defended in Cohen (1999). But also the linear increase of acceptability with growing $P(f|G)$ that Hypothesis 2 predicts is not completely supported by our data.

Finally, there is one more curious feature of the behaviour of participants in study 3. Even though the few datapoints we recorded do not allow us to test for it, notice that the size of group *noCon* appears to increase with growing $P(f|G)$. Using the terminology of our proposal, the higher the absolute probability of f given G the less relevant alternatives to G seem to be. There is some evidence from related domains, as studies of causal judgments, showing that actually $(P(f|G)$ counts more for the acceptability of such judgements (Wasserman et al. 1993, Anderson and Sheu 1995). Using a measure that takes this into account and, for instance, weights $P(f|G)$ more the larger this factor is, could explain the tentative observation just made. The higher $P(f|G)$, the less the contrastive value $P(f|Alt(G))$ would count for acceptability,

and, hence, the smaller the difference between the contrastive and the non-contrastive condition. Consequently, more people would look like belonging to the group *noCon* instead of the group *Con*. Thus, if this tentative observation just made could be confirmed by a study suitable to test it, it might give us an important hint for how to improve the proposal made here.

3. Conclusions

In this paper, we discussed the relevance of alternative groups for the interpretation of generic sentences *Gs are f*. This has led us to a first and preliminary formal description of the meaning of generic sentences, given in Definition 2. According to this approach, we have to distinguish two readings for generic sentences: a relative reading that does take alternatives to G into account, and an absolute reading that does not. The proposal is basically that in Cohen (1999).

We motivated the relevance of the relevance reading of generics, and thus of alternatives to G, by linking this meaning to associative learning. Building on theories of learning from psychology, we formulated a new and final version of our approach. According to this proposal, acceptability of a generic 'Gs are f' should be measured in terms of the strength of association of the group G with the feature f. To have a concrete proposal that we could test, we used contingency to measure this strength of association. This proposal differed from the approach we formulated at the end of the first part of the paper in two important respects. First of all, it predicts the acceptability of generics to come in degrees. More concretely, this means that our proposal does not assume strict cut-off points for the truth or acceptability of generics. Secondly, the proposal assumes not two, but only one (context-dependent) reading for generic sentences. This reading is the relative reading of Definition 2. The reading can in certain circumstances – if the alternative set the interpreter assumes for G is empty – collaps to the absolute reading of Definition 2.

In the second and main section of the paper, we reported on three studies that tested our proposal. In these studies participants were presented with a visual scene and asked to judge the acceptability of a generic sentence 'Gs are f'. We manipulated the presence of the alternatives and the frequency with which members of group G carried feature f. The results allowed us to confirm the relevance of G-alternatives for the meaning of generic sentences in the population in general. We also observed some evidence for the correlation between acceptability of generic sentences and $P(f|G)$. However, not all particular predictions made by the proposal were borne out.

Interestingly, there seemed to be at least two groups of participants based on the acceptability scores they gave in different conditions. One group did seem to take into account the alternative to G, whereas another group did not seem to do it. This difference can be explained from the perspective of the theory tested here: some participants did not accept the presented alternative as salient and adopted an absolute reading of the generic sentence. But it also hints at a weak point of the proposal: it remains silent on the question what the relevant alternatives are that a speaker considers. Why is it that the alternatives that we presented were ignored by more than half of the participants?

Also, we did not obtain the exact acceptability scores in different conditions that the theory predicted. Here, the question of how we transform the values from the theory to the scores given by participants is relevant (see below). Finally, based on the scores given by participants

in study 3, one group of participants exhibited behaviour compatible with there being a certain threshold, albeit not exactly what is expected under Definition 1. All of these empirical observations call for further work on the theory proposed here, but also certain methodological questions need to be addressed by future research.

One aspect of the used methodology that is important to note is that we did not model sequential learning in our experiments. A central idea of the theory proposed here is that acceptability of a generic sentence is equated with the strength of association built based on the frequency with which the agent observed members of a group carrying a particular feature. However, we did not allow the learning of the association to observe these occurrences sequentially. Instead, we just gave the participants of the studies the information in one batch. It would be good if we found an experimental setup that modelled learning in a more natural way.

Another aspect of the methodology applied that we would like to improve on in future work is the way we mapped the responses we get from the participants to the very precise numerical values of the theory used. Whether our data does or does not correspond to theory depended largely on how exactly one transforms these scores. Note that it is rather unusual for experimental psychology to formulate predictions in terms of specific scores as we did here, because it is assumed that there is too much uncertainty about what people are doing to have such precise predictions; traditionally, only presence or absence of differences between conditions is tested instead. We believe that formulating and testing more specific numerical predictions is a good way to reduce the gap between theories like the one about the meaning of generics presented here and experimental findings with human participants. But we also realise that methodologically this presents a number of challenges that we haven't solved completely yet.

Though the most pressing challenges for future work on the topic explored here are arguably methodological in nature – we need a solid empirical basis in order to direct further theoretical work – there are also a couple of interesting theoretical questions that we want to explore in future work. Just to mention one example, we picked contingency to measure associative learning. However, there are other measures of strength of association discussed in the literature. We should test those as well on the data-set gathered here and compare the predictions made with those of contingency.

References

Anderson, J. R. and Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition*, **23**: 510-524.

Cohen, A, (1999), *Think generic! the meaning and use of generic sentences*, CSLI Publications, Stanford.

Gluck, M. A. and Bower, G. H. (1988), 'From conditioning to category learning: An adaptive network model', *Journal of Experimental Psychology: General*, **117**: 227-247.

Krifka, M., F. Pelletier, G. Carlson, A. ter Meulen, G. Chierchia, and G. Link (1995), 'Genericity: An introduction', In G. Carlson and F. Pelletier (eds.) *The Generic Book*, pp. 1-124. University of Chicago Press, Chicago.

Leslie, S.J, (2008), 'Generics: cognition and acquisition', *The Philosophical Review*, **117**: 1-47.

JASP Team (2018). JASP (Version 0.9) [Computer software]. Retrieved from: https://jasp-stats.org/.

Palan, S., and Schitter, C. (2018). Prolific. A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22-27.

Rescorla, R.A. (1968), 'Probability of shock in the presence and absence of CS in fear conditioning', *Journal of Comparative and Physiological Psychology*, **66**: 15.

Rooij, R. and K. Schulz (2019), 'Generics and typicality: A bounded rationality approach', *Linguistics and Philosophy*. DOI: 10.1007/s10988-019-09265-8

Tessler, M. and N. Goodman (2019), 'The language of generalization' *Psychological Review*, 126(3), 395-436.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., and Baker, A. G. (1993). Rating causal relations: The role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**: 174-188.