The Semantics–Pragmatics Interface: An Empirical Investigation

Igor Douven
Sciences, Normes, Décision (CNRS)
Paris-Sorbonne University
igor.douven@paris-sorbonne.fr

Karolina Krzyżanowska
Munich Center for Mathematical Philosophy
Ludwig-Maximilians University
k.krzyzanowska@lmu.de

Abstract

Linguists and philosophers commonly distinguish between semantics and pragmatics, where the former concerns the truth or falsity of linguistic items and the latter concerns aspects of the use of such items that may make them unassertable even when true. Common though the distinction is, there is an ongoing controversy about where exactly the line between semantics and pragmatics is to drawn. We report two experiments meant to investigate empirically whether there is any pre-theoretic distinction that might help settle the debate. The same experiments are meant to shed light on a related question, namely, whether pragmatic aspects of language use pertain only at the level of assertability and not at that of believability. Our results suggest that ordinary people do not reliably distinguish among truth, assertability, or believability. We argue that this has consequences for the methodology of experimental semantics and pragmatics.

1 Introduction

Compare these sentences:

(1) a. Mike was thirsty, but the beer was warm.

b. The beer was warm, but Mike was thirsty.

According to most linguists and philosophers, these sentences have the same truth conditions—both are true if and only if Mike was thirsty and the beer was warm—but they make contradictory suggestions or, to use the technical term, they generate contradictory *implicatures*. While (1a) suggests that Mike did not drink the beer, (1b) suggests that he did. Therefore, at most one of (1a) and (1b) will be assertable. For instance, if Mike was thirsty, the beer was warm, and Mike drank the beer, then by asserting (1a) one would mislead one's audience—which cooperative speakers will want to avoid.

Implicatures and related phenomena that may render a sentence unassertable even if it is true are the objects of study of the field of *pragmatics*, while *semantics* is concerned

with the truth and falsity—rather than the assertability—of sentences. It is uncontested that semantics and pragmatics occupy themselves with different, although related, aspects of language and language use. However, where the line between the two is to be drawn is a matter of ongoing controversy. For example, it is not universally accepted that the contribution "but" makes to (1a) and (1b)—the suggestion of a contrast between the conjuncts, as most theorists would say—is to be thought of as affecting these sentences' assertability conditions but not their truth conditions. Many more, and partly more subtle, issues have been raised in the debate about the location of the semantics–pragmatics divide (see Levinson 2000, Ch. 3, for an overview).

That no agreement on this question is in sight, despite years of intense debate, raises the question to what extent the semantics–pragmatics distinction has a pre-theoretic basis. Do ordinary people reliably distinguish between truth and assertability, or are we confronting a purely theoretical issue?

We present two experiments meant to address the question of the semantics–pragmatics divide empirically. In the experiments, participants were asked to assess various sentences that, according to mainstream semantics, qualify as true but, according to mainstream pragmatics, carry a false implicature. Some participants were asked to judge the truth values of these sentences, whereas others were asked to judge their assertability. Given the just-described nature of the materials, and supposing the semantics–pragmatics divide to be rooted in ordinary linguistic practice, we should expect the judgments of the two groups of participants to differ markedly.

Besides truth and assertability judgments, we were also interested in judgments of believability, which we elicited from a third group of participants in both experiments. It is a common view among philosophers (even if not among linguists) that implicatures are relevant to the issue of assertability but not to that of acceptability or believability (e.g., Edgington 1986). But, first, if ordinary people do not distinguish systematically between truth and assertability, then, given that truth matters to believability, so might assertability. Second, quite independently of how truth and assertability are connected in actual practice, Douven (2010) raised the possibility that many or even most factors that may make a sentence unassertable may also make it unbelievable, the idea being that holding as a belief a sentence carrying a false implicature may mislead one's future self who may retrieve that belief from retentive memory, much in the way in which asserting the sentence may mislead one's audience.[1] If this is correct, we should not find significant differences between the responses from the groups judging the assertability of our items and the responses from the groups judging the believability of those items.

1.1 Theoretical background

Grice (1989a) was the first to systematically argue that, because participants to a conversation assume each other to be clear and provide the amount of information commensurate to the purpose of the conversation, by their utterances they can convey more than just the information contained in the truth-conditional content of those utterances. *That* we assert a specific sentence in a specific conversational setting can itself be a source of information, over and above the information semantically encoded in *what* we assert. Thus, we infer from a speaker's assertion of

(2) Some of Harriet's children are blond.

[1] For applications of this idea, see Douven (2008), (2016, Ch. 4), and Capone (2011), (2016).

that, at least as far as the speaker knows, not all of Harriet's children are blond; if they were, and if the speaker were aware of that, she could, with no additional effort, have been more informative by asserting:

(3) All of Harriet's children are blond.

That not all of Harriet's children are blond is said to be an implicature of (2). Similarly,

(4) Jim has four children.

is generally taken to implicate that Jim has *exactly* four children. If Jim had five children, (4) would still be true—someone who has five children also has four children—but the sentence would be underinformative: with just as much effort, someone asserting (4) could have provided more information.

A large part of Grice's own work as well as that of his followers concerns the typology of implicatures. The most general distinction Grice made is that between *conversational* and *conventional* implicatures. The former are those that, as in the case of (2) and (4), derive from what a speaker says in a given context in conjunction with the presumption that the speaker aims to be cooperative; the latter are related to the conventional meanings of words, as in the case of "but" encountered earlier, which by convention implicates the presence of a contrast.

It is nowadays common to further distinguish among conversational implicatures on the basis of the principles—maxims, in Grice's terminology—that underly their production.[2] For our present concerns, the important categories of conversational implicatures to be distinguished are the *scalar* implicatures and the *order* implicatures. The former exploit Grice's Maxim of Quantity, which implores speakers to make their contribution to a conversation as informative as is required, given the goal of the conversation. The latter exploit the Maxim of Order, which Grice actually presents as falling under the Maxim of Manner; according to the Maxim of Order, the speaker should be orderly, in particular, relate events in the order in which they occurred.

Scalar implicatures involve an expression that can be naturally put on a scale together with other expressions the speaker could have used but did not use, and they arise because the Maxim of Quantity gives the hearer grounds to presume that the speaker has gone as far out on the scale as his or her knowledge warrants, and as the purpose of the conversation requires. Scalar implicatures have given rise to a typology of their own, on the basis of the various scales that may be involved. Doran et al. (2009) distinguish between the following types:

Quantificational items: These involve a scale of quantifiers, such as ⟨some, many, most, all⟩ or ⟨sometimes, often, always⟩ or ⟨possibly, probably, certainly⟩; the "not all" implicature normally generated by (2) is of this kind.

Gradable adjectives: These involve a scale of adjectives admitting of degrees, such as ⟨small, middle-sized, big, gigantic⟩ or ⟨soft, audible, loud, blaring⟩ or ⟨somewhat sweet, sweet, very sweet⟩; for instance, an implicature of this kind is generated by asserting that Bill Gates is relatively rich (which implicates that he is not extremely rich).

[2] Grice (1989a, p. 37 ff) also distinguishes between *generalised* conversational implicatures and *particularised* conversational implicatures, where the former are supposed to be generated by default while the latter require special contextual assumptions. The present paper is concerned only with generalised conversational implicatures.

Ranked orderings: These involve orderings like ⟨beginner, intermediate, advanced⟩ or ⟨teenager, adult, senior⟩ or ⟨income under €50,000, income between €50,000 and €100,000, income between €100,000 and €200,000, income over €200,000⟩; for instance, asserting that people who earn more than €200,000 have to pay taxes implicates that people with a lower income are exempt from paying taxes.

Cardinal numbers: These involve some cardinal number scale; the "exactly four" implicature normally generated by an assertion of (4) is an instance of this type.

In our summary of Gricean pragmatics, we have been following the practice of many textbooks in pretending that there is always a clear-cut distinction between *what is said*, which is supposed to be the focus of semantics, and *what is implicated*, which is supposed to be the focus of pragmatics. That is an oversimplification, however.

We already noted that it is not completely obvious that the contrast that a use of "but" typically suggests *must* be thought of as being part of the word's pragmatic, rather than its semantic, meaning. Similarly, most theorists agree that the conditional-forming operator "if" suggests the existence of a connection between the parts which it connects (the antecedent and consequent), but while some have argued that this suggestion is a matter of conversational or conventional implicature, others hold that it flows from the semantics of "if" (e.g., Braine 1978; Braine and O'Brien 1991; Kratzer 1986; Krżyzanowska, Wenmackers, and Douven 2014; Douven 2016, Ch. 2). And some authors oppose the view that sentences like (4) carry the "exactly n" ("exactly four," in the given case) reading as a matter of implicature, advocating instead that the "exactly" reading is part of the semantics of numerals; see, for instance, Scharten (1997) and Breheny (2008).

More generally, Levinson (2000, p. 195) gives a schematic representation of the broad range of positions on the semantics–pragmatics interface to be found in the philosophical and linguistic literature. As the schema shows, there is disagreement about the divide between what is part of a sentence's semantic contribution and what is part of its pragmatic contribution between basically any pair of the most influential authors writing on the matter in the last decades of the previous century. And we add that the more than one and a half decades that have passed since Levinson's book appeared have failed to bring any convergence on the issue. All this illustrates Levinson's (2000, p. 165) claim that "[the Gricean] program . . . renders problematic and 'up for grabs' the correct division of labor between semantics and pragmatics in the explanation of many aspects of meaning."

What might explain the conspicuous lack of consensus on the location of the semantics–pragmatics interface? When one reflects on the previously mentioned questions of what "but," "if," numerals, and so on, contribute semantically, and what they contribute pragmatically, it seems that introspection gives little guidance on how to answer them. This raises the suspicion that the distinction is ultimately a theoretical one without a real grounding in how ordinary people think about language. We, as ordinary speakers, might simply not distinguish between truth and assertability in any systematic way, other than perhaps for nonlinguistic reasons (such as reasons of politeness or diplomacy). That would swiftly explain why we, as theorists, have no firm intuitions to rely upon in attempting to delineate the semantic from the pragmatic.

1.2 Hypotheses

Thus, our first hypothesis is that there is no folk distinction between truth and assertability, at least not one that is systematic enough for theorists to safely build upon. If borne out by the data, this would be of some importance. After all, it would mean that there is no hope for theorists to resolve issues concerning the semantics–pragmatics interface by

tapping into ordinary speakers' intuitions about where truth and assertability come apart. It is to be noted that this hypothesis has a largely *exploratory* character: there is nothing in the literature to suggest that asking people to judge the *truth* of sentences generating a false implicature will not lead to results significantly different from the results obtained by asking people to judge the *assertability* of those same sentences. Rather, the hypothesis is motivated by observation of the continuing lack of agreements among theorists about the location of the semantics–pragmatics interface, as well as about our own lack of introspective clarity on the same issue.

Our second hypothesis concerns the distinction between believability and assertability rather than that between truth and assertability. It is exactly parallel to the first hypothesis, stating that the folk do not systematically distinguish between believability and assertability. This second hypothesis is partly inspired by the same considerations that led us to postulate the first: if in the minds of ordinary people truth and assertability largely coincide, then there would not seem to be much room for truth to affect believability—as on any standard epistemology it does—in a way that does not entail assertability affecting believability as well. But, as previously mentioned, the second hypothesis also follows from a theory to be found in the literature, namely, the pragmatics of belief account as proposed in Douven (2010), and for all that follows from that account, there is a clear dividing line between truth and assertability.

There are two subsidiary questions that we are interested in. First, we will also look at differences between types of implicature. Previous experimental work on implicatures (e.g., Doran et al. 2009) found truth judgments for sentences carrying false implicatures to differ significantly depending on the type of implicature involved. We investigate this matter not only for truth judgments but also for believability and assertability judgments.

Second, we are interested in individual differences among participants. Spychalska, Kontinen, and Werning (2016) report the results of an EEG study investigating whether people tend to consider the "not all" implicature that according to standard pragmatic theorising is thought to be generated by default by the existential quantifier "some" as really being part of the *semantical* meaning of that quantifier. They found that, as far as their participants' truth value judgments went, the group of participants could be almost evenly split into "logicians" and "pragmatists" if the former were defined as responding in at least 70 percent of the cases with "true" to an existentially quantified sentence with false "not all" implicature and the latter as responding in at least 70 percent of those cases with "false." This at least hints at the possibility that we find a persistent disagreement about the location of the semantics–pragmatics divide because we are, possibly by nature, split up into two groups responding in opposite ways to questions concerning what is said and what is implicated by sentences, or at any rate by particular sentences. In that case, the two groups might each systematically distinguish between truth and assertability, but a failure to recognise that there are actually two groups might have made it seem as though there were no systematic intuitive view on the semantics–pragmatics interface. As mentioned, however, Spychalska et al.'s material was restricted to existentially quantified sentences. It remains to be seen whether their finding generalises once other types of implicature are taken into consideration.

2 Experiment 1

This experiment was designed to test both of our hypotheses. Participants were divided into three groups, which were asked to judge various sentences generating false implicatures in

terms of truth, believability, and assertability, respectively. The hypotheses were evaluated by comparing the responses in the three conditions.

2.1 Method

Participants

There were 349 participants in the experiment. They were recruited via CrowdFlower (`http://www.crowdflower.com`), which directed them to the Qualtrics platform (`http://www.qualtrics.com`) on which the experiment was run. The participants were paid a modest fee in return for their time and effort. Repeat participation was prevented.

All participants were from Australia, Canada, the United Kingdom, or the United States. We excluded data from the 5 percent slowest and 5 percent fastest participants, then from non-native speakers of English, participants who were colour blind (given that some of our materials involved colour stimuli), and participants who answered negatively to the question of whether they had responded seriously to the questions in the experiment (it was explicitly stated that their answer to this question would not affect payment; this followed a suggestion from Aust et al. 2014).

This left us with 290 participants whose responses were used for the final analysis. These participants spent on average 9.36 minutes on the survey (SD = 3.12 m). Their mean age was 36 (SD = 12). Of these participants, 184 were females; 198 indicated university as their highest education level, 85 high school, and 7 a lower education level.

Design

We used a 3×24 mixed design with three levels of type of question (true / believable / assertable). Participants were randomly assigned to one of three groups, each corresponding to one of the levels of the type of question variable. All participants judged 24 test items together with 43 filler items.

Materials and procedure

The materials for the experiment were all in English, the participants' native language. The 24 test items are shown in Table 1. They were so chosen that, according to any current semantic theory, they qualified as true but, according to standard pragmatic thinking, also generated a false implicature. In particular, four test items are generally taken to generate a false quantifier scalar implicature (items 1–4), four a false gradable adjective scalar implicature (items 5–8), four a false ranked ordering scalar implicature (items 9–12), four a false cardinal number scalar implicature (items 13–16), four a false order implicature (items 17–20), and four a false conventional implicature (items 21–24). The falsity of the implicatures was taken to arise either from their inconsistency with readily available world knowledge or from the visual context the experiment provided, for instance, by presenting "Some patches are blue" to participants while stating that the sentence is about the colour patches shown on the same screen, where these patches are then *all* blue (see the notes of Table 1).

The participants in all three groups were asked a two-alternatives forced choice question after each item, but the type of question differed among the groups. Participants in the first group (N = 95) were asked whether they deemed the item true or false; participants in the second group (N = 106) were asked whether they deemed the item believable or unbelievable; and participants in the third group (N = 89) were asked whether they deemed the item assertable or unassertable. Five items concerned a visual stimulus, which either consisted of a series of colour patches or of a short comic strip. These items were,

Table 1: Items used in Experiments 1 and 2.

1. Some patches are blue.[a]
2. Some roses are flowers.
3. Most patches are red.[b]
4. Most laptops are computers.
5. The tiger finds the boy's cereal moderately sweet.[c]
6. The female basketball player Margo Dydek (7 ft 2 in / 2.18 m) was tall for a woman.
7. Bill Gates is relatively rich.
8. On the North Pole, winter temperatures are somewhat cold.
9. In the UK, people over the age of 85 have the right to retire.
10. In principle, all American citizens over the age of 25 have the right to vote in federal elections.
11. In the UK and the US, children under the age of 15 are prohibited from buying hard drugs.
12. In the US, people who earn more than $200,000 a year are obliged to pay taxes.
13. Alfred Hitchcock made two movies.
14. President Obama has one daughter.
15. In the last Olympic games, the US won four medals.
16. At the height of its power, Great Britain owned 12 ships.
17. The tiger looks for the bread in the toaster and the boy puts a piece of bread into the toaster.[d]
18. Princess Diana died in a car accident and she divorced Prince Charles.
19. The man comes up with a bogus answer and the boy asks how the load limit on bridges is determined.[e]
20. Kate Middleton gave birth to a son and she married Prince William.
21. Although Prince William had fallen in love with Kate Middleton, the 2014 Winter Olympics will be in Russia.
22. *Harry Potter and the Sorcerer's Stone* was a box office hit, therefore Obama is the president of the US.
23. Although Obama won a second term as president, dolphins are mammals.
24. Mitt Romney lost the 2012 presidential election, therefore U2 is a rock band.

[a] Shown with a series of only blue patches. [b] Shown with a series of only red patches. [c] Shown with a comic strip in which a tiger is seen finding a boy's cereal extremely sweet. [d] Shown with a comic strip in which a boy first puts bread in a toaster and then a tiger looks into the toaster. [e] Shown with a comic strip in which a boy first asks the question and then the man answers it.

together with the associated visual stimulus, presented on a separate screen. The same was true for the 5 filler items that concerned a visual stimulus (also either a series of colour patches or a comic strip). The 19 remaining (test) items were each shown on a screen together with two filler items. The order in which the one test item and two filler items appeared on the screen was randomised per participant. The same was true for the order in which the screens appeared.

2.2 Results

Figure 1 shows the proportion of positive responses for each item, split by group. As is suggested by the figure, proportions of positive responses for any condition of type

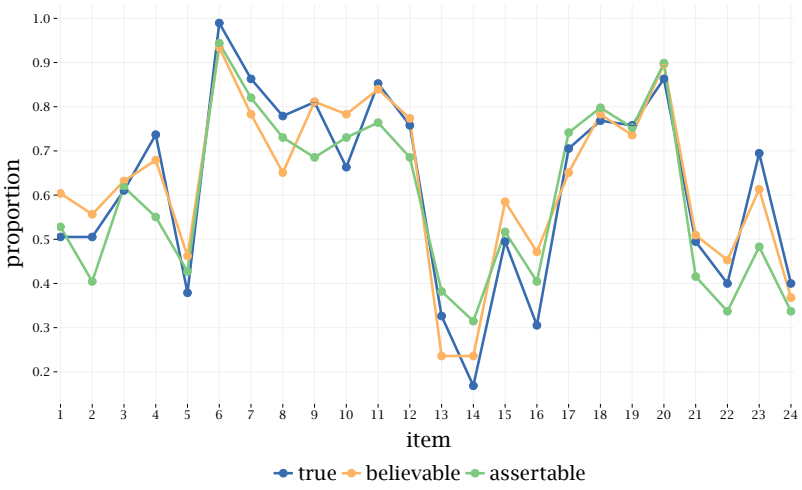

Figure 1: Proportions of positive responses for the 24 test items. Item numbers correspond to the numbering in Table 1.

of question were highly correlated with the proportions of responses for either of the remaining conditions: all rs $> .91$, all associated ps $< .0001$.

To investigate whether there was a main effect of type of question, we fitted two binomial generalised linear mixed models with logistic link functions, as recommended in Jaeger (2008) for categorical data generally. The models were fit using the `lme4` package (Bates et al. 2015) for the statistical computing language `R` (R Core Team 2015). Both models had participants' responses as independent variable (with the positive response coded as 1 and the negative as 0) and participants and items as crossed random effects (see Baayen, Davidson, and Bates 2008). One model had type of question as fixed effect while the other was an intercept-only model. A likelihood ratio test showed that adding type of question as predictor did not lead to a significant improvement of model fit: $\chi^2(2) = 1.62$, $p = .445$.

Because we were interested in whether type of question might still have a significant effect for specific types of implicature when considered separately, we defined type of implicature as a factor with six levels (quantificational / gradable adjective / ranked ordering / cardinal number / order / conventional) and fitted two further models, one with type of question and type of implicature as fixed effects, and another with the same variables and their interaction as fixed effects; both models had the same random effects structure as the previous two models. Significance of effects was again determined via likelihood ratio tests. The model with type of question and type of implicature as fixed effects fit the data significantly better than the model with only type of question as fixed effect: $\chi^2(5) = 23.73$, $p < .001$. And adding the interaction term led to a further significant improvement of fit: $\chi^2(10) = 34.23$, $p < .001$.

We followed up the finding of a significant interaction effect in the absence of a main effect of type of question by conducting post-hoc comparisons with Tukey's HSD adjusted p-values using the `lsmeans` package (Lenth 2015). These showed that, for all levels of type of implicature, the responses of none of the groups differed significantly from those of either of the other two groups. This was confirmed by building, for each type of implicature

Table 2: Summary statistics from Experiment 1 for types of implicatures.

	true		believable		assertable		all	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
quantificational	0.59	0.11	0.62	0.05	0.53	0.09	0.58	0.07
gradable adjective	0.75	0.26	0.71	0.20	0.73	0.22	0.73	0.23
ranked ordering	0.77	0.08	0.81	0.03	0.72	0.04	0.77	0.04
cardinal number	0.32	0.13	0.38	0.18	0.40	0.08	0.37	0.13
order	0.77	0.07	0.77	0.10	0.80	0.07	0.78	0.08
conventional	0.50	0.14	0.50	0.10	0.39	0.07	0.46	0.10

separately, two models with the same random effects structure as in the models above, where one model was an intercept-only model and the other had type of question as fixed effect, and where the independent variable consisted of the participants' responses to the items belonging to the given type of implicature only. For no type of implicature did the larger model fit the data significantly better than the intercept-only model.

These findings suggest that it basically makes no difference whether people are asked to judge the truth, believability, or assertability of a sentence that is true according to standard semantics but that generates a false implicature, and that this holds across all main types of implicature.

We next turned to an investigation of the question of whether it makes a difference for the truth / believability / assertability judgments to which type the false implicature belongs that a sentence carries, in line with but also extending the research reported in Doran et al. (2009). To that end, we carried out four one-way ANOVAs, one for each condition of type of question (true / believable / assertable) separately and one for the results of all three conditions collapsed. All ANOVAs had participants' responses as outcome variable and the factor type of implicature (with the previously mentioned levels) as predictor variable.

All four ANOVAs revealed a significant effect of type of implicature: $F(5, 2274) = 61.11$, $MSE = 0.21$, $p < .0001$, $\eta^2 = .12$ for true; $F(5, 2538) = 55.47$, $MSE = 0.21$, $p < .0001$, $\eta^2 = .10$ for believable; $F(5, 2130) = 51.50$, $MSE = 0.22$, $p < .0001$, $\eta^2 = .11$ for assertable; and $F(5, 6954) = 161.40$, $MSE = 0.21$, $p < .0001$, $\eta^2 = .10$ for the conditions collapsed. The effect size is in each case in the medium range. Post-hoc comparisons using Tukey's HSD indicated that in the separate conditions as well as in the conditions taken together, all pairs of types of implicature were significantly different from each other at $\alpha = .05$ except for, in the case of truth, each pair of gradable adjective, ranked ordering, and order (all ps $> .99$); in the case of believable, the pair order and gradable adjective ($p = .47$) and the pair order and ranked ordering ($p = .76$); in the case of assertable, each pair of gradable adjective, ranked ordering, and order (all ps $> .18$); and in the collapsed condition, each pair of gradable adjective, ranked ordering, and order (all ps $> .12$). The means and standard deviations for each type of implicature in each condition are shown in Table 2.

Finally, we looked at individual differences among the participants. Recall that Spychalska, Kontinen, and Werning (2016) found that their group of participants could be divided almost exactly into two subgroups on the basis of their truth value judgments, with one group giving mostly "logical" responses—meaning that they answered with "true" to an existentially quantified sentence with a false "not all" implicature—and the other

group giving mostly "pragmatic" responses, meaning that they answered with "false" to those same existentially quantified sentences.

Among our items were two existentially quantified statements with false "not all" implicatures (items 1 and 2 in Table 1), and one easily verifies that, based on the above-mentioned study, we should expect to find no more than 39.9 of our 95 participants in the true condition to have responded differently to the two items, and no less than 27.55 in either group of same responders. Our data are consistent with this: there were 32 participants who deemed both items 1 and 2 false, and 33 who deemed them both true, leaving 30 who deemed one of the items true and the other false.

Importantly, however, this finding does not generalise across types of implicature. If we consider the 95 participants' responses to all items, we find that there were 38 logicians (according to Spychalska et al.'s standards) and 3 pragmatists, leaving a majority of 54 participants who deemed between 30 and 70 percent of the items true, thus qualifying neither as logicians nor as pragmatists.

One might hope to find at least *some* natural division among people on the basis of how they take implicatures to bear on the truth values of sentences generating those implicatures, even if it is not the clean division between logicians and pragmatists proposed by Spychalska and coauthors. That the prospects for this are bleak becomes manifest when we look at the correlations among participants' truth judgments of our 24 items. It turns out that quantificational implicatures, like Spychalska et al. used in their experiment, are rather special, together with conventional implicatures. For as Figure 2 shows, for these types we see at least modest correlations among participants' responses. But the same figure shows that neither the responses to the quantificational items nor the responses to the conventional items correlate even moderately with virtually any of the other items, nor do the responses to those other items tend to correlate even moderately among themselves. (The correlation matrices for the believability and assertability responses reveal an almost identical pattern.)

2.3 Discussion

Given our relatively large number of participants, we should have been able to find a main effect of type of question if it existed. That we did not find one corroborates both our hypotheses. According to the first hypothesis, the semantics–pragmatic divide, while widely assumed to exist by linguists and philosophers, has no basis in ordinary linguistic practice, and according to the second, broadly the same considerations that pertain to the assertability of a sentence also pertain to its believability. The absence of a main effect of type of question corroborates these hypotheses because if people distinguished between either truth and assertability or between assertability and truth, then we should have registered significant differences in the responses of our three groups, given that according to standard theorising the items in Table 1 are all evidently true (and hence, according to mainstream epistemology, believable) but also unassertable.

The differences we found among types of implicatures are consistent with findings reported in Doran et al. (2009). These authors only considered scalar implicatures, and their experiment differed in important respects from ours. Nevertheless, we found that the percentage of "false" designations for cardinal number items was significantly higher than that for quantificational items, which in turn was significantly higher than those for gradable adjectives and ranked orderings, while the percentages of "false" responses for gradable adjectives and ranked orderings did not differ significantly from one another. This is exactly what Doran and coauthors report.

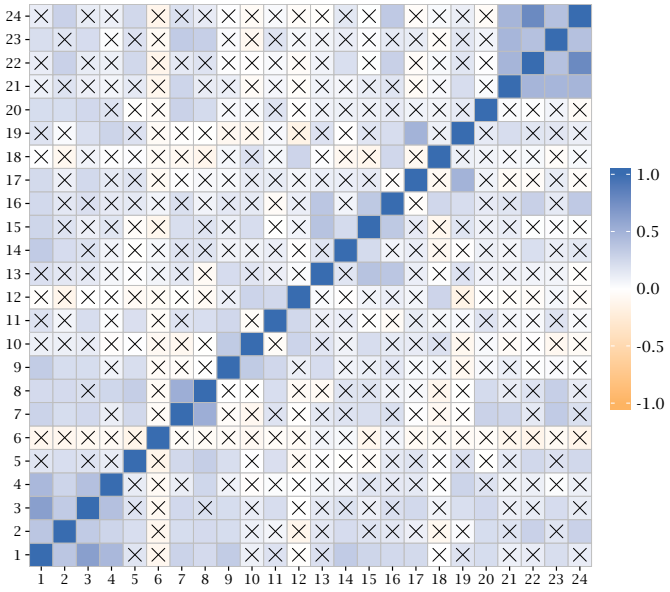

Figure 2: Correlations among participants' truth judgments for the 24 items. (A cross indicates failure to reach significance at $\alpha = .05$.)

Finally, our results concerning individual differences showed that it may prove difficult to make any useful divisions between participants according to how false implicatures affect their appreciation of the sentences generating those implicatures. In any event, we could not replicate Spychalska, Kontinen, and Werning's (2016) result concerning existentially quantified sentences for our more inclusive set of materials.

3 Experiment 2

That Experiment 1 failed to find a main effect of type of question may have been due to the fact that it used 2AFC tasks, which are known to have smaller discriminatory power than tasks offering Likert-scale responses, all else being equal (Preston and Colman 2000). For that reason, we conducted a further experiment which was like Experiment 1 in every respect except for offering Likert-scale response options.

3.1 Method

There were 363 persons participating in the experiment. Participants were recruited and tested in the same way as in Experiment 1. They were also paid the same fee for their participation. Here, too, the participants were from Australia, Canada, the United Kingdom, or the United States. Exclusion criteria were the same as in Experiment 1, which left 298 participants for the final analysis. It took these participants on average 10.02 minutes to complete the survey ($SD = 3.26$ m). Their mean age was 35 ($SD = 12$); 195 of them were females; 218 indicated university as their highest education level, 73 high school, and 7 a lower education level.

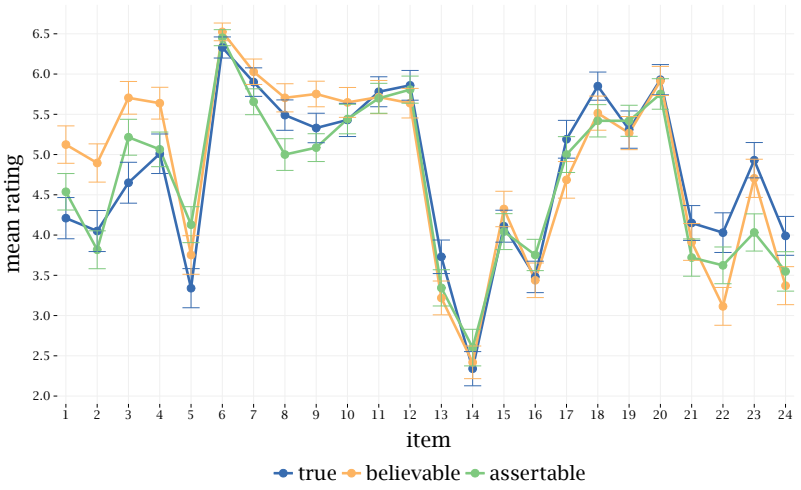

Figure 3: Average ratings of the 24 test items, with error bars indicating one *SE* from the mean. Item numbers correspond again to the numbering in Table 1.

The design, method, and procedure were almost exactly as in the previous experiment, the only difference being that now participants were asked to give their responses on a 7-point Likert-scale. In particular, participants in the true ($N = 100$), believable ($N = 105$), and assertable ($N = 93$) conditions were now asked to judge the truth / believability / assertability on a 7-point scale, with only the anchors being labeled (True / False, Believable / Unbelievable, and Assertable / Unassertable, respectively).

3.2 Results

Figure 3 shows the average ratings of the test items. Comparison with Figure 1 shows a striking similarity between those ratings and the proportions of positive responses obtained in Experiment 1. We also find that not only do the average responses from the three groups in Experiment 2 correlate highly with each other (all rs $> .9$, all associated p's $< .0001$), they also correlate highly with the proportions of positive responses from the corresponding group in Experiment 1 ($r = .97$, $p < .0001$ for true, $r = .93$, $p < .0001$ for believable, and $r = .94$, $p < .0001$ for assertable).

The analyses we conducted were similar to those conducted in Experiment 1.[3] Participants' ratings were coded from 1 to 7, with 1 coding the "negative" anchor of the relevant Likert-scale (False, Unbelievable, or Unassertable) and 7 the "positive" anchor (True, Believable, Assertable) and with the intermediate choice options being coded in the obvious way. Following a recommendation from Aiken and West (1991), we centred the ratings at the midpoint of the scale (i.e., 4) by subtracting 4 from each value. We then built two mixed-effects models with participants' ratings as independent variable and with participants and items as crossed random effects, one model having type of question as

[3] As explained in Field, Miles, and Field (2012, Ch. 14), for analyzing numerical data from a mixed design study one has a choice between performing a traditional ANOVA and using a mixed effects model, though they strongly recommend the latter type of analysis. We follow their recommendation here, also because it makes the outcomes more easily comparable to those from Experiment 1.

fixed effect and the other being an intercept only model. Here, too, we found that adding type of question as predictor did not result in a significantly better model fit: $\chi^2(2) = 1.60$, $p = .449$.

Again we were interested in whether type of question might have an effect for specific types of implicatures, and so we built two further models, one with type of question and type of implicature as fixed effects, and one with those two variables as well as their interaction as fixed effects, both models having the same random effects structure as the previous models. The former model fitted the data significantly better than the model with only type of question as fixed effect: $\chi^2(5) = 27.84$, $p < .0001$; and adding the interaction term resulted in a still better model: $\chi^2(10) = 65.29$, $p < .0001$.

Post-hoc tests with Tukey's HSD adjusted p-values showed that, with the exception of quantificational, for none of the levels of type of implicature did the mean rating of any one group differ significantly from that of either of the other groups. For quantificational items, the mean rating from the believable group was significantly higher than that from the true group ($p < .001$) as well as from that from the assertable group ($p < .05$); the means from the true and assertable groups did not differ significantly from one another. This was again confirmed by comparisons between an intercept-only model and a model with type of question as fixed effect for the various types of implicature, separately. Only for quantificational items did the larger model yield a significantly better fit to the data. This finding is hard to interpret, given that to believe something is to believe it to be *true*, so that truth is a requirement for believability on analytical grounds.

As we did in the analysis of Experiment 1, we conducted four one-way ANOVAs, for the three conditions of type of question separately and for the conditions collapsed, to investigate the effect of type of implicature. The ANOVAs had ratings as outcome variable and type of implicature as predictor variable.

All ANOVAs revealed a significant effect of type of implicature: $F(5, 2394) = 60.95$, $MSE = 4.91$, $p < .0001$, $\eta^2 = .11$ for true; $F(5, 2514) = 87.45$, $MSE = 4.80$, $p < .0001$, $\eta^2 = .15$ for believable; $F(5, 2226) = 71.81$, $MSE = 4.19$, $p < .0001$, $\eta^2 = .14$ for assertable; and $F(5, 7146) = 207.70$, $MSE = 4.70$, $p < .0001$, $\eta^2 = .13$ for the conditions collapsed. The η^2-values in each case indicate an effect size in the medium range. Post-hoc comparisons using Tukey's HSD showed that in the conditions separately and also in the conditions collapsed, all pairs of types of implicature were significantly different from each other at $\alpha = .05$ except for the following: (i) in the case of truth, the pairs quantificational and conventional, order and gradable adjective, ranked ordering and gradable adjective, and ranked ordering and order (all ps $> .27$); (ii) in the case of believable, the pairs conventional and cardinal number ($p = .06$), order and gradable adjective ($p = .91$), quantificational and gradable adjective ($p = .89$), ranked ordering and gradable adjecitve ($p = .82$), quantificational and order ($p = 1$), ranked ordering and order ($p = .21$), and ranked ordering and quantificational ($p = .20$); (iii) in the case of assertable, each pair of gradable adjective, ranked ordering, and order (all ps $> .77$) as well as the pair conventional and cardinal number ($p = .36$); and (iv) in the collapsed condition, each pair of gradable adjective, ranked ordering, and order (all ps $> .07$). The means and standard deviations for the types of implicature in each condition are stated in Table 3.

Spychalska, Kontinen, and Werning's (2016) definition of logicians and pragmatists does not carry over directly to Likert-scale responses, but the fact that of the 100 participants in the true condition, 52 had an average rating smaller than 5 but greater than 3 is enough to suggest that here, too, we fail to find a neat split between participants inclined to judge semantically true sentences with false implicatures as true and participants inclined to judge such sentences as false. Looking at the correlations among participants' truth

Table 3: Summary statistics from Experiment 2 for types of implicatures.

	true		believable		assertable		all	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
quantificational	4.48	0.43	5.34	0.39	4.66	0.63	4.84	0.47
gradable adjective	5.27	1.33	5.50	1.21	5.31	0.99	5.36	1.17
ranked ordering	5.60	0.26	5.69	0.05	5.51	0.32	5.60	0.17
cardinal number	3.42	0.76	3.35	0.78	3.44	0.63	3.40	0.71
order	5.57	0.37	5.35	0.51	5.40	0.31	5.44	0.39
conventional	4.28	0.44	3.78	0.70	3.73	0.21	3.93	0.45

ratings for the test items yields very much the same picture as was seen in Figure 2 for the judgments from Experiment 1; see Figure 4. (Here, too, the patterns for correlations among believability and assertability responses are almost identical to the one seen in Figure 4.)

3.3 Discussion

The results from the second experiment are in broad agreement with those from the first experiment. Most importantly, in spite of the fact that now participants could give more fine-grained responses, we again failed to find a main effect of type of question, supporting both of our hypotheses. To be sure, we found a significant interaction between type of question and type of implicature, and follow-up tests revealed that, for quantificational

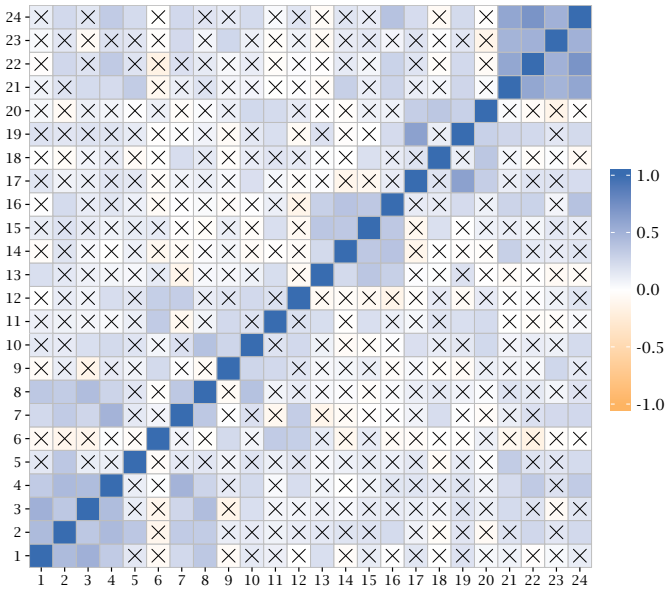

Figure 4: Correlations among participants' truth ratings for the 24 items. (A cross indicates failure to reach significance at $\alpha = .05$.)

items, it did make a difference whether people were to judge their truth or assertability, on the one hand, or their believability, on the other: mean ratings for believability were significantly higher than mean ratings for truth or for assertability. But because believability is considered to entail truth—we cannot take anything to be more believable than we take it to be true—that finding is perhaps best set aside as a fluke.

4 General discussion

Our main research questions were whether laypeople distinguish between the truth and the assertability of a sentence, and whether they distinguish between the assertability and the believability of a sentence. We reported the results from two experiments, which warrant answering both main research questions in the negative. Our materials involved sentences that, on standard semantic thinking, all qualified as true but that, on standard pragmatic thinking, all carried false implicatures, where the implicatures were of six types: four different types of scalar conversational implicature, order conversational implicatures, and conventional implicatures. There were no reliable differences among assessments of the truth of those items, assessments of their believability, and assessments of their assertability, a conclusion that was seen to hold across all types of implicature (with one minor exception that we discussed). We also found no support for the existence of a natural divide between "logical" responders and "pragmatic" responders that previous research had suggested.

We have investigated the semantics–pragmatics interface experimentally by asking participants *directly* for their judgments of truth, believability, or assertability. It could be argued that the better way to proceed here is to make use of indirect measures, such as measuring reaction times (Bott and Noveck 2004) or event-related potentials (Spychalska, Kontinen, and Werning 2016); perhaps it took our participants longer to process our materials that carried false implicatures, than it would have taken them to process similar sentences carrying *true* implicatures, or sentences not carrying any implicatures at all, or perhaps such different materials would have evoked different brain responses. While we do not deny the value of such indirect approaches, it is to be realised that the decision whether to classify whatever causes the differences in reaction times or brain responses as belonging to the realm of semantics or rather to that of pragmatics is not itself something that rolls out of such studies. To make *that* decision, it seems that we will ultimately have to resort to our judgments of whether the implicatures affect the assertability of the items or (also) their truth.

The main lesson to be learned from our experiments is methodological, namely, that in general it will not be possible to settle a debate about what is said by a given sentence and what is implicated by it by consulting the supposedly untainted intuitions concerning truth and assertability of ordinary speakers. It appears that there are no such reliable intuitions that we, as theorists, might be able to exploit. The lesson is decisively *not* that the semantics–pragmatics distinction is illusory or useless. Even if the distinction is strictly theoretical, it may help us make progress in our thinking about language and language use.

To see how it might be helpful, consider that semantics as we currently know it began with the work of Tarski (1935), who was the first to develop a mathematical model of truth in formalised languages. The word "formalised" is crucial here. Specifically, Tarski—a Polish mathematician—was interested mainly in the language of mathematics, which bears similarity to natural languages but also, in many respects, differs from them. The kind of

misleading implicatures that can arise in ordinary conversations, and which motivated the development of pragmatics, are simply nonexistent in the language of mathematics. For instance, the statement of an existentially quantified sentence in a mathematics text does not carry as an implicature the falsity of the corresponding universally quantified sentence.

In the meantime, however, richer mathematical models of language have been developed that were explicitly devised with natural languages in mind. A prominent example of this kind of semantics is the dynamic semantics proposed by Heim (1982), Groenendijk and Stokhof (1991), and others. The crucial observation to make here is that dynamic semantics allows one to model *semantically* phenomena that were traditionally deemed to belong to pragmatics (Gillies 2001). And dynamic semantics is not meant to be the end of all semantic theorising. We may expect to see mathematical models that are still more inclusive and that allow us to represent further aspects of meaning that are currently relegated to pragmatics.

Just think of how important it is for computer scientists and AI researchers to know which aspects of meaning we can, and which we cannot, currently model mathematically. The former might be declared to fall under the heading of semantics, and the latter under the heading of pragmatics. In other words, the semantics–pragmatics interface could be conceived as marking where we are in the process of mathematising language.[4]

References

Aiken, L. S. and West, S. G. (1991) *Multiple regression: Testing and interpreting interactions*, Newbury Park CA: Sage Publications.

Aust, F., Diedenhofen, B., Ullrich, S., and Musch, J. (2013) "Seriousness checks are useful to improve data validity in online research," *Behavior Research Methods* 45:527–535.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008) "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language* 59:390–412.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014) `lme4`*: Linear mixed-effects models using eigen and S4*, `R` package version 1.1–0, `http://lme4.r-forge.r-project.org/`.

Bott, L. and Noveck, I. A. (2004) "Some utterances are underinformative: The onset and time course of scalar inferences," *Journal of Memory and Language* 51:437–457.

Braine, M. D. S. (1978) "On the relation between the natural logic of reasoning and standard logic," *Psychological Review* 85:1–21.

Braine, M. D. S. and O'Brien, D. P (1991) "A theory of *if*: Lexical entry, reasoning program, and pragmatic principles," *Psychological Review* 98:182–203.

Breheny, R. (2008) "A new look at the semantics and pragmatics of numerical quantified noun phrases," *Journal of Semantics* 25:93–140.

Capone, A. (2011) "Knowing how and pragmatic intrusion," *Intercultural Pragmatics* 8:543–570.

Capone, A. (2016) "Indirect reports and slurring," in his *The pragmatics of indirect reports*, Basel: Springer, pp. 145–169.

[4]The first author would like to thank Alessandro Capone for his encouragement over many years to probe more deeply into the pragmatics of belief.

Doran, R. Baker, R.E., McNabb, Y., Larson, M., and Ward, G. (2009) "On the non-unified nature of scalar implicature: An empirical investigation," *International Review of Pragmatics* 1:211–248.

Douven, I. (2008) "The evidential support theory of conditionals," *Synthese* 164:19–44.

Douven, I. (2010) "The pragmatics of belief," *Journal of Pragmatics* 42:35–47.

Douven, I. (2016) *The epistemology of indicative conditionals*, Cambridge: Cambridge University Press.

Edgington, D. (1986) "Do conditionals have truth-conditions," *Crítica* 18:3–30.

Field, A., Miles, J., and Field, Z. (2010) *Discovering statistics using* R, London: Sage Publications.

Gillies, A.S. (2001) "A new solution to Moore's paradox," *Philosophical Studies* 105:237–250.

Grice, H.P. (1989) "Logic and conversation," in his *Studies in the ways of words*, Cambridge MA: Harvard University Press, pp. 22–40.

Groenendijk, J. and Stokhof, M. (1991) "Dynamic predicate logic," *Linguistics and Philosophy* 14:39–100.

Heim, I.R. (1982) *The semantics of definite and indefinite noun phrases*, PhD dissertation, University of Massachusetts Amherst.

Jaeger, T.F. (2008) "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," *Journal of Memory and Language* 59:434–446.

Kratzer, A. (1986) "Conditionals," in A.M. Farley, P. Farley, and K.E. McCollough (eds.), *Papers from the parasession on pragmatics and grammatical theory*, Chicago: Chicago Linguistics Society, pp. 115–135.

Krzyżanowska, K., Wenmackers, S., and Douven, I. (2014) "Rethinking Gibbard's riverboat argument," *Studia Logica* 102:771–792.

Lenth, R. (2015) `lsmeans`*: Least-squares means*, R package version 2.20–23, `http://cran.r-project.org/package=lsmeans`.

Levinson, S.C. (2000) *Presumptive meanings*, Cambridge MA: MIT Press.

Preston, C.C. and Colman, A.M. (2000) "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica* 104:1–15.

R Core Team (2015) R*: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, `http://www.R-project.org/`.

Scharten, R. (1997) *Exhaustive interpretation: A discourse-semantic account*, PhD dissertation, University of Nijmegen.

Spychalska, M., Kontinen, J., and Werning, M. (2016) "Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials," *Language, Cognition and Neuroscience* 31:817–840.

Tarski, A. (1935) "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Philosophica* 1:261–405.