Are All Generics Created Equal?

Francis Jeffry Pelletier
Simon Fraser University

Introduction

Generic statements of the sort under consideration in this article are those such as

(1) a. Potatoes contain vitamin C
b. Basketball players are tall
c. Predatory animals have sharp teeth
d. Birds fly

Such genericity is a feature of entire sentences (or at least of entire clauses), and is not attributable to any subpart of the sentence, such as to the subject noun phrase or to the verb phrase. It is instead somehow an interaction between the two. This sets it apart from another type of genericity, which is attributed to the noun phrase's referring to an abstract genus (or kind) and not as making a statement about the individual members of that genus, as in

(2) a. The potato was first grown in South America
b. Dodos are extinct
c. Mosquitos are common in northern Canada
d. Man landed on the moon in 1969

Although there are various commonalities between the two types, and certain relations that hold between them, we are here interested only in the first phenomenon, which was designated *characterizing genericity* in Krifka et al. (1995).

One of the logically, semantically, and psychologically interesting features of generic statements is that they "tolerate exceptions." That is, they can be true despite the fact that there are some particular instances that do not possess the generically-predicated property. For example, (1-a) is true despite some spoiled potatoes, and (1-d) is true despite ostriches, emus, penguins, kiwis, cassowaries, and the like. This feature is covered extensively in Krifka et al. (1995), and will simply be assumed in the present article.

Characterizing genericity can be expressed not only by the "bare plural" formulation given in (1) but also with definite (3-a) and indefinite count noun phrases (3-b), and as well, by mass nouns ("bare singulars", (3-c)). In addition, the subject term can be a proper name (3-d), although in that case, the characterizing genericity phenomenon is usually called 'habitual'.

(3) a. The potato is highly digestible
    b. A snake is a reptile
    c. Snow is white
    d. Fred has wine with dinner

Not only can the item that is being claimed "to generically have a property" occur in the subject position as all these examples suggest, but it can also occur in object positions and indeed, there can be more than one in a sentence, as these examples show

(4) a. Cats chase mice
    b. Canadians are taller than Mexicans
    c. Paul smokes cigars after meals
    d. Italians love pasta

Many of these types of generics raise special puzzles within formal semantics, and their position within the pantheon of generics is not yet settled. For example, it was pointed out in Lawler (1973) that the indefinite formulation as in (3-b) only works when the predicate is somehow "essential" to the subject term. And comparatives like (4-b) seem to call for some special treatment that involves comparisons between averages (Carlson and Pelletier, 2002). Habituals like (3-d) and (4-c) seem to generalize over events, rather than over individuals, as the others do. Furthermore, mass terms themselves bring a set of unique problems (Pelletier and Schubert, 2003), and it might be best to await a semantic treatment for them before tackling generics like (3-c). Finally, there might be important semantic differences between definite generics as in (3-a) and the bare plural formulations in (1), so we perhaps should reserve them for separate treatment.

Thus, the present article discusses only variants of the bare plural formulation, and then only in subject position of simple sentences that do not involve a relational statement. We furthermore reserve for a separate treatment the types that are mentioned in (2), even when they employ bare plurals. (The common attitude is that the types of predicates in these sentences are of a different sort than the more traditional generic predicates in (1), but we do not wish to engage that discussion here.)

But even within just the indicated type of generics, one can find distinct formulations. For example, we might see (1-c) stated in any of these ways (and others):

(5) a. Predatory animals have sharp teeth. (BP)
    b. Most predatory animals have sharp teeth. (Q)
    c. The typical predatory animal has sharp teeth. (M)
    d. Predatory animals usually have sharp teeth. (Adv-1)

e. Predatory animals normally have sharp teeth. (Adv-2)

The (a) version is called the "bare plural" formulation (BP), the (b) version is the "quantificational" formulation, the (c) version called the "noun modifier" version, and the final two are labeled two different "adverbial" formulations. There are other forms that can be found in the literature, but these are representative of the variety that are mentioned. Mostly authors will mention all of these as merely alternative formulations of the same semantic item—that is, they are claimed to have the same generic force or meaning.

There has been, however, an opinion that there is a difference in the ways that these generic sentences have been given a theoretical footing. This view is commonly attributed to Greg Carlson (see, e.g., Carlson, 1980, 1982, 1995), who calls them "the inductivist approach" and "the rules and regularities approach". The former approach has a driving intuition that our generic sentences basically express inductive generalizations, where the basis of the generalization is some observed set of instances. The idea is that, after observing a number of instances of predatory animals with sharp teeth, a sentence like (5-a) is generated as a covering description. In describing the philosophical backdrop of this view of generic sentences, Carlson (1995, p. 225) says "the most natural bedfellows of this approach would be empiricists, verificationists, and nominalists of varying stripes." A different background view is the latter approach, which does not hold that generic are truly asserted on the basis of *any* array of instances, but rather that they depend for their truth or falsity upon whether or not there is a causal organization within the world that corresponds to them. The philosophical perspective of those who would adopt this view are those who, in Carlson's words (1995, p. 225), "admire properties and propositions as real entities,... as would many realists."

The present paper is an investigation into the status of these sorts of claims. There are two related aspects: first is the view, taken by almost all writers on generics, whether within philosophy or linguistics or artificial intelligence, that the sentences in (5) all express genericity equally; and second is the view expressed by Carlson that there is a difference between two theoretical attitudes towards the grounds for the truth of generic sentences. Note that Carlson's view is that each of the two theoretical viewpoints would hold that all the sentences in (5) express genericity equally; it is just that they differ in what genericity amounts to.

This study aims to challenge the view that all genericity is viewed the same way by ordinary speakers of natural language (in this case, English). Intuitively, a sentence such as (5-b) ought to be a paradigm case of the inductivist approach since *most* seems clearly to call for some sort of explicit comparison of the number of predatory animals with versus without sharp teeth. And again, the version worded as (5-a) seems most straightforwardly to embody the rules and regularities approach, since we seem to be appealing to the very kind itself, Predatory Animals, and its regulatory properties. The other versions in (5) can be seen as going either way, or perhaps even changing their interpretations depending on the context or example under discussion. We show that there are at least two different types of generic statements, and that they are lexically differentiated. A natural interpretation might be that the two types correspond to Carlson's two viewpoints, and that each viewpoint finds itself most clearly represented by different ones of the sentences in (5). But that is a further interpretation to put upon the present findings. What we can say is that our findings are consistent with the view that the two different background interpretations for characterizing generics find expression in two syntactically different sentence-types.

Background to Default Reasoning

Default reasoning occurs whenever the evidence available to the reasoner does not guarantee the truth of the conclusion being drawn; that is, does not deductively force the reasoner to draw the conclusion under consideration. ('Force' in the sense of being required to do it if the reasoner is to be logically correct). But nonetheless the reasoner does draw the conclusion, and is correct in doing so. For example, from the statements 'Linguists typically speak at least three languages' and 'Kim is a linguist', one might draw the conclusion, by default, 'Kim speaks at least three languages'. What is meant by the phrase 'by default' is that we are justified in making this inference because we have no information which would make us doubt that Kim was covered by the generalization concerning linguists or would make us think that Kim was an abnormal linguist in this regard. Of course, the inference is not deductively valid: it is possible that the premise could be true and the conclusion false. So, one is not forced to draw this conclusion in order to be logically correct. Rather, it is the type of conclusion that we draw "by default"—the type of conclusion we draw in the ordinary world and ordinary circumstances in which we find ourselves.

Default reasoning is non-monotonic, i.e., adding new premises can make us withdraw previously-generated conclusions without withdrawing any of the previous premises. For example, were we to add to our list of premises the further fact that Kim graduated from NewWave University, which we know has revoked all language requirements, we then would wish to withdraw the earlier conclusion, even though we would not withdraw any of our other premises.

Default reasoning makes its appearance in many academic disciplines. Pelletier and Elio (2005) cite examples from ethics, philosophy of science, conditional and counterfactual logics, relevance logics, studies of prototypical and stereotypical schemata, causal reasoning, medical and fault diagnosis, reasoning in the social sciences, judgments under uncertainty, reasoning involving (Gricean) implicatures, lexical defaults in linguistics, "natural" logic (as pursued, e.g., by Lakoff, 1972, 1973), knowledge representation in AI, and some argumentation about the nature of cognition in Cognitive Science. In fact, the leading examples of all these fields are what would be called characterizing generics by Krifka et al. (1995).

Lifschitz (1989) published a set of "benchmark problems" that all formalisms for default reasoning are supposed to follow. These examples covered a wide variety of different applications; the ones I am interested in for the present study form the first part of his paper: three of the four problems he called "Basic Default Inference". Here are the three, as presented in Lifschitz (1989):

Benchmark 1.
Blocks A and B are heavy.
Heavy blocks are normally located on this table.
A is not on this table.
Therefore, B is on this table.

Benchmark 2.
Blocks A and B are heavy.
Heavy blocks are normally located on this table.
A is not on this table.

B is red.
Therefore, B is on this table.

Benchmark 3.
Blocks A and B are heavy.
Heavy blocks are normally located on this table.
Heavy blocks are normally red.
A is not on this table.
B is not red.
Therefore, B is on this table.

Each of these problems concerns two objects governed by one or more default rules. Additional information is given to indicate that one of the objects (at least) does not follow one of the default rules. We refer to this as the exception object (for that default rule). The problem then asks for a conclusion about the remaining object, which we refer to as the object-in-question. For all these problems, Lifschitz endorses the conclusion that the object-in-question (Block B) obeys the default rule concerning location. According to the collective wisdom of researchers into nonmonotonic theories, the existence of an exception object for a default rule (Benchmark Problem #1), or additional information about that exception object, should have no bearing on a conclusion drawn about any other object when using that rule. Extra information about the object in question itself (e.g., Block B's color in Benchmark Problem #2) should also have no bearing on whether a default rule about location applies. And being an exception object for some other default rule should have no bearing on whether it does or does not follow the present default rule (Benchmark Problem #3). I call these the "AI-approved answers" to the Benchmark Problems.

It will be noted that the default rules used in Lifschitz's problems are in fact the Adv-2 versions of our variant ways of stating generics given in (5). So we might ask whether people treat this version of a generic statement the same or different from the other versions of the same content. One problem with using Lifschitz's problems directly is that the generic statement (i.e., the default rule) is what might be called a nonce-generic, that is, a spur-of-the-moment invention. We should wish to investigate generics that are more "stable" in their interpretations. But in doing so we need to avoid well-known generic truths (and falsehoods), since this knowledge could affect subjects' answers about whether or not the conclusion follows. For instance, we shouldn't ask about whether 'Tweety the penguin flies' follows from 'Birds normally fly' and 'Penguins are (always) birds', since it is a piece of stored knowledge about penguins that they don't fly. Hence, subjects wouldn't be calling upon any reasoning module to answer such a question, but would instead just be engaged in some sort of memory look-up.

Some theorists have also thought that generic statements about natural kinds might be internalized differently by people than ones about artifacts, since the former maybe are governed by "natural laws" while the latter could just have "accidental facts" be true about them. (Or maybe the other way around: things in the artifactual world are made for a purpose, whereas things in the natural world exhibit just whatever behavior happens to hold for them.)

For these reasons I made up example generics using both made-up "natural" kinds and made-up "artifactual kinds". A cover story was used to introduce the kinds, and to make clear whether or not they were natural or artifactual kinds. The story went on to characterize the kinds along

the lines of the Benchmark problems, by giving the "generic information", and the subjects were asked to evaluate how well the story supported the Benchmark conclusion. The goal then was to investigate whether the five different ways of giving the generic information mentioned in (5) above would affect how strongly the subjects thought the conclusion followed.

Method and Results

Two experiments were conducted: the first was a large-ish pilot study that attempted to determine whether there were likely to be detectable effects in asking questions about the Benchmark Problems when using only slightly syntactically different formulations. Despite the lack of controls for various factors, the results of this pilot study are interesting for their descriptive data. A follow-up experiment attempted to measure more carefully the factors involved. Both experiments used the same general type of setup, to which we now turn.

A Pilot Study

Materials in Pilot Study

As remarked above, subjects were presented with "cover stories" that characterized some kind — natural or artifactual — which they would not have heard about before (since they were made up for the experiment). These cover stories were described in the instructions to subjects as "paragraphs taken from newspaper or magazine stories", and they described various features of these kinds. The things that were mentioned about the kinds mapped directly to the features mentioned in the original Benchmark Problems, and the subjects were then asked to describe the extent to which they thought the conclusion (which also mapped directly to the Benchmark Problem) followed from the information presented.

The pilot experiment was of a $3 \times 2 \times 6$ design: three Benchmarks (BM1, BM2, BM3), two types of Kinds involved (natural vs. artifactual: NK, AK), and six different types of information. Five of these types were generic information (BP, Q, M, Adv-1, Adv-2), and one further version, Inv, used an existential quantifier rather than a type of generic (for instance, 'Some predatory animals have sharp teeth' rather than 'Predatory animals usually have sharp teeth'). The intent was that this sort of information — rather than a generically-stated premise — should make the subjects decide that the conclusion *did not* follow, and that the argument was invalid. Generally speaking, this was to serve as a check that the subjects really were attending to the content of the generic information premise of the problems.

A sufficient number of different contents were constructed so that no subject received the same content with only a different version of the generic premise. For example, if a subject answered the BM1.NK.BP problem with a content that dealt with (made-up) desert plants, then that subject would be given a BM1.NK.Q problem with, say, content dealing with (made-up) ocean current types, and a BM1.NK.M problem having planets as content, and so forth. This was enforced so as to prevent boredom and familiarity from becoming factors in subjects' answers to problems that

differed only in the linguistic form of their generic information. The different contents of problems were balanced so that other subjects received their BM1.NK.BP problem about ocean currents, their BM1.NK.Q problem about planets, and so forth.

Subjects were presented with these "newspaper stories", each on a separate sheet of paper, and were to answer according to the strength that they thought the story supported the conclusion by marking a location on a 1—7 scale. The BM1.NK.M problem using the "conifer" story, the BM2.AK.Q problem using the "medieval musical instruments" story, and the BM1.AK.Inv (invalid) problem using "pillows" are given in the Appendix.

Some of the problems given to subjects had the "AI approved" answer on the left side (low numbers) of the answer scale, while others had it on the right side (high numbers) of the answer scale. In the results reported, all scores have been normalized so that the "AI approved" answers are high. Intuitively, then, a score above 4 (after normalization) are the AI-approved answers; scores below 4 say that the object-in-question follows the exception object; and scores of about 4 say that one can't choose between them.

Had this been carried out correctly, all subjects would therefore have been given 36 problems to solve, with their order randomized between subjects. But due to an error in the collocation of the answer sheets, not all subjects received every condition. Subjects took between 30 and 55 minutes to answer the set of problems, averaging about 45 minutes.

Subjects in Pilot Study

There were 72 subjects in the pilot study, 41 female and 31 male, all native speakers of English. Their average age was slightly higher than 22 years. They were paid $10 for participating, and drawn from the population of Simon Fraser University. Most were students (both graduate and undergraduate), but 10 of them were visitors to campus during a summer break.

Results of Pilot Study

Although the design of the experiment, as well as the error in collocation of some answer sheets and a typo in one group of problems, didn't allow for detailed analysis of the data, there are nonetheless some very interesting descriptive statistics that can be stated. First, TABLE 1 shows scores from the invalid versions of the six BM × Kind problems contrasted with the averages of the other five versions of the same problem (the versions with some form of generic premise).[1] All these differences are significant (ignoring the issue with BM3.AK). Thus, subjects did, in fact, attend to the generic aspect of the premises in these sorts of problems, and did recognize that they were different from the similar premise with an existential quantifier in it. I take this to show the general viability of this direction of research.

[1] An unfortunate but unnoticed typo in the BM3.AK.Inv problem made it not be appropriate as a contrast problem.

type:	BM1.AK	BM1.NK	BM2.AK	BM2.NK	BM3.AK	BM3.NK
Invalid	2.94	3.80	4.92	3.82	*	4.21
Generics	5.75	5.79	5.72	5.36	5.17	5.23

Table 1: Benchmark Problems: Invalid arguments vs. all valid generics averaged

TABLE 2 gives overall ranking of the three different Benchmark problems, averaged across all the different Kind and Linguistic Types (except the Invalids, of course). Note that each is significantly higher than the 4.0 "can't tell" midpoint score, and so each are even further from the "follows the exception object" answer. Therefore, people do in fact follow the AI-approved answer to Benchmark questions, no matter what sort of generic formulation is used in the premise.[2]

type:	BM1	BM2	BM3	total
Average:	5.77	5.73	5.27	5.59
n:	463	454	448	1365

Table 2: Average "normalized" scores on Benchmark Problems, number of answers

TABLE 2 also shows that the overall scores for BM1 and BM2 are not significantly different from one another, but they are both significantly different from BM3. Recall that BM3 has the form: A is an F; F's are usually G; F's are usually H; A is not an H; Is A a G? In related previous research (Elio and Pelletier, 1993, 1996 summarized in Pelletier and Elio, 2002, 2005), we used a similar methodology with the purpose of investigating whether subjects treated the three Benchmark Problems the same. In those studies, unlike the present ones, we only used the Adv-1 (*usually*) formulation; and we used nonce generics such as *books required for my course* and the like, rather than the made-up but "real" generics that we use in the present studies. Those studies also found that BM1 and BM2 were treated the same by subjects, but that BM3 was different. The AI-approved answer is 'yes, A is a G', but we found that subjects were significantly less likely to say this than in BM1 and BM2. Various further manipulations (Elio and Pelletier, 1996) concerning the degree of similarity of the "extra property" mentioned in Benchmark #3 to a presumed reason the exception object violated the default rule led us to a notion of "explanation-based default reasoning". In turn this led us to speculate in our later summaries that an increased number of violated generic statements (adding "F's are usually J; A is not a J" and so on), and would make subjects were less likely to give the AI-approved answer. We called this a second-order default ("The more defaults that are violated, the more likely to violate defaults"), or the rotten egg principle ("Once a rotten egg, always a rotten egg"), which seems to be endorsed by the results of the present Pilot Study.

FIGURE 1 gives the averages across all conditions (except Inv) for Artifactual vs. Natural Kinds. We see that although the AK condition seems to be slightly higher than the NK condition throughout all Benchmark problems, the difference is not significant.

Surprisingly, there was no main effect for Linguistic Type of generic. The averages for the different types across Benchmark × Kind are given in TABLE 3.

[2]The different numbers of subjects in the different conditions illustrates the errors in collocation of the answer sheets and the problem with one set of answers. Nonetheless, it shows that there was a large number of responses for each condition.

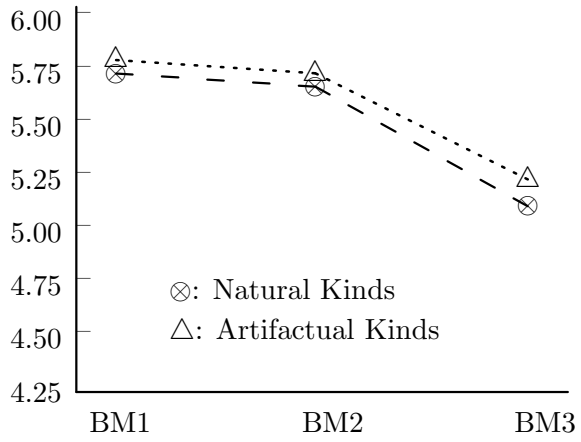


Figure 1: Benchmark × Kind Interaction

type:	Q	M	Adv-2	BP	Adv-1
mean:	5.41	5.42	5.45	5.64	5.65

Table 3: Linguistic Types, across all conditions

There were between 260 and 300 answers for each of the types in TABLE 3, and the differences were close to being significant. Using the Tukey HSD ("Honestly Significant Difference"), a difference of 0.25 would justify significance at the .05 level. Given the missing data in the cells, a between-subjects comparison had to be carried out here, and it seems plausible to think that a within-subjects set of comparisons could more cleanly give appropriate data. One can note that the group {Q,M,Adv-2} versus the group {BP,Adv-1} is very close to significantly different. (However, from the point of view of an intuitive difference, one might find it strange that bare plural generics and *usually*-generics should be so close).

Given these suggestive results, a more carefully controlled experiment was developed.

Main Experiment

Materials in Main Experiment

The materials used in this experiment were a subset of those used in the pilot study, with some alterations to be discussed below. So, they were again the Benchmark problems presented as "paragraphs taken from newspaper or magazine stories" that we saw in the pilot study. The major design change was to eliminate two of the Linguistic Mode forms: the modifier (M) and the second Adverb (Adv-2) forms were deleted so as to leave only four syntactically different formats. One of these four was the "invalid" (Inv) form, so we tested only three different ways to say generics: the bare plural form (BP), the quantificational form (Q), and an adverbial form (Adv). (The adverbial form in this study was *usually*). The experiment now has a $3 \times 2 \times 4$ design, so there are 24 different conditions being tested, and all subjects were tested in all conditions. As before, no subject had

the same content in any two test conditions. Another change saw the Inv problems more closely imitate the generic information premise of the regular benchmarks, rather than being a specially constructed problem. Also, some of the exact wording for some problems was changed, as a result of comments obtained from participants in the pilot study, and the lengths of some of the problems were altered, so that they were all closer in length to each other.

Other changes involved using a randomized block design. As before, it is impossible to measure whether the different contents that a subject gets (when answering different levels of the Linguistic Mode for the same Benchmark $\times$ Kind problem) produces an effect. But the effects of different factors we measure here are based on within-subjects measurements.

Other features of the experiment, such as having some questions with a 1-answer meaning "the AI-approved answer" and others with a 7-answer having that meaning (and then normalizing so that 7 always means the AI-answer when we report the data), are the same in the two experiments. This being a shorter task, subjects almost always were done within a half-hour.

Subjects in Main Experiment

There were 108 subjects in the main experiment, 65 female and 43 male, all native English speakers. One subject (a female) had to be eliminated, due to missing data on the answer sheets. Their average age was just over 21.5 years. Again, as in the pilot study, they were paid $10 for participating, and were drawn from the population of Simon Fraser University. Unlike the pilot study, all these were undergraduate students. None of these subjects had participated in the pilot study.

Results of Main Experiment

First we show that, as in the pilot experiment, subjects reliably distinguish the formulations with generic information—whether they be phrased "Birds fly", "Most birds fly", or "Birds usually fly"— from the invalid ones, existential ones like "Some birds fly." And they can reliably do this even with novel types of entities, not just birds, and in complex scenarios. Figure 2 shows the means for the four Linguistic Modes (including the Inv mode). It can be seen that the Inv mode is significantly lower than the three versions of generics.[3]

However, the presence of an Inv premise affects the Benchmark Problems in a different way than do the generic premises, as Figure 3 shows. When averaged over all the various generics, BM1 and BM2 are not significantly different from one another, but they are significantly different from BM3. But in the Inv case, the three problems are not significantly different from one another.

Once again, as in the Pilot Study and in the earlier work of Elio and Pelletier (1993, 1996), these data show that Benchmark #3 is indeed treated quite differently from the other two Benchmark Problems when using generic premises, with subjects finding it a considerably worse argument

[3] All claims of of significant and non-significant differences are determined by a 3-way repeated-measures ANOVA, with significance set at $\alpha = .01$. When testing for factors with more than two levels, the Mauchly test of sphericality was run, and when it was not met, the Huynh-Feldt correction was used.

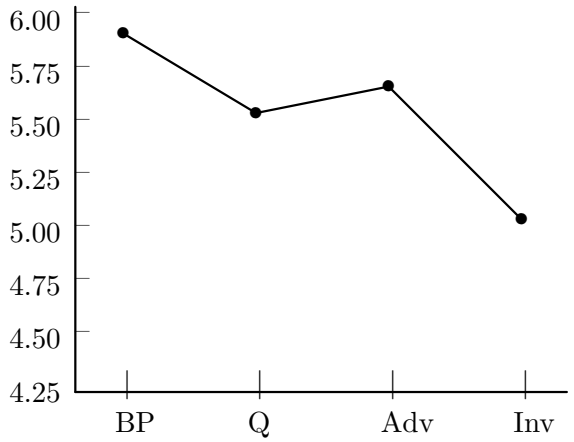


Figure 2: Averages of the Four Linguistic Forms

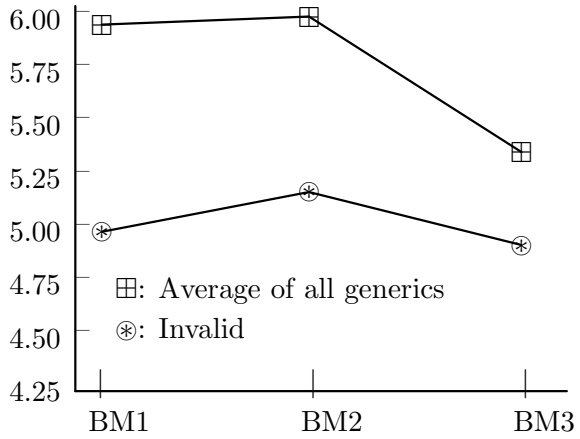


Figure 3: How the Benchmark Problems Treat Generics vs. Inv

when there is an "irrelevant" generic truth that the object-in-question disobeys, even though the conclusion does not mention that violation.

Having therefore shown that the various generic formulations actually tap something different than the invalid, existential formulation of the same premise, we turn our attention to the way natural vs. artifactual kinds affect subjects' reasoning (Figure 4). Confirming the result of the Pilot Study, there is no significant effect for natural vs. artifactual kinds in the three generic types of information. However, with the invalid versions there *is* an effect, with subjects recognizing the invalidity more easily when the kind in question is an artifact. The fact that natural–artifactual distinction affects non-monotonically invalid arguments but not non-monotonically valid arguments is itself a very interesting phenomenon for which I have no clear explanation. Certainly this deserves further study.

Turning our attention now to the way the three different Benchmark Problems deal with the linguistically different generic formulations, we see in Figure 5 that the problems treat the formulations differently. (That is, there is an interaction between Benchmark Problem and Linguistic Form). In

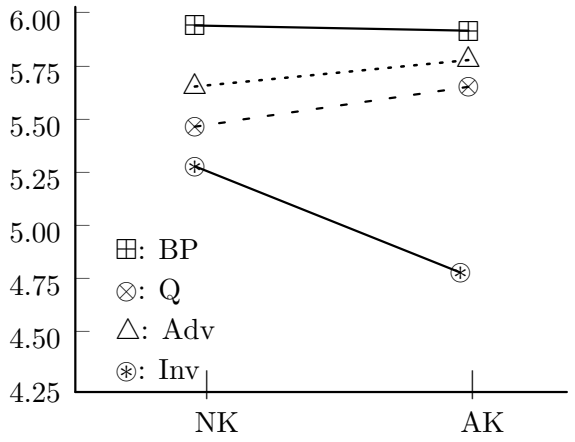


Figure 4: Linguistic Form with different Content-Types

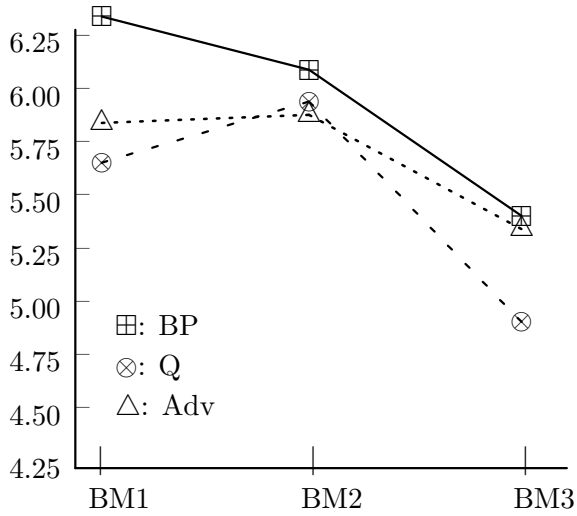


Figure 5: Linguistic Form at different Problems

Benchmark 1, the BP formulation is significantly different from the Q and the Adv formulations, which are not significantly different from each other. In Benchmark 2, none are significantly different from one another. And in Benchmark 3, the Q formulation differs significantly from the others, which do not differ from each other. It seems therefore clear that the three linguistic formulations are tapping different aspects of generic information, and that their differing types of information come to the fore in the different problems.

Finally, let us return to the overall data, which includes the Inv linguistic mode, and consider the magnitude of effect of the various factors that were found to be significant in the Main Study. These magnitudes are estimated by the proportion of variance in the subjects' judgment of validity that is explained by its relationship with one of these factors, after the effects of the blocking factor (subjects) and the interactions of the blocking factor with other factors have been removed. Thus, we estimate the size of the effect of Problem Type (i.e., which Benchmark Problem), after removing the blocking factor and its interactions. Similarly, we can estimate the magnitude of the effect of the Linguistic Mode on the judgment of validity, the magnitude of the effect of the interaction

of Content Kind (natural vs. artifactual) with Linguistic Mode on the judgment of validity, and the the magnitude of the effect of the interaction of Problem Type with Linguistic Mode on the judgment of validity. TABLE 4 gives these results.

Factor:	Magnitude:
Linguistic Mode	55.6%
Problem Type	30.4%
Problem × Mode	06.5%
Problem × Kind	04.9%
(other)	02.6%

Table 4: Estimated Magnitude of Effect

Discussion

It should be re-emphasized that the manipulation investigated in this study actually yields interpretable results. Informal discussions with other researchers have revealed that many of them believe that whatever semantic differences there might be among the various ways of expressing generic statements, they are extremely minor and easily overwhelmed by all sorts of other factors. And it is thought that tests involving the vague notion of default reasoning and asking for judgments of the validity of arguments will simply invoke too coarse of a methodology for testing these minor differences. But the results of the experiments here are robust, both in the Pilot Study and in the Main Study: there is in the first place a clear difference between generically-presented information and existentially-presented information, and the subjects reliably detect this difference. Furthermore, there are detectable significant differences within the various linguistically different generic formulations.

The results, both from the Pilot Study and the Main Study, also demonstrate that the use of a natural kind vs. an artificial kind does not affect subjects' judgments concerning the validity of default argumentation that employs these kinds in generic statements. This might be somewhat surprising to those who have views concerning the way that natural-kind information comes into our consciousness as opposed to the way artificial-kind information comes in. But we also saw that this distinction *does* make a difference in the perception of *invalidity* of default arguments. Now, it is not at all clear why the natural–artifactual distinction should play a role with invalid (existential) premises but not with generic premises, in these sorts of default reasoning. The answer could lie in the idea that people think of artifacts as created for a purpose and their properties as leading to that purpose and therefore having certain regularities, while natural items are simply "presented" to an observer and hence no purpose for their properties is clear. Or, perhaps natural items are assumed to have regularities that obey natural laws, while artifacts are created without any concern for natural regularities. These speculations are suggestive, each in its own way. But such explanations require more clear rationales and a closer tie to theories of learning (involving how kinds are learned), and to how default reasoning proceeds from these learned kinds, before the present results can be said to support one or the other of the speculations.

It is also heartening (for me) to note that both the Pilot Study and the Main Study supported our

earlier finding (Elio and Pelletier, 1993, 1996) that Benchmark Problem #3 is significantly different from the others. A look at Figure 3 shows that the average of the generics on Benchmark #3 is approaching the level of the responses to the Invalid versions of the Benchmark Problems. So, not only do subjects treat Benchmark #3 differently from the other two, but in fact they seem to view it as (almost) invalid. The earlier studies did not conclude this, but rather claimed that subjects saw the increase in the number of violated generic traits as being more and more telling against the AI-approved answer. But since there were no results for invalid default problems, we could not make the claim that subjects actually found these problems to be invalid. But we can now see that they (almost) do; and we might speculate that when yet further violated generic traits (which are nonetheless supposed to be irrelevant to the argument) are also mentioned, subjects will find the arguments as invalid as the ones that they acknowledge as invalid, by applying the second-order default principle of the "rotten egg". This second-order default rule seems to be a clear contender for being added to all normative theories of default reasoning—as well as being a psychologically-successful predictor of human judgments.

A central goal of the Main Study was to investigate whether there are any subtle differences in meaning among three ways of expressing generics. This would form part of the wider study that was started in the Pilot Study that explored the five different ways mentioned in (5), which in turn is a part of a still wider study that in addition looked at such ways to express generics as those mentioned in (3). We did discover that there are differences among our three ways, although it remains far from clear what are the underlying semantic features that could possibly give rise to these differences. We discovered that the Bare Plural formulation contained information that is different from that contained in the generics using a *most* quantifier or the *usually* adverb, but that this diference displayed itself only with the simplest of the Benchmark Problems—the one that asserts the mere existence of some other exception object. So, we discovered that

> [BP] Birds fly
> Tweety and Polly are birds
> Polly doesn't fly
> Therefore, Tweety flies

is judged "more valid", or "as valid more often", then when the (BP) premise is replaced by either (Q) "Most birds fly" or (Adv) "Birds usually fly". But in the rather similar Benchmark Problem 2, where the only obvious difference is in the presence of an "irrelevant" property of the object-in-question,

> [BP] Birds fly
> Tweety and Polly are birds
> Polly doesn't fly
> Tweety is yellow
> Therefore, Tweety flies

there is no significant difference in subjects' evaluations of the validity of this argument-type among the (BP), (Q), and (Adv) variants of the generic premise. And in the final problem

[BP] Birds fly
[BP] Birds eat birdseed
Tweety and Polly are birds
Polly doesn't fly
Tweety doesn't eat birdseed
Therefore, Tweety flies

subjects think of the (BP) version and the (Adv) version (where both the BP premises are replaced by their *usually* variants) as being the same, and these both as "more valid" than the (Q) version (where both the BP premises are replaced by their *most* variants). Just why this should happen with these sorts of problems (which seem so similar to each other) is a question whose answer lies deep in a theory of information, and cannot be answered here. But it is clear that there is a difference among the three different generic formulations.

Conclusion

So: are all generics created equal? The data presented here say no—the particular linguistic form in which they are couched will determine subtle differences that show up when the information being put forward by the generic statements are employed in other tasks, such as everyday, default reasoning about the commonsense knowledge that is given by means of these statements. So, the presumption by both the holders of the inductivist view of generics and the rules-and-regularities view of generics seem to be shown wrong: there is no one interpretation that is correct for all generics.

But: is the proposed variant of Carlson's thesis thereby affirmed—that the lexically different formulations of generic statements correlates with the the two different sources for information that Carlson identified as forming the basis for different theories of generics? This seems not so clear, since the differences that show up in interpretation of the different linguistic expressions of genericity do not seem to straightforwardly map onto Carlson's inductivist vs. rules-and-regularities distinction. And in any case, these results identify *three* different patterns that the generic statements obey, not just the two patterns that the Carlsonian hypothesis suggests. So, before such an identification can be made firmly, a better and deeper theory of the sort of information that is conveyed by generics is needed, and a better and deeper theory of the way information is employed in default reasoning is required. Such theories should set an agenda for researchers involved with the formal (and informal) semantics of linguistic generics and generalizations about the world, and an agenda for researchers involved in the ways people employ this sort of knowledge in everyday reasoning.

Acknowledgments

Thanks to student research assistants Cam Clark and Masha Tkatchouk, and to both (Canadian) NSERC grant 5525 and (SFU) President's Research Grant for support to them and this research.

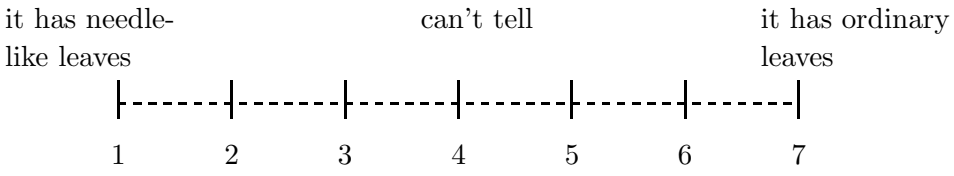
Also, I offer sincere thanks to the many comments from the conference participants that I received at the 2006 Conference, where I presented the results of the pilot study. And thanks to Alasdair Urquhart for technical assistance.

Appendix

The BM1.NK.M problem using the conifer story:

A conifer is a type of tree or shrub that bears reproductive structures called cones. Among the different conifers are the pines and cedars that grow so commonly in the Vancouver area, as well as throughout the world. But not all conifers are so widespread. In northwestern Saskatchewan there is a conifer commonly called pigwaisa that is found nowhere else, and in the Sonoran desert of northern Mexico is another conifer that is called sumunaro, which is found only in a few other locations around the world. The typical conifer has needle-like structures rather than ordinary leaves, although the sumunaro is an exception to this because it does have medium-sized, ordinary leaves.

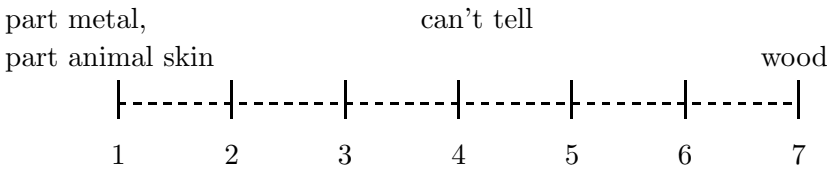
According to the information in this paragraph, what kind of leaves do you conclude that pigwaisa has?



The BM2.AK.Q problem using the medieval musical instrument story:

The middle ages saw a great proliferation of musical instruments, especially stringed instruments. Included in these new stringed instruments were the ganbaz and the tanfir. Most of the new stringed instruments were made of wood, although the gambaz was an exception, being partially metallic and partially animal skin. The tanfir had doubled strings and frets. Both instruments enjoyed wide popularity for more than a century.

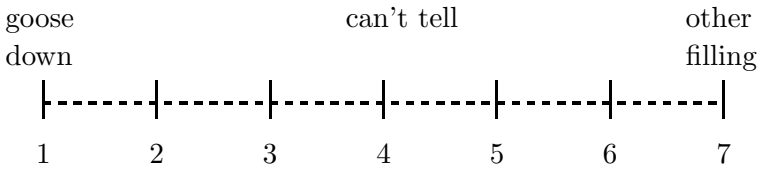
What do you think the tanfir was made of, given the evidence presented in this paragraph?



The BM1.AK.Inv (invalid) problem using the pillows story:

Over the years many different styles of pillows have been developed. Not only has the filling been changed from the traditional feather or down, but also there have been many new shapes for pillows. Two new shapes have been the certival support and the lumnaxious pillows. The new shape pillows are often triangular. Some of the certival support pillows have been filled with the traditional goose down.

What do you think the lumnaxious pillows are filled with, according to the evidence presented in this article?



Notes

References

Carlson, G. (1980). *Reference to Kinds in English*. NYC: Garland Press. Originally a University of Massachusetts PhD dissertation, 1977.

Carlson, G. (1982). Generic terms and generic sentences. *Journal of Philosophical Logic 11*, 145–181.

Carlson, G. (1995). Truth-conditions of generic sentences: Two contrasting views. In G. Carlson and F. J. Pelletier (Eds.), *The Generic Book*, pp. 224–237. Chicago: University of Chicago Press.

Carlson, G. and F. J. Pelletier (2002). The average American has 2.3 children. *Journal of Semantics 19*, 73–104.

Elio, R. and F. J. Pelletier (1993). Human benchmarks on ai's benchmark problems. In *Proceedings of the 15th Congress of the Cognitive Science Society*, Hillsdale, NJ, pp. 406–411. Laurence Erlbaum.

Elio, R. and F. J. Pelletier (1996). On reasoning with default rules and exceptions. In *Proceedings of the 18th Congress of the Cognitive Science Society*, Hillsdale, NJ, pp. 131–136. Laurence Erlbaum.

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

Lakoff, G. (1972). Linguistics and natural logic. In D. Davidson and G. Harman (Eds.), *Semantics of Natural Language*, pp. 232–296. Dordrecht: D. Reidel.

Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic 2*, 458–508.

Lawler, J. (1973). *Studies in English Generics*. Ph. D. thesis, University of Michigan.

Lifschitz, V. (1989). 25 benchmark problems in nonmonotonic reasoning, v. 2.0. In M. Reinfrank, J. de Kleer, and M. Ginsberg (Eds.), *Nonmonotonic Reasoning*, pp. 202–219. Berlin: Springer Verlag.

Pelletier, F. J. and R. Elio (2002). Logic and cognition. In P. Gärdenfors, J. Wolenski, and K. Kijainia-Placet (Eds.), *In the Scope of Logic, Methodology, and Philosophy of Science, Vol. 1*, Dordrecht, pp. 137–156. Kluwer.

Pelletier, F. J. and R. Elio (2005). The case for psychologism in default and inheritance reasoning. *Synthèse 146*, 7–35.

Pelletier, F. J. and L. Schubert (1989/2003). Mass expressions. In F. Guenthner and D. Gabbay (Eds.), *Handbook of Philosophical Logic, 2nd Ed.. Vol. 10*, pp. 249–336. Dordrecht: Kluwer. Updated version of the 1989 version in the 1st edition of *Handbook of Philosophical Logic*.