University of Reading

Linguistic experiments and ordinary language philosophy

Article

Accepted Version

Hansen, N. and Chemla, E. (2015) Linguistic experiments and ordinary language philosophy. Ratio, 28 (4). pp. 422-445. ISSN 1467-9329 doi: https://doi.org/10.1111/rati.12112 Available at https://centaur.reading.ac.uk/33469/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.
Published version at: http://onlinelibrary.wiley.com/doi/10.1111/rati.12112/abstract
To link to this article DOI: http://dx.doi.org/10.1111/rati.12112

Publisher: Wiley

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

LINGUISTIC EXPERIMENTS AND ORDINARY LANGUAGE PHILOSOPHY

Nat Hansen and Emmanuel Chemla

Abstract

J.L. Austin is regarded as having an especially acute ear for fine distinctions of meaning overlooked by other philosophers. Austin employs an informal experimental approach to gathering evidence in support of these fine distinctions in meaning, an approach that has become a standard technique for investigating meaning in both philosophy and linguistics. In this paper, we subject Austin's methods to formal experimental investigation. His methods produce mixed results: We find support for his most famous distinction, drawn on the basis of his 'donkey stories', that 'mistake' and 'accident' apply to different cases, but not for some of his other attempts to distinguish the meaning of philosophically significant terms (such as 'intentionally' and 'deliberately'). We critically examine the methodology of informal experiments employed in ordinary language philosophy and much of contemporary philosophy of language and linguistics, and discuss the role that experimenter bias can play in influencing judgments about informal and formal linguistic experiments.[1]

1 Introduction

J.L. Austin criticized traditional approaches to philosophical problems for ignoring and distorting the 'ordinary' meaning of philosophically significant expressions. Austin is still regarded as having an especially acute ear for fine distinctions of meaning overlooked by other theorists.[2] He employed an informal experimental approach to gathering evidence of these fine distinctions in meaning, an approach that has become a standard technique for investigating meaning in both philosophy and linguistics.

Austin's method of demonstrating fine distinctions between words or phrases with similar meanings involves constructing a type of linguistic experiment, in which we are

[1]Thanks to Zed Adams, Jonas Åkerman, Emma Borg, Aidan Gray, Chauncey Maher, participants at the conference on Empirical Data and Philosophical Theorizing at the University of Barcelona, the conference on the Contemporary Significance of Ordinary Language Philosophy at Åbo Akademi, and at the *Ratio/CCR* conference on Semantics and Science at the University of Reading for comments on this paper. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n^o 229 441 - CCC.

[2]A recent example of this view can be found in Livingston (2012): 'In a footnote to "A Plea for Excuses", J. L. Austin offers one of the precisely chosen examples that illustrate the keen ear for the language, and almost unmatched capacity for noting its fine distinctions, for which he and his method of reflection on ordinary language were justly notorious'. See also Williams (2014, p. 44), who says 'Austin was fascinated by, and extraordinarily gifted in detecting, nuances of ordinary speech', Searle (2001, p. 226), who writes that 'Austin did indeed have a genius for spotting linguistic differences and distinctions where most people would have thought there were none', and Cavell (1994, p. 21), where the ability to produce examples of ordinary language use is compared to perfect pitch.

either asked to judge whether it is better to use one or the other of the words being investigated in a particular situation, or to judge whether a particular expression better fits one or the other of a pair of situations that differ only in limited ways from one another. Austin himself says that this is a method of gathering 'experimental data', where the explanation of that data will be an account of the meaning of the expressions under investigation:

> First let us consider some cases. Actual cases would of course be excellent: we might observe what words have actually been used by commentators on real incidents, or by narrators of fictitious incidents. However, we do not have the time or space to do that here. We must instead imagine some cases (imagine them carefully and in detail and comprehensively) and try to reach agreement upon what we should in fact say concerning them. If we can reach this agreement, we shall have some data ("experimental" data, in fact) which we can then go on to explain. Here, the explanation will be an account of the meanings of these expressions, which we shall hope to reach by using such methods as those of "Agreement" and "Difference": what is in fact present in the cases where we do use, say, "deliberately", and what is absent when we don't. Of course, we shall then have arrived at nothing more than an account of certain ordinary "concepts" employed by English speakers: but also at no less a thing. (Austin 1966, p. 429)[3]

Such an investigation of the meaning of words like 'deliberately' is interesting in itself, as a contribution to a descriptive theory of meaning for English (as the passage from Austin just quoted suggests), but it is also interesting insofar as the words and concepts being investigated figure in philosophical disputes. It is this latter project that prompts Austin to investigate the meaning of expressions like 'mistake', 'accident', 'deliberately' and 'intentionally', specifically the role that these expressions play in discussions of responsibility, 'what actions are good or bad, right or wrong' (Austin 1957, p. 4), the nature of action, and 'the problem of Freedom' (Austin 1957, p. 6).

In the middle of the 20th century there was a heated debate about whether ordinary language philosophy involved a method of investigating meaning that was distinct from the methods of empirical linguistics. Mates (1958) argues that ordinary language philosophers used unreliable methods of gathering data, while Cavell (1958) defends the distinctiveness of ordinary language philosophy, comparing claims about what 'we' say to instances of 'Transcendental Logic', which are subject to different standards of criticism and justification than ordinary claims about how people use language. Rejecting this idea, Fodor and Katz (1963) argue that there is no defensible distinction between

[3] Austin was reflecting on how to gather evidence bearing on our understanding of meaning at the dawn of the turn to using introspective, intuitive judgments (of acceptability, e.g.) as evidence for linguistic theories. Though the paper from which this quote is taken, 'Three Ways of Spilling Ink', was first published in 1966, after Austin's death, the lecture on which the paper is based was delivered in 1958. See the editor's note in Austin (1966, p. 427).

the kind of justification for statements about ordinary language sought by the linguist and those sought by the ordinary language philosopher. However, there have been various subsequent attempts (Henson 1965, Friedman 1969, Bates and Cohen 1972, Hanfling 2000, Sandis 2010, Baz 2012) to defend the distinctiveness of ordinary language philosophy. In this paper, rather than entering directly into that debate, we assume that experimental data is relevant to the ordinary language approach to the investigation of meaning on the grounds that Austin says that it is. There may still be something else ordinary language philosophers are doing besides collecting and explaining 'experimental data' (as Austin says), but that is a topic for another paper.[4]

Austin's most famous experiment concerns the difference between two expressions that, at first glance, may not seem to differ significantly in meaning: 'mistake' vs. 'accident'. He says that the choice between these expressions can '*appear* indifferent…Yet a story or two, and everybody will not merely agree that they are completely different, but even discover for himself what the difference is and what each means' (Austin 1957, pp. 10–11). To distinguish the meaning of these expressions, Austin sets up an experiment, involving his well-known 'donkey stories' which makes it seem clear that 'by mistake' better describes the action in one situation, and 'by accident" better describes the action in the other, thereby providing evidence that the meanings of the two expressions are indeed distinct. Cavell (1965, p. 211) says that Austin 'inspire[s] revelation' with the donkey stories, which go as follows:

> You have a donkey, so have I, and they graze in the same field. The day comes when I conceive a dislike for mine. I go to shoot it, draw a bead on it, fire: the brute falls in its tracks. I inspect the victim, and find to my horror that it is *your* donkey. I appear on your doorstep with the remains and say—what? 'I say, old sport, I'm awfully sorry, &c., I've shot your donkey by *accident*? Or '*by mistake*'? Then again, I go to shoot my donkey as before, draw a bead on it, fire—but as I do so, the beasts move, and to my horror yours falls. Again the scene on the doorstep—what do I say? 'By mistake'? Or 'by accident'? (Austin 1957, p. 11 n. 4)

Austin's donkey stories are so compelling, in fact, that he does not even need to say what the most appropriate response to each situation is, and yet most of those who read his example reach the same conclusion about which term to apply to which situation: 'by mistake' better suits the first story, and 'by accident' the second.[5]

In other places Austin tells similar stories with the same aim of drawing subtle distinctions between the meaning of certain phrases, but accompanies those stories with

[4]See Hansen (2014a) and Jackman (2001) for discussion of different conceptions of the project of ordinary language philosophy.

[5]For explicit endorsements of the standard response to the donkey stories, see Gustafsson (2005, p. 368), 'We all agree that in the first scenario the donkey was shot by mistake, whereas in the second scenario it was shot by accident' and Hanfling (2000, p. 64), 'The first case is "by mistake", the second "by accident'".

his own powerful glosses on 'what we should say'. For example, the gocart story from Austin (1966) is intended to distinguish the meaning of 'intentionally' and 'deliberately':

> I am summoned to quell a riot in India. Speed is imperative. My mind runs on the action to be taken five miles down the road at the Residency. As I set off down the drive, my cookboy's child's new gocart, the apple of her eye, is right across the road. I realize I could stop, get out, and move it, but to hell with that: I must push on. It's too bad, that's all: I drive right over it and am on my way. In this case, a snap decision is taken on what is essentially an *incidental* matter. I did drive over the gocart deliberately, but not intentionally—nor, of course, unintentionally either. It was never part of my intention to drive over the gocart. At no time did I intend to drive over it. It was incidental to anything I intended to do, which was simply to get to the scene of the riot in order to quell it. However "odd" it may sound, I feel little doubt that we should say here that we did run over the gocart deliberately *and* that we should not care to say we ran over it intentionally. We never intended to run over it. (Austin 1966, p. 432)

After reading the story and Austin's gloss, the fact that the narrator ran over the gocart deliberately but not intentionally seems convincing.[6] But is it as obvious as it is in the donkey story what the right thing to say about the story is? Or is Austin's gloss significantly affecting our response? Would a different gloss have produced different judgments about the story? What would happen if the gocart story were presented without any gloss at all? Would judgments still align with Austin's reading of the story?

The same questions about the role of Austin's gloss apply equally to the racing-car story from Austin (1958, p. 272):

> A boy in an arm-chair is making tugging and twisting movements with his arms, accompanied by gear-change and other raucous noises. He is 'pretending to be driving a racing-car', but scarcely 'pretending to drive a racing-car'. Why? A *possible* answer is this. In neither case is the behavior of the pretending party sufficiently like the genuine article…for it to be in point to mark the distinction between the two. To pretend to drive a racing-car, he would need a racing-car…

And once one begins to worry in general about the techniques that Austin uses to get his readers to agree about what to say about his gocart and racing-car stories, subtler worries about his methods of gathering data arise as well. Consider the donkey stories again. There is an interesting asymmetry in the way Austin presents the options 'by mistake' and 'by accident' in the two versions of the story he tells.[7] In both versions, the

[6] At least, it has seemed convincing to some. For examples of favorable citations of Austin's reading of the gocart story, see Ferguson (2003, p. 93), Searle (2001, p. 223), Williams (2009, p. 24).

[7] This possibility was suggested in conversation by Aidan Gray and Sören Häggqvist.

response that people tend to give is presented *second*. One might worry that the asymmetry in the response options is influencing judgments about the donkey stories. What would happen if the options were reversed?

In this study, we aim to answer those questions, and in so doing we will investigate some easily overlooked aspects of the experiments used in some paradigm cases of ordinary language philosophy and discuss the ways that they may corrupt the results of otherwise legitimate experimental investigations of meaning.

2 Linguistic Experiments: Formal vs. Informal

Linguists and philosophers of language employ various methods of gathering data: Corpus studies draw on naturally occurring uses of language and linguistic field work aims to collect examples of language use by transcribing or recording language use 'in the wild', while linguistic experiments 'create, produce, refine and stabilize phenomena' (Hacking 1983, p. 230) in controlled circumstances. The control that experimenters have over the data generated by experiments confers an enormous practical advantage over straightforward observation, but it brings with it a corresponding risk that experimenters are merely creating *artifacts* that emerge only because of the particular experimental design they are employing. So it is the responsibility of experimenters to pay close attention to the design of their experiments to ensure that they are uncovering evidence of the phenomena they take themselves to be investigating.

The experiments that Austin employs, like many of the experiments employed in contemporary investigations of meaning, are *informal*—they do not involve gathering judgments from large groups of participants, running statistical tests of significance on the data gathered, and so on. Instead, they involve the theorist reporting her own judgments about situations that she herself describes, and presenting those judgments alongside the situations in the context of an academic paper. Though there are many differences of procedure between informal and formal experiments, these informal experiments are similar in an essential respect to the more formal experiments conducted by linguists.[8] Both formal and informal linguistic experiments aim to create circumstances in which it is possible to observe the effect of an independent variable (some feature of the context, for example) on a dependent variable (acceptability judgments or truth-value judgments, for example).

One central difference between informal and formal experiments concerns how rigorously the experimenter tries to control for particular types of biasing factors, such as:

(a) the *order* in which conditions or response options are presented[9]

(b) contrast (this includes both whether participants are allowed to see contrasting

[8] See Sprouse et al. (2013) for an illuminating discussion of similarities and differences between formal and informal linguistic experiments.

[9] See Schwitzgebel and Cushman (2012) for discussion of the effects of order of presentation of conditions.

experimental conditions—as in a within-subjects design—or not—as in a between-subjects design—and whether there are contrasting response options)[10]

(c) experimenter bias[11]

(d) effects of response scales[12]

Controlling for those potentially biasing factors might involve systematically varying the order in which objects of evaluation or response options are presented, employing a between-subjects design (or a design that allows experimenters to collect responses in both circumstances with and without potential contrast effects, as in Hansen and Chemla 2013), removing or systematically varying the experimenter's own preferences, and employing different types of response options (binary t/f judgments, Likert scales, magnitude estimation, and so on). An experimenter might refrain from going to the trouble of employing those types of controls if she believes the phenomenon she is investigating is so pronounced that experimental biases could not obscure it. (And this is typically what happens when informal experiments are conducted and not supplemented with formal experiments.)

But if a particular phenomenon is disputed, or if an effect is weak enough that small, external fluctuations might obscure it, or if there is a specific reason to think that the presence of some particular bias is distorting the data generated by an informal experiment, then there is good reason to conduct a more formal experiment. We think there is reason to worry about the effects of two particular types of bias in Austin's informal experiments—namely the role that glosses play in generating a form of experimenter bias, and the role played by the order in which response options are presented in Austin's donkey stories. We conducted more formal linguistic experiments with the aim of evaluating the role such biases play in Austin's informal experiments.

3 Summary of results

In this study, we evaluated two hypotheses:

1. Glosses play an essential role in experiments *à la* Austin; if the glosses are removed or reversed, different judgments will be obtained.

2. The order in which response options are presented in experiments *à la* Austin (e.g., with Austin's donkey stories) may have an effect on response preferences.

In brief, we found evidence in support of hypothesis 1 but not hypothesis 2.

[10]For discussion of the role that contrasting experimental conditions play in recent work in experimental philosophy, see Hansen (2014b) and Phelan (2013).

[11]See Hansen (2013) and Strickland and Suben (2012) for recent discussions of experimenter bias.

[12]See Cullen (2010).

Re 1. We found evidence that glosses have an impact on judgments. We reproduced Austin's claims when the gocart and racing-car stories (from Austin 1966 and 1958, respectively—to be discussed below) were presented with Austin's glosses. But when the gocart and racing-car stories were presented with a gloss that suggested the *reverse* of Austin's claims about those stories, participants reversed their responses to both stories. And crucially, when presented *without* an accompanying gloss, the gocart story generated no preference at all, and the racing-car story generated the *opposite* of Austin' claims.

Re 2. We did not find evidence that the order of response options has any effect on judgments about the donkey stories. Under experimental conditions, we reproduced the expected difference in judgments about the two donkey stories, both when the order of responses matched and did not match the original order proposed by Austin. Note that we obtained these judgments from participants who could not contemplate and compare the two stories, which shows that the judgments obtained by Austin were not the result of the explicit contrast between the stories.

Our findings indicate that Austin's methodology is a mixed bag: On one hand, he was able to generate robust results (as with the donkey stories), on the other, some of his results are undermined by the effects of bias.

4 Racing-Cars and Gocarts

Our interest in the gocart and racing-car stories was prompted by a worry about the role that *experimenter bias* plays in influencing judgments about informally presented linguistic experiments. Experimenter bias is an effect generated when experimenters disclose (even unconsciously) their own expectations about the outcome of an experiment.[13] Because the experiments employed by ordinary language philosophers are presented informally, they are susceptible to certain forms of experimenter bias, the most obvious being the fact that in some cases the theorist simply states what his preferred judgment about the story being evaluated is. Our first experiment gathers evidence about the role played by experimenter bias in influencing responses to two of Austin's experiments.

To evaluate the effects of experimenter bias, we offered participants one of three versions of one or the other of the gocart and racing-car stories (the versions differed from one another only in terms of their accompanying glosses—the details of the story, which contain the factors that are supposed to be those influencing responses, remained exactly the same):

1. Austin's original gloss
2. Reversed gloss
3. No gloss

[13] See Doyen et al. (2012), Intons-Peterson (1983), and Rosenthal (1976, Ch. 8) for discussions of unconscious forms of experimenter bias.

4.1 Material for the gocart experiment

We made some minor amendments to the original version of the gocart story to make it easier to understand (e.g., replacing 'cookboy' with 'cook', and replacing the phrase 'right across the road' with 'in the middle of the road'), and, to simplify presentation, the story was changed from a first-person narration to a third person account of 'George'.[14] It's important to note that the changes we made to the stories were constant across conditions so that any overlooked influence they may have should be the same in all conditions. The story that remained constant across the three conditions read as follows:

Gocart Story

> George is summoned to quell a riot in India. Speed is imperative. His mind runs on the action to be taken five miles down the road at the Residency. As he sets off down the drive, his cook's child's new gocart, the apple of her eye, is in the middle of the road. George realizes that he could stop, get out, and move it, but to hell with that: he must push on. It's too bad, that's all: He drives right over it and is on his way.

The following is our revised version of the original gloss, which accompanied the story in the first experimental condition:

> **Original Gloss:** In this case, a snap decision is taken on what is essentially an incidental matter. George did drive over the gocart deliberately, but not intentionally. It was never part of his intention to drive over the gocart. At no time did he intend to drive over it. It was incidental to anything he intended to do, which was simply to get to the scene of the riot in order to quell it. However 'odd' it may sound, we should say here that George did run over the gocart deliberately and that we should not say he ran over it intentionally. He never intended to run over it.

And the following is the reversed gloss, which accompanied the story in the second experimental condition:

> **Reversed Gloss:** In this case, a snap decision is taken on what is essentially an incidental matter. George did drive over the gocart intentionally, but not deliberately. At no time did he deliberate whether to drive over the gocart. He simply intended to get to the scene of the riot in order to quell it. However 'odd' it may sound, we should say here that George did run over the gocart intentionally and that we should not say he ran over it deliberately. He never deliberated about running over it.

[14] A critic might worry that by changing from the first to the third person, we are altering a significant feature of Austin's stories, and thereby not really experimenting with the same stories. If that's right, then the critic should think of the experiments we run as posing a challenge to Austin's original stories in virtue of the fact that they show the importance of glosses in closely related stories; the burden is then on a defender of the original experiments to show that the glosses are not playing the same role in the original stories.

The third experimental condition, corresponding to the absence of a biasing gloss, consisted solely of the revised story plus the prompt (which appeared in all three conditions):

> Question: Which of the following best describes what George did?
>
> (a) George ran over the gocart intentionally.
> (b) George ran over the gocart deliberately.[15]

4.2 Materials for the racing-car experiment

The racing-car experiment was identical in structure to the gocart experiment, giving separate groups of participants one of three versions of the following racing-car case (drawn from Austin 1958, p. 272):

Racing-car story

> A boy in an arm-chair is making tugging and twisting movements with his arms, accompanied by gear-change and other raucous noises.
>
> **Original Gloss:** He is 'pretending to be driving a racing-car', but not 'pretending to drive a racing-car'. To pretend to drive a racing-car, he would need a racing-car.
>
> **Reversed Gloss:** He is 'pretending to drive a racing-car', but not 'pretending to be driving a racing-car'. To pretend to be driving a racing-car, he would need a racing-car.

The first two conditions in the racing-car experiment featured the same prompt:

> Question: Which of the following best describes what the boy is doing?
>
> (a) The boy is pretending to be driving a racing-car.
> (b) The boy is pretending to drive a racing-car.

The third condition featured a slightly different prompt to give participants relevant information that appears in both glosses:

[15] What if participants have no preference between the two responses? They would be forced to choose between (a) and (b) arbitrarily, presumably leading to a 50/50 split in responses. Such a pattern could be interpreted as an absence of preference (although one should avoid drawing strong conclusions from a null result). One might worry that the absence of a third, no-preference option could force participants to search for a preference even when they don't see one immediately, and that might lead to participants relying on irrelevant factors to make a choice between (a) and (b). But that would distort our results only if there were a non-arbitrary factor that participants were relying on—a factor that we would have failed to detect and which would apply consistently across our items instantiating the conditions and across our participants. That is possible, but unlikely. And if indeed there is such a salient factor that can help participants so robustly, it's likely that adding a third "no preference" option would not have blocked participants from finding it. Thanks to Emma Borg for raising questions about the limitations of the forced-choice response option.

Question: Given that he is not in an actual racing-car, which of the following best describes what the boy is doing?

(a) The boy is pretending to be driving a racing-car.
(b) The boy is pretending to drive a racing-car.

4.3 Participants

We recruited 370 participants over Mechanical Turk (see Sprouse 2011 for discussion). Each participant was paid $0.05. Each of the participants saw only one of either the go-cart or racing-car scenario, in only one of the glossing conditions described above (original, reverse, neutral). Participants were also asked to answer three simple control questions to ensure they were paying attention ('What is nine minus three?', 'What is three plus three?', 'What is two times three?') and to report their native language (with no incentive in favor or against reporting English as a native language). Participants who did not report English as their native language (10 participants) or who failed to answer one of the three simple questions (3 more participants) were paid but excluded from the analyses.

4.4 Results

Figure 1 shows the number of participants who opted for each option in each condition. We will describe these results from left to right. First, and unsurprisingly, with the original gloss we obtained significantly more answers corresponding to Austin's judgments both for the racing-car scenario ($\chi^2(1) = 16, p < .001$) and for the gocart scenario ($\chi^2(1) = 30, p < .001$).[16]

*Number of answers of each type for the **racing-car** scenario:*

	Original gloss	Reversed gloss	No gloss
to be driving	46	2	9
to drive	15	61	48

*Number of answers of each type for the **gocart** scenario:*

	Original gloss	Reversed gloss	No gloss
deliberately	50	5	27
intentionally	8	54	32

Figure 1: Number of participants who responded in accordance with the original judgments about the stories are indicated by red bars, while the number who responded contrary to the original judgments are indicated with blue bars.

Second, and importantly for our purposes, we found that this preference was reversed when the gloss was reversed, both for the racing-car scenario ($\chi^2(1) = 55, p <$

[16]We report the results of Chi-square tests with Yates' continuity correction. We obtained similar results with exact binomial tests for pairwise comparisons where it applies.

.001) and the gocart scenario ($\chi^2(1) = 41, p < .001$).[17] Of course, this reversal could be the mere result of our participants trying to conform with 'instructions'. In the context of our experiment, participants may not read glosses as opinions from colleagues that are open to discussion, but rather simply as instructions indicating the correct answers.

The gloss-free condition reveals that the original gloss distorts how ordinary speakers respond to the story. For the racing-car scenario, participants show a clear preference even in the absence of a biasing gloss ($\chi^2(1) = 27, p < .001$). If there was no fact of the matter and no difference between the phrases (in this case 'to be driving' vs 'to drive') and if the glosses were the only guide available to our participants, we would expect to find no preference in the condition without a gloss.[18] But, strikingly, the preference obtained in this neutral condition is *the opposite* of Austin's judgment and the one obtained with the original gloss ($\chi^2(1) = 27, p < .001$). This is our most solid result: removing the gloss reveals preferences that are the opposite of how they appear with the gloss.

Responses to the gocart scenario when it was not accompanied with a gloss are also interesting. In the gloss-free condition, we found no preference between judgments that the gocart was run over 'intentionally' or 'deliberately' ($\chi^2(1) = .42, p = .52$). The parallel gloss-free condition with the racing-car scenario shows that our setting is able to reveal preferences, when they exist. Hence, the absence of preference for either response in the gocart scenario suggests that at best the preference was over-estimated by Austin and those who endorse his claims about the gocart story: our data does not indicate that there is a preference beyond the effect of the gloss.[19]

Note as well that despite the absence of a preference for either 'intentionally' or 'deliberately', there is a tangible effect here: the results obtained without a gloss are statistically different from those obtained with the original gloss ($\chi^2(1) = 20, p < .001$) and the reversed gloss ($\chi^2(1) = 19, p < .001$). That confirms that the gloss may be corrupting the data in Austin's original gocart story. (That is, there may be no difference between 'intentionally' and 'deliberately', but the presence of the gloss makes it look as if there is.)

---

[17] The 'interaction' between response and original/reversed gloss, is also found to be significant both for the racing-car ($\chi^2(1) = 65, p < .001$) and for the gocart scenario ($\chi^2(1) = 68, p < .001$).

[18] This remark only applies to the gloss-free condition. It remains possible, as mentioned above, that *when* there is a gloss, participants follow the gloss.

[19] Independent effects may artificially raise or decrease the acceptability of describing the action in the gocart scenario as done 'intentionally' or 'deliberately'. One salient candidate is the so-called 'Knobe effect' (see, e.g. Knobe 2006). In a nutshell, the Knobe effect would make the use of 'intentionally' more acceptable in scenarios with a negative moral valence than in scenarios with a positive valence. One might then reason that if the gocart scenario has a negative valence, preferences for describing the running over of the gocart as done 'intentionally' would receive a boost, which could erase the preference there exists otherwise in favor of saying that the gocart was run over 'deliberately'. There are two things to say about this suggestion. First, note that Austin's judgments should also have been affected by such an effect. So the existence of the Knobe effect would not explain a difference between Austin's judgments and the judgments we report. Second, for all we know, 'deliberately' may be subject to an equivalent Knobe-style effect (see Pettit and Knobe 2009 and Egré 2014 for arguments that the effect goes beyond judgments about the word 'intentionally'). If that's right, then Knobe-style effects would have parallel consequences for both phrases, and they would not alter the preference we uncovered.

4.5 Discussion

First, we found that responses to the neutral stories, unaccompanied by glosses, were either the *reverse* of Austin's verdict (for the racing-car story), or that, contrary to Austin's verdict, there was no significant difference between responses (for the gocart story). Second, reversing the accompanying glosses reversed participants' judgments about both the racing-car and gocart stories. The lesson of these findings is that glosses can influence judgments about the meaning of words in two different ways: (1) The presence of a gloss can make it look like there is a preference one way, when in fact the preference goes the other way (as in the racing-car story); or (2) the presence of a gloss can make it look like there is a preference one way, when in fact there is no preference (as in the gocart story). The fact that we found positive evidence that responses to the racing-car story are the reverse of Austin's original judgment about that story makes it plausible that the second error is also possible.

The role played by glosses for those participating in a formal experiment is probably different from the role it plays for philosophers reading an academic article. The 'workers' who are the participants in the experiment are motivated to *perform correctly*, and likely understand the accompanying glosses as a kind of instruction for how to respond. With that in mind, it's not surprising that we found reversals of judgment about the stories when they were accompanied with reversed glosses. Philosophers *might* be more resistant to the glosses than the Mechanical Turk 'workers' are, but there is evidence that philosophers are no less influenced by other, extraneous features of experiments than ordinary participants (e.g. see Schwitzgebel and Cushman 2012, who show that philosophers's judgments about well-known moral thought experiments are affected by order of presentation). And, for what it's worth, we find that we can feel our own responses to the stories being swayed by the different glosses. But the key result in the gocart and racing-car examples is the fact that the presence of glosses can obscure, rather than reveal, judgments about key features of the stories themselves. The key result thus comes from the racing car scenario that is not accompanied by a gloss. In that condition, there is no appearance of an incentive (as there may be in the cases accompanied with glosses) for workers to prefer one answer over the other. Yet they do provide one of the answers more robustly than the other—and their preference goes in a direction opposite to Austin's judgment about the story.[20]

[20]Jennifer Nagel remarked in conversation that we should leave open the possibility that our participants and Austin (and other scholars who have responded to his scenarios) speak different dialects (1950s Oxford English in Austin's case, and a variety of other dialects for contemporary participants). However, our main claims are not affected by any dialectal differences. We show that a minimal change in a given experimental situation (namely the presence or absence of a gloss) alters its outcome, and we conclude that experimenters should beware of the possibility of experimenter biases.

We are convinced that there are indeed dialectal differences. But we do not know how to assess accurately their consequences for our task. One may compare frequencies of the relevant phrases in corpora from Austin's era and from ours. But even large frequency changes would not show that semantic changes have taken place. In fact, it is worth recalling that one important feature of language infinitude is that phrases that appear with extremely low frequency are not beyond the reach of coherent seman-

5 Shooting Your Neighbor's Donkey

The second hypothesis we were interested in evaluating was that responses (to the donkey scenario) are coerced through a subtle form of order of presentation bias, concerning the order in which the response options ('by mistake', 'by accident') are presented in Austin's original stories. To test this hypothesis, we presented subjects with slightly revised versions of the original donkey stories, again switching first personal elements to third personal ones, and in which participants saw only one or the other of the two stories, thereby eliminating any contrast effects. Each story was accompanied with either a response option featuring the original order ('mistake…accident' in the first story, and 'accident …mistake' in the second), or the reversed order. So participants saw one of four possible combinations of story and response pairs, as follows:

1. *Mistake* story + original order of responses
2. *Mistake* story + reversed order of response
3. *Accident* story + original order of responses
4. *Accident* story + reversed order of responses

The two, lightly revised donkey stories read as follows:

Mistake

> John has a donkey, so does Mary, and the donkeys graze in the same field. The day comes when John comes to dislike his donkey. He decides to shoot it, draws a bead on it, fires: the brute drops dead. He inspects his victim, and finds to his horror that it is Mary's donkey.

Accident

> John has a donkey, so does Mary, and the donkeys graze in the same field. The day comes when John comes to dislike his donkey. He decides to shoot it, draws a bead on it, fires—but as he does so, the beasts move, and to his horror Mary's donkey drops dead.

Contrary to our hypothesis, we did not find evidence that order made any difference to how participants responded to the donkey stories.

---

tic judgments. A best case scenario for this line of inquiry would be one according to which frequency biases affect Austin's judgments and our participants differently (because the relative frequency of the relevant phrases is different for Austin and for our participants). But even if that's the case, we would still be a long way from knowing how frequency affects judgments. Furthermore, this thesis and its putative demonstration is less parsimonious than our own explanation of the difference in judgment in terms of the experimenter bias we originally targeted, which we would expect to be equally active across dialects.

5.1 Procedure and participants

We recruited 248 participants on mechanical turk, who participated for $0.05. As with the previous study, we excluded 3 participants from the analyses because they did not report to be native speakers of English, and 4 more because they failed to answer simple control questions.[21]

5.2 Results

Figure 2 presents the number of participants in each condition who opted for each response choice ('by mistake' vs. 'by accident').

Mistake Scenario:

	Aggregated results	Original order	Reverse order
by accident	22	15	7
by mistake	96	43	53

Accident Scenario:

	Aggregated results	Original order	Reverse order
by accident	58	30	28
by mistake	65	31	34

Figure 2: Number of participants who responded with 'by mistake' and 'by accident' in different conditions.

Before we consider order effects, consider the aggregated results (the first column of results in the tables from Figure 2). Participants' answers reveal a preference for 'by mistake' in the *Mistake* scenario ($\chi^2(1) = 46, p < .001$). This preference confirms the standard response to the first donkey story. In the *Accident* scenario, no preference emerges ($\chi^2(1) = .40, n.s.$). Such an absence of a preference, a so-called *null-result*, is in general difficult to interpret, because:

> ...there are many reasons why a study may fail to uncover a relation between variables even when the relation does in fact obtain. One may be relying on instruments that do not have the necessary degree of resolution to detect the relevant relation, for example. And every experimental result is noisy to some degree. An absence of difference cannot establish that the difference does not exist, unless one also proves the counterfactual claim that the experiment would have been sufficiently powerful to detect it (Hansen and Chemla 2013, p. 7).

[21] We recruited one set of participants for the whole study, including gocart scenarios, racing-car scenarios and the donkey scenarios. Participants were distributed randomly between the different conditions presented in the whole study. The target was to obtain around 60 participants in each condition. Because the assignment to conditions was random and because some participants were disqualified from the analyses numbers are not perfectly smooth.

Let us list three possible ways of interpreting the null result that we found in the *Accident* version of the donkey story. First, it may be due to the absence of an effect. Second, an absence of result may also be obtained if the *methodology* we employed is not suited or powerful enough to detect the effect. Third, and more specific to the example under consideration, the absence of a visible preference may be the result of the combination of (i) an actual preference, say for 'by accident', as expected, and (ii) an independent bias against responding 'by accident', e.g., because the phrase is less frequent than 'by mistake'.[22,23] In the absence of a reliable preference, it is not possible to exclude any of these various possibilities. The upshot is that it is not possible to determine whether or not there exists a preference between 'by mistake' and 'by accident' in the *Accident* version of the donkey story.

The data we gathered do contain a tangible, second-order effect, however. Participants reliably *distinguish* the two scenarios. Their preference pattern for 'by mistake' vs. 'by accident' varies from one scenario to the other: the two cells in the 'aggregated results' are different, the preference *changes* (although it is not fully reversed) from the *Mistake* scenario to the *Accident* scenario ($\chi^2(1) = 21, p < .001$). As a result, we can conclude that some difference between the scenarios is affecting judgments about the (relative) appropriateness of the two expressions. Hence, some difference in the scenario corresponds to some difference of meaning in the two expressions. This conclusion is possible even in the absence of a clear preference in the second scenario, because we can make use of the powerful two-way contrasts between both phrases and scenarios employed in this experiment.

However, despite the power of such a design, for the sake of methodological rigor, it is worth discussing two issues which remain undecidable given our data.

- First, the preference observed in the *Mistake* scenario does not show that participants judged that one phrase is correct and the other is incorrect. It is possible that participants are judging both to be correct or incorrect, and the contrast we observed is due to there being a slight preference for one over the other. That is, since we asked participants to judge which of the two phrases *best* describes the

[22]Google searches actually return more hits for 'by accident' (28M) than for 'by mistake' (20M). Hence, frequencies go against (ii) in this particular case. So one may disregard this hypothesis, but the point is more general: for all we know (and for all we tested), our setting may generate a bias *against* frequent phrases.

[23]In principle, a visible preference could also be the result of a bias rather than a genuine preference. Concretely, the preference observed for 'by mistake' in the first donkey scenario could in principle be the result of (i) an absence of preference (or even a reversed preference) and (ii) a bias in favor of responding 'by mistake' or against responding 'by accident'. The reason why such an hypothesis is not entertained (and generally in similar situations) is that the result conforms with the expectation. Suppose a theory T predicts the presence of an effect E and we do not find the effect or find an opposite effect E'. It is then necessary to set up a new theory or to supplement theory T so as to explain effect E'. (At this point, supplementing theory T may simply explicate the biases that may be at play to turn E into E'). If on the contrary we do find effect E, nothing calls for an explanation. In Bayesian terms: if an experiment is set up with the underlying belief that some cause C would generate an effect E, then actually observing E should reinforce our prior belief that C is the (actual) cause of E.

situation, they may be merely recording a small difference in preference between the two phrases. At the theoretical level, this highlights the fact that judgments of preference are not directly linked with truth-conditional distinctions.

- Second, in the *Accident* scenario, in addition to not being able to conclude that participants have a preference for one phrase or the other, our data do not indicate whether both phrases are equally correct or equally incorrect or somewhere in between. As a result, it is not possible to use our results to argue that one phrase is more appropriate in one scenario than in the other. We can only conclude that the *preference* for one phrase disappears when we move from one scenario to the other, but whether the felicity/appropriateness of the corresponding phrases increases or decreases cannot be determined.

In summary, appropriateness judgments cannot be unambiguously derived from preference judgments. Different possible situations may give rise to a preference for A over B: (a) A is good, B is bad; (b) A is very good, B is good; (c) A is bad, B is very bad, etc. Similarly, different situations may give rise to an absence of preference (independently of biases hiding an underlying real preference): (d) A is good, B is good; (e) A is bad, B is bad; (f) A is intermediate, B is intermediate. The point is that, in a preference paradigm, (a), (b) and (c) are not distinguishable and (d), (e) and (f) are not distinguishable. Possibility (a) is a somewhat privileged case of preference, and in informal experiments there may be a hope that when they find themselves in situations (b) or (c) rather than (a), experimental participants will speak up when they express their preference, indicating that while they have a preference for one option over the other, they find both options are acceptable or unacceptable (as the case may be). Overall, the preference ('by mistake' vs. 'by accident') and contrast (*Mistake* and *Accident* scenarios) design employed in the donkey experiment is very powerful, but the interpretation of results remains a delicate matter.

Let us now turn to the order effects we were primarily interested in. There is no evidence that the results in the last two columns are different, either for the *Mistake* scenario ($\chi^2(1) = 3.0, p = .08$), or for the *Accident* scenario ($\chi^2(1) = .07, p = .80$). In other words, there is no evidence that the order of presentation of the response choices have an influence on participants' eventual decision: their preference for 'by mistake' is independent from this option occurring first or second, and the absence of a visible preference is also independent of order.

5.3 Discussion

We found that in controlled experimental circumstances, responses aligned with standard judgments about the donkey stories, and contrary to our hypothesis, this remains true even when order effects are factored out. It may appear at first glance that only responses to the 'mistake' story align with standard judgments about that story, since there was no significant difference between the 'by mistake' and 'by accident' responses to the 'accident' story. But what we found was a significant *contrast* between responses to the two stories that aligns with standard existing judgments about the stories.

Notice that the contrast between the donkey stories emerges even though participants only saw one or the other story, not both as in Austin's original presentation. And the donkey stories are not accompanied by any gloss. Both of those factors indicate that responses to the stories are tracking some underlying difference in meaning between 'by mistake' and 'by accident' (though see the warnings about difficulties involved in interpreting the data mentioned above in §5.2).

6 Conclusion

The project of dissolving traditional philosophical debates by paying close attention to the ordinary use of philosophically significant expressions has come under withering criticism in the 60 or so years since its heyday. But the experimental approach to the study of meaning employed by Austin, whether conducted informally or formally, remains the dominant methodology in both philosophy and linguistics. Contemporary experimental studies in semantics and pragmatics employ more sophisticated, controlled versions of what are essentially Austin's methods, in which participants are asked to imagine some situations and data is collected regarding what they say about those situations. Hypotheses about the meaning of certain expressions can be confirmed or disconfirmed based on what participants say in response to the situations presented in experimental conditions.[24]

Austin says that he's interested in experimental linguistic data, with the aim of coming to a better understanding of subtle distinctions in the meaning of philosophically significant expressions. Looking in detail at the methods that he and many others use to gather that data reveals that while some experiments generate robust results independently of possible biases (cf. the absence of order effects in the donkey stories), some features of the design of these experiments (such as the presence of a gloss in Austin's go-cart and racing-car experiments) actually obscure the phenomena they aim to uncover. Our aim in this paper has been to focus attention on the experimental methods Austin (and many others) employ to investigate fine distinctions in meaning and indicate what features of those experiments could be problematic and where those methods can be made more rigorous.

Nat Hansen
University of Reading
Reading RG6 6AA
U.K.
n.d.hansen@reading.ac.uk

Emmanuel Chemla
École Normale Supérieure
Laboratoire de Sciences Cognitives et Psycholinguistique

[24] Debates between Austin and 'empirical semanticists' in the 1950s that foreshadow some of the concerns raised in this paper are discussed in Chapman (2011) and Murphy (2015).

Paris 75005
France
chemla@ens.fr

References

Austin, J. (1956–1957). A plea for excuses. *Proceedings of the Aristotelian Society 57*, 1–30.
Austin, J. (1958). Pretending. *Proceedings of the Aristotelian Society 32*, 261–294.
Austin, J. (1966). Three ways of spilling ink. *The Philosophical Review 75*(4), 427–440.
Bates, S. and T. Cohen (1972). More on what we say. *Metaphilosophy 3*(1), 1–24.
Baz, A. (2012). *When Words Are Called For: A Defense of Ordinary Language Philosophy*. Cambridge, Massachusetts: Harvard University Press.
Cavell, S. (1958). Must we mean what we say? *Inquiry 1*, 172–212.
Cavell, S. (1965). Austin at criticism. *The Philosophical Review 74*(2), 204–219.
Cavell, S. (1994). *A Pitch of Philosophy: Autobiographical Exercises*. Cambridge, MA: Harvard University Press.
Chapman, S. (2011). Arne naess and empirical semantics. *Inquiry 54*(1), 18–30.
Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology 1*(2), 275–296.
Doyen, S., O. Klein, C.-L. Pichon, and A. Cleeremans (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE 7*(1), 1–7.
Egre, P. (2014). Intentional action and the semantics of gradable expressions (on the knobe effect). In B. Copley and F. Martin (Eds.), *Causation in Grammatical Structures*, pp. 176–204. Oxford University Press.
Ferguson, K. (2003). Three ways of spilling blood. In F. Debrix (Ed.), *Language, Agency, and Politics in a Constructed World*, pp. 87–100. Armonk, New York: M.E. Sharpe.
Fodor, J. A. and J. J. Katz (1963). The availability of what we say. *The Philosophical Review 72*(1), 57–71.
Friedman, P. (1969). The availability of ordinary-language philosophy. *Man and World 2*(3), 410–422.
Gustafsson, M. (2005). Perfect pitch and austinian examples: Cavell, mcdowell, wittgenstein, and the philosophical significance of ordinary language. *Inquiry 48*(4), 356–389.
Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.
Hanfling, O. (2000). *Philosophy and Ordinary Language: The Bent and Genius of our Tongue*. London: Routledge.
Hansen, N. (2013). A slugfest of intuitions: Contextualism and experimental design. *Synthese 190*(10), 1771–1792.
Hansen, N. (2014a). Contemporary ordinary language philosophy. *Philosophy Compass 9*(8), 556–569.
Hansen, N. (2014b). Contrasting cases. In J. Beebe (Ed.), *Advances in Experimental Epistemology*, pp. 71–95. New York: Bloomsbury.
Hansen, N. and E. Chemla (2013). Experimenting on contextualism. *Mind & Lan-*

guage 28(3), 286–321.
Henson, R. G. (1965). What we say. *American Philosophical Quarterly 2*(1), 52–62.
Intons-Peterson, M. J. (1983). Imagery paradigms: How vulnerable are they to experimenters' expectations. *Journal of Experimental Psychology: Human Perception and Performance 9*(3), 394–412.
Jackman, H. (2001). Ordinary language, conventionalism and *a priori* knowledge. *dialectica 55*(4), 315–325.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies 130*(2), 203–231.
Livingston, P. M. (2012). Review of john roberts, *The Necessity of Errors*. *Notre Dame Philosophical Reviews 2012.06.43*.
Mates, B. (1958). On the verification of statements about ordinary language. *Inquiry 1*(1–4), 161–171.
Murphy, T. S. (2015). Experimental philosophy: 1935-1965. In T. Lombrozo, J. Knobe, and S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy*, Volume 1, pp. 325–. Oxford: Oxford University Press.
Pettit, D. and J. Knobe (2009). The pervasive impact of moral judgment. *Mind & Language 24*(5), 586–604.
Phelan, M. (2013). Evidence that stakes don't matter for evidence. *Philosophical Psychology* (iFirst), 1–25.
Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research* (Enlarged ed.). New York: Irvington Publishers, Inc.
Sandis, C. (2010). The experimental turn and ordinary language. *Essays in Philosophy 11*(2), 181–196.
Schwitzgebel, E. and F. Cushman (2012). Expertise in moral reasoning? order effects on moral judgments in professional philosophers and non-philosophers. *Mind & Language 27*(2), 135–153.
Searle, J. (2001). J.l. austin. In A. Martinich and D. Sosa (Eds.), *A Companion to Analytic Philosophy*, pp. 218–230. Oxford: Blackwell.
Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods 43*(1), 155–167.
Sprouse, J., C. T. Schütze, and D. Almeida (2013). A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua 134*, 219–248.
Strickland, B. and A. Suben (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology 3*(3), 457–467.
Williams, B. (2014). *Essays and Reviews: 1959–2002*. Princeton, NJ: Princeton University Press.
Williams, M. A. (2009). *Henry James and the Philosophical Novel: Believing and Seeing*. Cambridge: Cambridge University Press.