Oxford Handbooks Online

Experimental Philosophy of Language

Nathaniel Hansen

Subject: Philosophy, Philosophy of Language Online Publication Date: Nov 2015
DOI: 10.1093/oxfordhb/9780199935314.013.53

Abstract and Keywords

Experimental philosophy of language uses experimental methods developed in the cognitive sciences to investigate topics of interest to philosophers of language. This article describes the methodological background for the development of experimental approaches to topics in philosophy of language, distinguishes negative and positive projects in experimental philosophy of language, and evaluates recent experimental work that concerns the reference of proper names and natural kind terms. The reliability of expert linguistic judgments versus the linguistic judgments of ordinary speakers, the role that different forms of ambiguity play in influencing responses to experiments, and the reliability of metalinguistic judgments are also assessed.

Keywords: experimental philosophy, expertise, philosophy of language, meaning, reference, proper names, natural kind terms, psychological essentialism, intuitions, philosophical methodology

1 Introduction

Experimental philosophy of language applies experimental methods used in the cognitive sciences (experimental psychology, psycholinguistics) to topics of interest to philosophers of language, such as the meaning of particular kinds of expressions (names, determiners, natural kind terms, adjectives, and so on), pragmatic phenomena (implicature, presupposition, metaphor, the semantics-pragmatics boundary, for example), and methodological issues (the reliability of informal versus formal experimental methods, the reliability of expert judgments versus the judgments of ordinary speakers, for example).[1]

Experimental philosophy of language has become a topic of intense interest in the past decade, due to the rapid growth of experimental philosophy and the roughly contemporaneous “linguistics turn” in philosophy of language, which has involved much contemporary work in philosophy of language being informed by—or in some cases being

indistinguishable from—work in contemporary linguistics. Indeed, it is not always clear what distinguishes experimental philosophy of language from experimental work in linguistics. Sometimes the difference is sociological: The distinguishing feature of research in "experimental philosophy of language" might merely be that it is published in philosophy journals or is written by theorists who are employed in philosophy departments, rather than departments of linguistics or cognitive science. But there is also a more substantive difference: Work in experimental philosophy of language sometimes explicitly engages with traditional philosophical debates, such as investigations into the meaning of names or natural kind terms. This article will focus on the relevance of experimental philosophy of language for those debates.

2 Methodological Background

Bogen and Woodward (1988) distinguish the *data* that are generated in, and are specific to, experimental contexts from the underlying *phenomena* that the experiments are designed to investigate. Theories explain phenomena or facts about phenomena, which have "stable, repeatable characteristics which will be detectable by means of different procedures" (p. 317). Data are the observable results generated by experiments and are evidence for the existence of phenomena (p. 305). Examples of the kind of phenomena that have interested experimental philosophers of language are the following:

- The extension of "knows" can vary when certain features of context are varied.
- "Red" behaves more like "spotted" than like "tall" in certain entailment patterns.
- "Gödel" refers to Gödel, not whoever uniquely or best satisfies descriptions associated with the name.
- A speaker can have completely false beliefs about a natural kind like *gold* and yet still refer to gold with the natural kind term "gold."

Experimental philosophy of language investigates these phenomena by gathering experimental data. Data can include metalinguistic judgments or "intuitions" (about the acceptability of sentences, or the truth value of what is said by a use of a sentence, or whether one sentence entails another, for example), non-metalinguistic actions (how someone responds to a request to "hand me the blue one," for example), eye movements, reaction times, brain activity, and so on (Krifka, 2011).[2]

As the list of types of data discussed earlier indicates, it is a standard methodological assumption guiding experimental philosophy of language that the *behavior of speakers* is data that provides evidence for linguistic phenomena. ("Behavior" is here understood

broadly—that is, not *behavioristically*—to include truth-value judgments, patterns of inference, reaction times, eye movements, brain activity, and so on.) This methodological assumption has been resisted by some critics of experimental philosophy of language, who argue that facts about speakers' behavior "are data for a psychological theory," but are not data for theories about linguistic phenomena (like the reference of proper names) (Deutsch, 2009, p. 449). The criticism helps to make explicit the fact that experimental philosophers of language do assume that linguistic phenomena have psychological consequences, which are detectable in behavior.[3]

Negative and Positive Research Programs in Experimental Philosophy of Language:

Alexander et al. (2010) distinguish *negative* and *positive* research programs in experimental philosophy. The negative research program in experimental philosophy of language, epitomized by Machery et al. (2004) and Mallon et al. (2009), critiques what they consider to be a "widespread" methodology in the philosophy of language. That methodology involves testing the predictions of theories (paradigmatically, about the way the reference of proper names is determined) against speakers' intuitions about actual or hypothetical examples. Mallon et al. (2009, p. 338) call this methodology "the method of cases":

> *The method of cases*: The correct theory of reference for a class of terms T is the theory which is best supported by the intuitions competent users of T have about the reference of members of T across actual and possible cases.

According to advocates of the negative program, philosophers have naively assumed that their own intuitions are representative of the intuitions of all competent speakers. If it turns out that their intuitions are not representative, then there is reason to doubt whether the theory that was based on those nonrepresentative intuitions is correct. And if there is widespread variation in intuitions, then there may be reason to wonder whether there will be a single "correct" theory of reference that is best supported by the intuitions of competent speakers. Mallon et al. (2009, p. 342) seem to suggest that the right response to such variation would be to give up on the project of developing substantive theories of reference altogether.

But whether the "method of cases" is genuinely widespread is contested by Deutsch (2009) and Ludwig (2007). Deutsch (2009) rejects the idea that philosophers use intuitions as evidence for theories of reference.[4] Ludwig (2007, 2010) also denies that intuitions are evidence for theories of reference—he holds that they are expressions of "conceptual competence," which involves the *correct* application of concepts. Variation in intuitions is only evidence that something is interfering with conceptual competence. A less radical challenge is posed by Devitt (2011, 2012), who argues that intuitions can be

evidence for theories of reference, but the intuitions of experts are better evidence than the intuitions of ordinary speakers. Variation in intuitions about reference is then only a problem if it is the intuitions of experts that vary. This objection will be discussed in detail in section 3.2.2.

The *positive* program in experimental philosophy of language is less concerned with challenges to traditional philosophical methodology than with piecemeal investigations of particular linguistic phenomena like the meaning of determiners, gradable adjectives, scalar implicature, and so on. The branch of the positive program that will be discussed in this essay concerns the meaning of *natural kind terms* ("gold," "cat," "lemon," and so on). The positive program accepts that there can be variation in intuitions, but aims to explain that variation in terms of some shared cognitive mechanism(s). For example, the theories of natural kind terms offered by Braisby et al. (1996) and Nichols et al. (2015) (discussed in section 4.2) involve variations on the idea that natural kind terms are systematically ambiguous. Variation in intuitions about natural kind terms can then potentially be explained as the result of different ways of resolving the relevant ambiguity.

3 Reference and Proper Names

Ground zero of the explosion of recent interest in experimental philosophy of language concerns how *proper names* refer to the objects they name. The intense debate surrounding the experimental investigations of names and reference (sparked by Machery et al., 2004) is partly explained by the fact that proper names have been at the center of philosophical debates about language since the rise of analytic philosophy in the early twentieth century. But the intensity of the debate is also at least partly to do with the radical and contentious conclusions about philosophical methodology in general that have been drawn by experimental philosophers on the basis of data involving judgments about the reference of proper names.

3.1 "Cross-Cultural Semantics"

Inspired by studies in cultural psychology that provide some support for the idea that cultural differences between Westerners and East Asians affect individuals' "naive metaphysical systems," "tacit epistemologies," and the "nature of their cognitive processes" (Nisbett et al., 2001, p. 291), Machery et al. (2004) conducted an experiment

that they argue yields evidence that Westerners and East Asians have systematically different intuitions about the reference of proper names.

The experiment conducted by Machery et al. is designed to test a hypothesis about "descriptivist" and "causal-historical" views of reference. As described by Machery et al. (p. 2), a descriptivist view of the reference of proper names consists of two components:

> **D1** Competent speakers associate a description with every proper name. This description specifies a set of properties.
> **D2** An object is the referent of a proper name if and only if it uniquely or best satisfies the description associated with it.… If the description is not satisfied at all or if many individuals satisfy it, the name does not refer.

A "causal-historical" view of the reference of proper names, in contrast, involves two different components that conflict with the descriptivist view (p. 3):

> **C1** A name is introduced into a linguistic community for the purpose of referring to an individual. It continues to refer to that individual as long as its uses are linked to the individual via a causal chain of successive users: every user of the name acquired it from another user, who acquired it in turn from someone else, and so on, up to the first user who introduced the name to refer to a specific individual.
> **C2** Speakers may associate descriptions with names. After a name is introduced, the associated description does not play any role in the fixation of the referent. The referent may entirely fail to satisfy the description.

An important test case for evaluating descriptivist versus causal-historical views would be one in which a speaker associates a description with a proper name that isn't satisfied by the individual to whom the name was originally applied and who is linked to current uses of the name by way of a causal chain. If, in that case, the name is taken to refer to whoever satisfies the description (or to no one, if nothing satisfies the description), rather than the person to whom the name was originally applied and who is linked to current uses of the name by a causal chain, that supports descriptivism and presents a challenge to the causal-historical view. If the name is *not* taken to refer to whoever satisfies the description (or is not taken to fail to refer, if no one satisfies it), but rather is taken to refer to the person to whom the name was originally applied and who is linked to current uses of the name in the right way, that supports the causal-historical view and challenges the descriptivist view.

Machery et al. created two types of story that were intended to present Western and East Asian experimental participants with these crucial test cases. The first type of story was modeled on Kripke's (1980) "Gödel" case, in which the person causally linked with the introduction of the name "Gödel" is associated with the description "the discoverer of the

incompleteness theorem," which he does not satisfy in the story, but which is uniquely satisfied by another person, Schmidt. In Kripke's story, Schmidt dies under mysterious circumstances, and Gödel takes credit for his work. The second type of story was modeled on Kripke's (1980) "Jonah" case, which involves a person causally linked with the introduction of the name "Jonah," but who does not satisfy any of the descriptions associated with the name, and which are not satisfied by any other referents either. For both the Gödel-type and Jonah-type stories, Machery et al. created a "Western" version, with Western names ("Gödel," for example), and an "East Asian" version, with East Asian names ("Tsu Ch'ung Chih," for example). In the Gödel case, participants were asked to respond to the following prompt (p. 6):

When John [the speaker in the story] uses the name "Gödel," is he talking about:

(A) the person who really discovered the incompleteness of arithmetic? or

(B) the person who got hold of the manuscript and claimed credit for the work?

Though they found no statistically significant cross-cultural difference in responses to the Jonah-type stories, Machery et al. did find that there was a statistically significant difference between responses given by Western participants to the Gödel-type stories and those given to those stories by East Asian participants: East Asians responses tended to be descriptivist ((A)-type responses), while Western responses tended to be "Kripkean" (causal-historical) ((B)-type responses) (see Table 1).[5]

Table 1: Percentage of "Kripkean" (B-type) responses to the Gödel-type stories in Machery et al. (2004); reported as percentages in Machery (2012)

	Westerners	**East Asians**
Gödel story	58%	29%
Tsu Ch'ung Chih story	55%	32%

The significant variation between the responses of Western and East Asian participants to the two versions of the Gödel story is the key finding of the study, but Machery et al. also found substantial variation *within* each cultural group as well.

Machery et al. argue that the variation they observe undermines philosophers' use of their own intuitions as evidence of what the correct theory of reference for proper names is. In Mallon et al. (2009), the same theorists argue that many theories in different areas of philosophy rely on what they call "arguments from reference," which presuppose

either a descriptivist or causal-historical view of the reference of certain expressions. Given their argument that variation in intuitions about the reference of proper names problematizes the idea that philosophers have evidence that there is a single, correct reference for proper names, they conclude that there is reason to be skeptical of any philosophical argument that relies on an argument from reference.

3.2 Criticisms of Machery et al. (2004) and Replies

Criticisms of Machery et al. (2004) can be classified as either (i) objections to the experimental design employed in the study, or (ii) objections to the philosophical significance of the data, even if it has been collected appropriately.[6]

3.2.1 Objections to the Experimental Design

Metalinguistic intuitions vs. use:

The prompt employed in Machery et al.'s (2004) Gödel stories asks participants in the experiment to decide who John [the speaker in the story] is talking about when he "uses the name 'Gödel.'" Martí (2009) refers to this type of task as "metalinguistic," and she contrasts metalinguistic intuitions about the reference of names with how we *use* names to refer (p. 44):

> [Machery et al.] test people's intuitions about *theories* of reference, not about the *use* of names. But what we think the correct theory of reference determination is, and how we use names to talk about things are two very different issues.

Because Machery et al. elicit metalinguistic intuitions, Martí argues, the current "experiment does not provide any evidence at all about name use" (p. 46).[7] And we already knew, from the history of debates between advocates of descriptivist and antidescriptivist views, that there is variation in intuitions about what the right theory of reference is for proper names (p. 45). We don't need experiments to tell us that.

As Ichikawa et al. (2012, n. 3) observe, Martí's criticism "overshoots." There is not a sharp separation between linguistic use and responses to metalinguistic tasks. Regarding a parallel debate in linguistics over the relative merits of metalinguistic intuitions about the acceptability of sentences versus the use of those sentences in conversation as evidence of theories of syntax, Schütze (1996, pp. 81–82) observes that there are two "extreme positions" one could take about the relation between use and metalinguistic intuition: The first position is that there is no difference between the two. The second extreme position is that use and metalinguistic intuition are "entirely separate and might differ in arbitrary ways."[8] A much more plausible intermediate view (defended at length

in Schütze, 1996) is that metalinguistic intuitions are shaped by the same linguistic competence that shapes production in ordinary conversation, plus a range of other, experiment-specific factors. On the intermediate view, metalinguistic intuitions are a source of evidence of the underlying linguistic competence (though it is important to control for a range of possible interfering factors).

Machery et al. (2009) attempt to respond to Martí's challenge with an experiment that asks participants to read a version of the Gödel story (involving the Chinese astronomer Tsu Ch'ung Chih), and varies whether participants read a "metalinguistic" prompt (identical to the prompt used in Machery et al., 2004), or a "linguistic" prompt that asks participants to make a truth value judgment about what is claimed by the speaker in the story (note, however, that the "linguistic" prompt is still metalinguistic: It asks participants to make a truth value judgment about a *claim*!) (p. 690):

> when Ivy says, "Tsu Ch'ung Chih was a great astronomer," do you think that her claim is: (A) true or (B) false?

Reactions were classified as "Kripkean" ((B)-type responses in either the linguistic or metalinguistic condition) or "non-Kripkean" ((A)-type responses in either condition). Machery et al. (2009) found no difference in the responses of participants to the metalinguistic and linguistic prompts. But given that the "linguistic" prompt is also metalinguistic, this result doesn't address the question of whether metalinguistic intuitions and "linguistic intuitions" are "largely congruent" (as Machery et al. claim it does).[9]

It is true that speakers can have false beliefs about how they actually use language (see Labov, 1996), and those false beliefs can influence their metalinguistic intuitions. But absent additional evidence, there is little reason to think that metalinguistic judgments about the reference of proper names diverge substantially from the way speakers use names to refer.

Semantic reference versus speaker reference ambiguity:

Deutsch (2009) argues that it is impossible to conclude from Machery et al.'s (2004) results that there is cross-cultural or intra-cultural variation in intuitions about the reference of proper names, because there is a crucial ambiguity in the prompt that could be affecting participants' responses.[10] Deutsch argues that when participants in Machery et al.'s (2004) experiment are asked to choose who John is "talking about" with his uses of "Gödel," they might reasonably interpret that question in either of the following ways (p. 454):

> (Q1) To whom does *John intend to refer* when he uses "Gödel"?

(Q2) To whom does *the name, "Gödel,"* refer when John uses it?

Q1 interprets the prompt as asking about something pragmatic, namely *speaker reference* (who speakers intend the name to refer to). Q2 interprets the prompt as asking about *semantic reference* (who the name refers to in virtue of the conventions of the language).[11] Machery et al.'s conclusions about cross-cultural and intra-cultural variation in intuitions about the reference of names presuppose that participants are interpreting the prompt as Q2, but Q1 seems like an equally reasonable interpretation.[12] It is impossible to tell, based on the results reported in Machery et al. (2004), how much of the observed variation in responses is due to different interpretations of the prompt and how much is due to actual variation in referential intuitions.

But revised versions of Machery et al.'s experiment that attempt to eliminate or control for the ambiguity replicate the earlier findings of variation. In Machery et al. (2015), Chinese and American participants read the "Gödel" story used in Machery et al. (2004), but with a modified prompt designed to exclude the speaker's reference interpretation ("When John uses the name "Gödel," regardless of who he might intend to be talking about, he is actually talking about…"). Using the revised version, they found significant variation between Chinese and American participants, with Chinese participants again tending to give responses consistent with a descriptivist theory of reference for proper names.

Ambiguity in epistemic perspective:

Sytsma and Livengood (2011) make the case that Machery et al.'s (2004) Gödel stories involve an ambiguity of "epistemic perspective." The only thing that the character in the Gödel story, John, is said to have heard about Gödel is that he is the discoverer of the incompleteness of arithmetic, while the narrator of the story has much more information about Gödel and Schmidt: namely, that Schmidt actually did the work in question, and that Gödel took credit for it. When participants are prompted with the question:

When John uses the name "Gödel," is he talking about:

(A) the person who really discovered the incompleteness of arithmetic? or

(B) the person who got hold of the manuscript and claimed credit for the work?

From whose epistemic perspective should they attempt to answer the question: John's, or the narrator's? If participants take up John's epistemic perspective in responding to the prompt, limiting themselves to the information that he has—that Gödel is the discoverer of the incompleteness of arithmetic—then they could conceivably choose the apparently descriptivist option (A). Crucially, participants could choose (A) even if from their *own*

epistemic perspective (that of the narrator) they would opt for the causal-historical option (B).

Sytsma and Livengood hypothesized that they could significantly lower the percentage of causal-historical ((B)-type) responses by modifying the prompt to emphasize John's epistemic perspective, and that they could significantly raise the percentage of (B) responses by modifying the prompt to emphasize the narrator's epistemic perspective. They demonstrated this in their first experiment, shifting (B)-type responses from the level of the original Machery et al. prompt (39.4%) down to 22% when the John's-perspective prompt was used, and up to 57.4% when the narrator's-perspective prompt was used (p. 322). A second experiment that used a prompt that made the narrator's perspective even more prominent raised the percentage of (B)-type responses to 73.8% among non-philosophers (p. 325). A third, within-subjects experiment revealed that each version of the prompt significantly affected participants' responses to the Gödel story (see Table 2).

Table 2: Percentage of (B)-type (causal-historical) responses to different prompts in a within-subjects experiment (Sytsma & Livengood, 2011, p. 325)

Original	**John's perspective**	**narrator's perspective**	**clarified narrator's perspective**
42.9%	31.4%	57.1%	74.3%

A fourth experiment asked participants to respond to the original Machery et al. prompt, and then choose one of two restatements of the prompt that best corresponded with how they understood the question. One restatement emphasized John's perspective, while one emphasized the narrator's perspective. Participants who gave (B)-type responses tended to choose the narrator's perspective restatement, while participants who chose (A)-type responses tended to choose the John's-perspective restatement (p. 327).

These patterns of responses are consistent with the existence of an ambiguity in Machery et al.'s original prompt. Sytsma and Livengood argue that the appearance of variation in semantic intuitions in the original study can be explained instead by participants disambiguating the prompt in different ways. That undermines Machery et al.'s conclusion that their results are evidence of variation in semantic intuitions, because it appears that facts about epistemic perspective, rather than intuitions about the semantic reference of "Gödel," could be driving participants' responses.

While Sytsma and Livengood's experiments do pose a serious challenge to Machery et al. (2004), further experimental work has uncovered statistically significant cross-cultural

variation in responses to versions of the Gödel story, even while employing Sytsma and Livengood's "clarified narrator's perspective" prompt (Beebe and Undercoffer, 2015; Sytsma et al., 2015).

Summary:

There are good reasons to question aspects of the experimental design used in Machery et al. (2004). However, experiments that attempt to disambiguate speaker's reference and semantic reference (Machery et al., 2015), and disambiguate relevant epistemic perspectives (Beebe and Undercoffer, 2015; Machery et al., 2015; Sytsma et al., 2014) replicate the findings of cross-cultural and intra-cultural variation in participants' responses to Gödel-type stories. If the results of these recent experiments hold up, then this would appear to be a case in which an imperfectly designed experiment uncovers evidence of a phenomenon that is confirmed by more carefully designed experiments.

3.2.2 An Objection to the Philosophical Significance of the Data: Expertise

Devitt (2011, 2012) argues against Machery et al. on the grounds that ordinary speakers are not expert judges of reference. He maintains that while the intuitions of ordinary speakers provide some evidence of facts about the reference of proper names, the intuitions of experts (linguists and philosophers of language) provide *better* evidence. We should expect variation between the intuitions of experts and laypeople, and the existence of variation among nonexperts does not pose a problem for the use of expert intuitions as evidence for theories of reference.

What Devitt calls "linguistic intuitions" are what Martí calls "metalinguistic intuitions" (discussed in section 3.2): "fairly immediate unreflective judgments about the syntactic and semantic properties of linguistic expressions" (p. 482).[13] Devitt's view is that linguistic intuitions are "theory-laden empirical opinions" or "empirical unreflective judgments" (p. 488). That is, judgments about whether a sentence is true or false, or whether an expression refers to one object or another, are akin to judgments about whether some animal is an echidna or some white thing sticking out of the ground is a pig's jawbone or not. Just as one would be warranted in treating a judgment that an animal is an echidna as evidence that the animal is indeed an echidna to the extent that the judge is an expert on echidnas, one is warranted in treating linguistic judgments as evidence that what they represent is true to the extent that the judge is an expert on linguistic matters. And Devitt says that, when it comes to linguistic intuitions, it is linguists (and philosophers of language) who are the most expert (pp. 499–500).[14]

Experimenting on the "expertise defense":

The idea that the intuitions of philosophers and linguists are more reliable indicators of linguistic facts than those of ordinary speakers has been scrutinized by both experimental

philosophers and experimental linguists.[15] Culbertson and Gross (2009) ran an experiment to test Devitt's claim that linguists have more reliable judgments about syntactic phenomena than ordinary speakers. Because the syntactic facts are disputed, Culbertson and Gross use intragroup *consistency* of judgments as a measure of reliability (intragroup consistency is a necessary but not sufficient condition for the reliability of the judgments of that group). Participants in Culbertson and Gross's experiment were categorized into four different groups, based on their experience with theoretical syntax and cognitive science: Ph.Ds in linguistics ("LOTS"), students with at least one class in generative syntax ("SOME"), students with no experience in syntax but experience in other areas of cognitive science ("LITTLE"), and students with no experience of cognitive science ("NONE"). Participants rated the acceptability of 73 sentences taken from an introductory linguistics textbook.

The experiment revealed that while all of the groups with at least some exposure to cognitive science (LOTS, SOME, and LITTLE) all "showed equally high intra-group average correlation values," participants in the group with no exposure to cognitive science were not well correlated with one another, and "the average correlation was significantly lower than the average correlation of the other three groups" (pp. 729–730). Furthermore, LOTS, SOME, and LITTLE were "highly correlated with one another, and more correlated with each other than the NONE group" (pp. 731–732). Culbertson and Gross interpret this result as indicating that once participants acquire a minimum degree of task-specific knowledge (familiarity with the acceptability task in this case), further expertise does not affect the reliability of their syntactic intuitions.[16]

In a separate evaluation of the "expertise defense," Machery (2012) gave the Tsu Ch'ung Chih version of the Gödel story to a variety of professional linguists, philosophers, and nonexpert holders of Ph.Ds. Machery formed two groups of specialists who he thought were likely to have expertise relevant to the reference of proper names: philosophers of language and semanticists (Group 1) and researchers in discourse analysis, historical linguistics, and sociolinguistics (Group 2). The central result of this experiment is that Group 1 has a significantly higher proportion of Kripkean intuitions than Group 2 (p. 48), and the proportion of Kripkean responses among nonexpert participants (Group 3) is intermediate between Group 1 and Group 2 (see Table 3).

Table 3: Percentage of “Kripkean” responses among semanticists and philosophers of language (Group 1), discourse analysts, historical linguists, and sociolinguists (Group 2) and nonexperts with Ph.Ds (Group 3) (Machery, 2012, p. 49)

Group 1	**Group 2**	**Group 3**
86.4%	68.7%	76.9%

Machery interprets these results as showing that expertise has an inconsistent effect on intuitions about reference: Sometimes it correlates with a greater degree of Kripkean intuitions than nonexperts (Group 1), but other times it correlates with a lesser degree of Kripkean intuitions than nonexperts (Group 2). Machery takes this inconsistent effect of expertise to undermine the expertise defense (p. 50), which should predict a uniform effect of expertise on intuitions (assuming that there is only a single type of linguistic expertise).[17]

Sprouse et al. (2013) gather a large set of data relevant to debates about the relative reliability of the judgments of ordinary speakers and linguistic experts. The theoretical background to Sprouse et al.’s study is a dispute in syntactic theory regarding the reliability of formal versus informal methods of collecting acceptability judgments, which parallels the debate over the reliability of expert versus ordinary speaker intuitions in philosophy.[18] The informal method of collecting acceptability judgments is what philosophers call an “armchair” method: it tends to involve small numbers of expert participants (sometimes just the theorist herself), it usually does not involve statistical tests of significance, and it usually does not control for known sources of bias (order of presentation bias, experimenter bias, and so on). Formal methods of collecting acceptability judgments “tend to involve substantially more participants, substantially more tokens per condition, substantially more response options, relatively naïve non-linguist participants, substantially more instructions, and substantially more statistical analyses” (p. 224).

Sprouse et al. collected a random sample of 300 sentence types used in informally collected acceptability judgments in the journal *Linguistic Inquiry* (“a leading theoretical journal among generative syntacticians … [intended] to stand as a proxy for the use of informal methods in syntax more broadly”) (p. 222) from 2001 to 2010. The 300 sentences consisted of 150 unacceptable sentence types and 150 more acceptable controls, forming 150 “pairwise phenomena” (p. 223) like the following:

> ?? Ginny remembered to have bought the beer.
>
> Ginny remembered to bring the beer. (p. 237)

The results of the experiment (936 participants distributed across three versions of the experiment evaluating different judgment tasks) indicate a 95% convergence rate between informal and formal methods of collecting acceptability judgments, with a margin of error of 5.3–5.8% (p. 230).

These results indicate that the choice of formal versus informal methods of gathering data in syntactic theory does not have an overwhelming effect on the empirical content of the data. While it doesn't bear directly on debates about armchair versus experimental methods of gathering data about theories of reference, Sprouse et al.'s experiment does suggest that in the vast majority of cases, linguistic expertise does not play a crucial role in shaping intuitions (which accords with Culbertson and Gross's findings).

But Sprouse et al.'s results also suggest that it would be a mistake to put too much evidential weight on intuitions that concern only a very small sample of nonrandomly chosen expressions ("Gödel" and "Tsu Ch'ung Chih"). That lends additional weight to a criticism of Machery et al. raised by Devitt (2011), Deutsch (2015), and Ichikawa et al. (2012), that the "Gödel" story is only one example of many that Kripke presents as counterexamples to descriptivism about proper names. Pro-descriptivist intuitions about a single example do not demonstrate that a speaker has pro-descriptivist intuitions in general. This indicates the need for experiments (or corpus studies) that consider how experts and ordinary speakers react to a larger sample of proper names.

4 Natural Kind Terms

4.1 Theoretical Background

There is substantial overlap between the experimental investigation of natural kind terms and proper names because both involve a dispute between descriptivist and nondescriptivist theories of reference. Putnam (1975a, p. 140) and Kripke (1980) critique what Putnam calls "the traditional view" of the meaning of natural kind terms like "gold," "lemon," and "tiger."[19] The traditional view of the meaning of these terms consists of the following components:

1. The meaning of "cat" (for example) is a conjunction of properties (*animal, carnivorous, has four legs*, and so on).
2. For each property P associated with "cat" (for example), "cats have property P" is an analytic truth (necessarily true, knowable a priori).

3. "Anything with all of the properties associated with cats is a cat" is also an analytic truth (necessarily true, knowable a priori).

Both Kripke and Putnam point out that there are counterexamples to the traditional theory. Consider the expression "tiger," and the conjunction of properties "large carnivorous quadrepedal feline, tawny yellow in color with blackish transverse stripes and white belly" (Kripke, 1980, p. 119). It seems very implausible to think that "tiger" wouldn't apply to a creature that shared all other properties with tigers but that had only three legs. But that is entailed by the traditional view. So the traditional view can't be right about the meaning of terms like "tiger."

A more sophisticated version of the traditional view introduces the idea of a "cluster" of properties associated with a natural kind term. If an object satisfies enough of the properties (some of which might be weighted differently), then it would count as a member of the relevant kind. The cluster view would avoid the obvious counterexamples that make the traditional descriptivist view unacceptable. But Kripke (1980) argues that even the cluster version of descriptivism is untenable. He argues that (1) the cluster view doesn't provide sufficient conditions for an object to belong to a natural kind—even if an object possesses *all* of the properties associated with "gold" or "tiger" it is possible that it wouldn't count as gold or a tiger (p. 120); and (2) that it isn't necessary to satisfy *any* of the properties associated with the kind term for an object to count as a member of the relevant kind (p. 121).

On Kripke's antidescriptivist, causal-historical view of natural kind terms, they refer to "the essence" of the relevant kind (p. 138), and they do so *not* by way of description. Speakers' beliefs about the kind can be completely incorrect, and yet the term will still refer, as long as it is linked in the right way to an original use by way of a causal-historical chain of use.

4.2 Experimenting on Natural Kind Terms

4.2.1 Braisby et al. (1996)

Braisby et al. (1996) ran two experiments to evaluate what they call the "essentialist view" (which they attribute to Kripke [1980] and Putnam [1975b]) of natural kind terms, which they characterize as follows (p. 248):

1. Essential properties determine reference.
2. Nonessential (or contingent) properties do not determine reference.
3. Reference is determined independently of people's beliefs about which properties determine reference.

For essentialism to yield predictions that can be empirically tested, it must be assumed that speakers implicitly believe (1–3) and that those beliefs will be manifested in their linguistic behavior (this turns essentialism into *psychological* essentialism).

Braisby et al. ran two experiments to test (psychological) essentialist predictions like the following:

> If beliefs about an essential property of a kind turn out to be false (cats turn out to be robots, not mammals, for example), speakers will still apply the kind term ("cat") to the same objects.

Braisby et al. found significant divergences from (psychological) essentialist predictions. For example, only 58% of participants in one experiment, and 76% in another, responded to a story based on Putnam's (1975a) Martian robot cat thought experiment in accordance with (psychological) essentialist predictions.

Braisby et al.'s experiments also indicated a tendency among some participants to respond to their stories with prima facie contradictions. So, for example, in response to stories modeled on Putnam's Martian robot cat thought experiment, 31% of participants in the first experiment, and 15% in the second, assigned statements like the following the same truth value:

> (+) Tibby is a cat, though we were wrong about her being a mammal.
>
> (–) Tibby is not a cat, though she is a robot controlled from Mars.

Neither essentialism nor the cluster theory predicts this pattern of apparently contradictory responses. Braisby et al. propose that a "representational change" theory, which holds that natural kind terms can have different senses (and extensions) in different contexts, can explain this puzzling pattern of data. Roughly, such a theory holds that speakers employ both descriptivist and a nondescriptivist ("particularist") interpretations of natural kind terms, and seemingly contradictory responses would be explained in terms of speakers switching between the two interpretations.

4.2.2 Jylkkä et al. (2009)

Jylkkä et al. (2009) aim to compare the plausibility of descriptivism, (psychological) essentialism, and Braisby et al.'s "representational change theory" (essentially a form of ambiguity theory). Participants were asked to respond to complex scenarios that involved a substance X that shares superficial properties with substance Y, and is believed by experts to share a deep structure with Y. Participants were first asked to judge whether substance X is substance Y. Then participants were then told that later discoveries showed that substance X in fact does *not* share a deep structure with Y. Participants were

asked whether their earlier judgment “X is Y” or “X is not Y” was (a) justified and (b) strictly speaking correct.

Jylkkä et al. (2009) found that participants tended to give answers to the first question that were compatible with essentialism: If X and Y shared the same deep structure, they tended to say that X is Y; if X and Y did not share the same deep structure, they tended to say that X is not Y.[20] That result undermines a traditional descriptivist position, but does not distinguish between essentialism and the ambiguous representational change theory (or indeed a cluster theory: see Haggqvist and Wikforss, 2015). In response to part (b) of the second question, which asked whether their earlier judgments were *strictly speaking correct*, 69% of responses were compatible with essentialism (and an ambiguity theory and a cluster theory), 28% were compatible with an ambiguity theory or a cluster theory, but not essentialism, and 3% were in the middle of the response scale and labeled as *compromises*.

A second experiment offered an explicitly ambiguous response option (“on the one hand yes, on the other hand no”) in order to more directly probe the ambiguity theory. As in the first experiment, participants were asked to say whether the earlier judgment that X and Y were the same substance is correct or not when it turns out that X and Y do not share the same deep structure (only superficially similar properties). According to Jylkkä et al. (2009), essentialism predicts “not correct” answers, descriptivism predicts “correct” answers, and representational change theory predicts explicitly ambiguous answers. The results of the second experiment are shown in Table 4.

Table 4: Results of Jylkkä’s second experiment

Q: When X and Y turn out not to share deep structure, was it correct to judge that “X is Y”?	**No**	**Yes**	**On the one hand yes … on the other hand no**	**Can’t say**
	48%	22%	17%	12%

Surprisingly, Jylkkä et al. (2009) take these results to support essentialism on the grounds that essentialist answers were the most common response. But this pattern of responses in fact both presents a serious challenge to essentialism, because of the large number of responses that are incompatible with essentialist predictions, and it suggests that there is substantial interpersonal variation in how one responds to the scenario, which includes a minority response ("on the one hand yes … on the other hand no") that is most easily explained by a representational change theory.

4.2.3 Genone and Lombrozo (2012)

Genone and Lombrozo (2012) present evidence that they take to problematize both "pure" descriptivist and "pure" causal-historical theories of the reference of natural kind terms, and that they take to motivate a "hybrid" theory that incorporates both descriptive and causal-historical components. They constructed an experiment that asked participants to judge whether two speakers using an invented kind term ("tyleritis") for a disease were "having a thought about the same disease" (p. 725) in four conditions that varied the descriptive information the speakers associated with the relevant expression and the causal origin of the expression used by each speaker, so that in some conditions they matched and in some conditions they were different (see Table 5).

Table 5: Percentages of "yes" responses indicating co-reference in Genone and Lombrozo's Experiment 1

	Description	***Causal origin***	**"Yes" responses**
Part I	Different	Same	44%
Part II	Same	Same	98%
Part III	Same	Different	53%
Part IV	Different	Different	2%

Genone and Lombrozo found that when descriptive information and causal origin coincided (Part II), responses clearly indicated co-reference, and when both factors differed (Part IV), responses clearly indicated lack of co-reference.

The interesting results occur in Parts I and III, neither of which is significantly different from a 50% response (p. 726). Genone and Lombrozo found no significant correlation between participants' answers to Part I and their answers to Part II, indicating that the near-50% responses were not due to participants splitting into groups with "pure

descriptivist" and "pure causal" responses (if there were such groups, "yes" responses in Part I should be correlated with "no" responses in Part III). They conclude that these findings suggest "that most participants utilize both descriptive and causal information in making reference judgments" (p. 727).

A second experiment that included more specific information about the causal origin of the relevant information replicated the findings of Experiment 1: Parts I and III again showed intermediate results, and responses in those parts were not correlated with each other. Parts II and IV showed clear agreement and disagreement, also as in experiment 1.

Genone and Lombrozo argue that the possibility of a hybrid theory of reference that their evidence supports poses a challenge to the skeptical methodological conclusions that Machery et al. (2004) and Mallon et al. (2009) draw from the observation of inter-cultural and intra-cultural variation in intuitions about reference. The variation Machery et al. observe is consistent with all speakers sharing a hybrid theory of reference, but differing "in their preferred strategy for combining causal and descriptive information in making reference judgments" (p. 732).[21] Variation would then not be due to variation in whether speakers treat names as having their reference determined in descriptivist or causal-historical terms, but in the effects of as-yet-unspecified contextual factors on a shared—part descriptivist, part causal-historical—semantics.

4.2.4 Nichols et al. (2015)

If, as the experimental evidence from Genone and Lombrozo (2012) suggests, whether speakers employ a descriptivist or causal-historical interpretation of natural kind terms can vary depending on contextual factors, then it should be possible to construct experiments in which varying such contextual factors affects which interpretation speakers use. Nichols et al. (2015) construct such experiments and find evidence in support of what they call an "ambiguity" view of natural kind terms. According to the ambiguity view:

> in some cases, the reference of a token is fixed by a causal-historical convention; in other cases, the reference of a token of the same type is fixed by a descriptivist convention. (p. 8)

Nichols et al. (2015) claim to find support for the ambiguity view in four experiments. The experiments concern the "catoblepas," a mythical creature described as having scales on its back, a head like a bull, and a gaze that, if met, causes instant death. According to the story used in the experiment, researchers think that descriptions of the catoblepas were based on reports of encounters with wildebeests. In their first experiment, Nichols et al. asked participants whether catoblepas are more like rabbits ("really exist") or like goblins ("don't really exist"). Participants who read a neutral story about triceratops

before reading the catoblepas story tended to say that catoblepas did not really exist—a *prima facie* descriptivist response.[22] But participants who were primed with a story about triceratops that ascribed many false beliefs to earlier scientists (that triceratops was a bison, for example) while implying that reference was to the same species of animal throughout had significantly less descriptivist responses to the catoblepas story.

A second experiment asked participants to register their agreement with two statements after reading the catoblepas story used in the first experiment (minus the fact about the killer gaze):

1. "Catoblepas" refers to wildebeests.
2. Catoblepas exist.

Participants agreed with statement 1 to a greater degree than statement 2. Nichols et al. find this result puzzling, arguing that if "catoblepas" refers to wildebeests, then catoblepas exist. Nichols et al. suggest that the ambiguity theory offers a way of explaining the puzzling result if the "refers to" statement primes the causal-historical interpretation of "catoblepas" and the "existence" statement primes the descriptivist interpretation.[23]

In a third experiment, Nichols et al. found further puzzling patterns of agreement and disagreement to statements about catoblepas. Participants tended to agree with statement 1 and disagree with statement 2:

1. Catoblepas are wildebeests.
2. Wildebeests are catoblepas.

Nichols et al. assume that on the causal-historical view, "catoblepas" and "wildebeests" are co-referential, so 1 and 2 should be equivalent.[24] Furthermore, responses to 1 were significantly different from responses to statement 3:

3. Catoblepas exist.

Each of these results is prima facie puzzling on a (psychologized) causal-historical view about the reference of "catoblepas," which should predict that, given the information in the stories, participants should treat "catoblepas" and "wildebeests" as co-referential, and that if they take "catoblepas" to refer, then they should infer that catoblepas exist.

Nichols et al. replicated the difference between agreement with "Catoblepas are wildebeests" and "Catoblepas exist" even in a within-subjects follow-up experiment, when participants saw both statements side by side. They argue that the ambiguity view can explain this pattern of responses: When it appears in subject position, "catoblepas" carries an existence presupposition, and participants aim to accommodate the

presupposition, which requires adopting the causal-historical interpretation of "catoblepas." But in statements 2 and 3, when it occurs in predicate position and in an existence statement, respectively, there is no presupposition of existence, and participants are free to assign the descriptivist interpretation to "catoblepas."

A fourth (within-subjects) experiment that used a version of the catoblepas story intended to make causal-historical components more salient replicated the earlier findings that participants would both agree with "Catoblepas are wildebeests" and disagree with "Catoblepas exist," and also yielded responses to "Catoblepas are wildebeests" that were significantly above the midpoint (unlike earlier experiments). Like Braisby et al., Nichols et al. explain the apparent contradictoriness of these responses in terms of an ambiguity (or "semantic indecision" [Lewis, 1999] that can be resolved in different ways) in natural kind terms.

4.2.5 Summary

Several studies (Braisby et al., 1996; Genone and Lombrozo, 2012; and Nichols et al., 2015) have found support for either "hybrid" descriptive and causal-historical interpretations of natural kind terms, or versions of an ambiguity theory that also incorporates both interpretations.[25] Even Jylkkä et al (2009), which attempts to defend a form of psychological essentialism, present evidence that indicates there is a minority of participants whose responses are most easily explained by an ambiguity theory.

5 Conclusion

The negative program in experimental philosophy of language has provoked a great deal of debate. That is at least partly to do with the contentiousness of its characterization of the allegedly "widespread" philosophical methodology that assumes and relies on the fact that philosophers' intuitions are representative of those of all competent speakers. The critical attention given to the negative program tends to overshadow the less radical positive program.[26] The positive program in experimental philosophy does not set out to undermine traditional philosophical methodology, but to supplement it with experimental methods.

This article has only touched on one branch of the positive program, namely research on the reference of natural kind terms. But the positive program is wide-ranging, diverse, and developing rapidly. It includes investigations of whether moral considerations affect the interpretation of the determiner "many" (Cova and Egré, 2015), how speakers understand the determiner "most" (Pietroski et al., 2009), and the role that world

knowledge plays in the interpretation of “donkey sentences” like “Every farmer who owns a donkey beats it” (Geurts, 2002). Philosophers have used experimental methods to investigate the understanding of vague terms (Raffman, 2014; Ripley, 2015), epistemic modals (Knobe and Yalcin, 2014), and whether (and to what extent) the context of use affects the extension of “knows” (Buckwalter, 2010; Buckwalter and Schaffer, 2013; Hansen and Chemla, 2013), color adjectives (Hansen and Chemla, 2013, 2014), and aesthetic adjectives (Liao and Meskin, 2015).[27]

The boundary between semantics and pragmatics has been investigated by looking at whether autistic speakers (who have significant pragmatic deficits) can understand “primary pragmatic processes” like quantifier domain restriction (de Villiers et al., 2007, 2012, 2013), whether minimal propositions play a role in linguistic understanding (Bezuidenhout and Cutting, 2002), and whether metaphors are any less paraphrasable than literal utterances (Phelan, 2010). Finally, philosophers pursuing the positive program in experimental philosophy of language have investigated scalar implicature (Geurts and Pouscoulous, 2009), a phenomenon that has received intense attention from linguists and psychologists.[28] The positive program will undoubtedly continue to expand both in terms of the range of topics that receive experimental treatment and in the sophistication with which those topics are investigated, advancing traditional debates and raising new questions for philosophers of language.

Acknowledgments

This paper was written with the support of British Academy grant SQ120050, “Quantitative Methods in Experimental Philosophy of Language.” Thanks to Zed Adams, Emma Borg, Daniel Cohnitz, James Genone, Sören Haggqvist, Richard Heck, Eliot Michaelson, Ángel Pinillos, Mark Pinder, and Åsa Wikforss for comments.

References

Alexander, J., R. Mallon, and J. M. Weinberg (2010). “Accentuate the Negative.” *Review of Philosophy and Psychology* 1(2), 297–314.

Beebe, J. R., and R. J. Undercoffer (2015). “Individual and Cross-Cultural Differences in Semantic Intuitions: New Experimental Findings.” Forthcoming in *Journal of Cognition and Culture*.

Bezuidenhout, A., and J. C. Cutting (2002). “Literal Meaning, Minimal Propositions, and Pragmatic Processing.” *Journal of Pragmatics* 34(4), 433–456.

Birdsong, D. (1989). *Metalinguistic Performance and Interlinguistic Competence.* Berlin: Springer-Verlag.

Bogen, J., and J. Woodward (1988). “Saving the Phenomena.” *Philosophical Review* 97(3), 303–352.

Braisby, N., B. Franks, and J. Hampton (1996). “Essentialism, Word Use, and Concepts.” *Cognition* 59(3), 247–274.

Buckwalter, W. (2010). “Knowledge Isn’t Closed on Saturdays: A Study in Ordinary Language.” *Review of Philosophy and Psychology* 1(3), 395–406.

Buckwalter, W., and J. Schaffer (2013). “Knowledge, Stakes and Mistakes.” *Noûs* 49(2), 201–234.

Cappelen, H. (2012). *Philosophy without Intuitions.* Oxford: Oxford University Press.

Chemla, E., and R. Singh (2014a). “Remarks on the Experimental Turn in the Study of Scalar Implicature: Part I.” *Language and Linguistics Compass* 8(9), 373–386.

Chemla, E., and R. Singh (2014b). “Remarks on the Experimental Turn in the Study of Scalar Implicature: Part II.” *Language and Linguistics Compass* 8(9), 387–399.

Cohnitz, D., and J. Haukioja (2013). “Meta-Externalism vs. Meta-Internalism in the Study of Reference.” *Australasian Journal of Philosophy* 91(3), 475–500.

Cohnitz, D., and J. Haukioja (2014). “Intuitions in Philosophical Semantics.” *Erkenntnis* 10.1007/s10670-014-9666-1.

Cova, F., and P. Egré (2015). “Moral Asymmetries and the Semantics of Many.” Forthcoming in *Semantics and Pragmatics*.

Culbertson, J., and S. Gross (2009). “Are Linguists Better Subjects?” *British Journal of Philosophy of Science* 60(4), 721–736.

Deutsch, M. (2009). “Experimental Philosophy and the Theory of Reference.” *Mind & Language* 24(4), 445–466.

Deutsch, M. (2015). “Kripke’s Gödel Case.” In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language.* London: Bloomsbury.

de Villiers, J., B. Myers, and R. J. Stainton (2013). “Revisiting Pragmatic Abilities in Autism Spectrum Disorders: A Follow-up Study with Controls.” *Pragmatics & Cognition* 21(2), 253–269.

de Villiers, J., B. Meyers, R. J. Stainton, and P. Szatmari (2012). “Differential Pragmatic Abilities and Autism Spectrum Disorders: The Case of Pragmatic Determinants of Literal Content.” In M. Macaulay and P. Garces-Blitvich (Eds.), *Pragmatics and Context*. Toronto: Antares, 1–24.

de Villiers, J., R. J. Stainton, and P. Szatmari (2007). “Pragmatic Abilities in Autism Spectrum Disorder: A Case Study in Philosophy and the Empirical.” *Midwest Studies in Philosophy* 31, 292–317.

Devitt, M. (2010). “Linguistic Intuitions Revisited.” *British Journal of Philosophy of Science* 61(4), 833–865.

Devitt, M. (2011). “Experimental Semantics.” *Philosophy and Phenomenological Research* 82(2), 418–435.

Devitt, M. (2012). “Whither Experimental Semantics?” *Theoria* 27(1), 5–36.

Genone, J. (2012). “Theories of Reference and Experimental Philosophy.” *Philosophy Compass* 7(2), 152–163.

Genone, J., and T. Lombrozo (2012). “Concept Possession, Experimental Semantics, and Hybrid Theories of Reference.” *Philosophical Psychology* 25(5), 717–742.

Geurts, B. (2002). “Donkey Business.” *Linguistics and Philosophy* 25(2), 129–156.

Geurts, B., and N. Pouscoulous (2009). “Embedded Implicatures?!?” *Semantics and Pragmatics* 2(4), 1–34.

Gross, S., and J. Culbertson (2011). “Revisited Linguistic Intuitions.” *British Journal of Philosophy of Science* 62(3), 639–656.

Haggqvist, S., and A. Wikforss (2015). “Experimental Semantics: The Case of Natural Kind Terms.” In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language*, pp. 109–138. London: Bloomsbury.

Hansen, N., and E. Chemla (2013). “Experimenting on Contextualism.” *Mind & Language* 28(3), 286–321.

Hansen, N., and E. Chemla (2014). “Color Adjectives, Standards and Thresholds: An Experimental Investigation.” Unpublished manuscript.

Ichikawa, J., I. Maitra, and B. Weatherson (2012). “In Defense of a Kripkean Dogma.” *Philosophy and Phenomenological Research* 85(1), 56–68.

Jylkkä, J., H. Railo, and J. Haukioja (2009). “Psychological Essentialism and Semantic Externalism: Evidence for Externalism in Lay Speakers’ Language Use.” *Philosophical Psychology* 22(1), 37–60.

Katz, J. J. (1975). “Logic and Language: An Examination of Recent Criticisms of Intentionalism.” In K. Gunderson (Ed.), *Language, Mind and Knowledge*, pp. 36–130. Minneapolis: University of Minnesota Press.

Knobe, J. (2012). “Experimental Philosophy.” In E. Margolis, R. Samuels, and S. P. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford: Oxford University Press.

Knobe, J., and S. Yalcin (2014). “Epistemic Modals and Context: Experimental Data.” *Semantics and Pragmatics* 7(10), 1–21.

Krifka, M. (2011). “Varieties of Semantic Evidence.” In C. Maienborn, K. von Heusinger, and P. Portner (Eds.), *Semantics: An International Handbook of Natural Language and Meaning*, pp. 242–267. Berlin: de Gruyter.

Kripke, S. (1977). “Speaker’s Reference and Semantic Reference.” *Midwest Studies in Philosophy* 2, 255–276.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Labov, W. (1996). “When Intuitions Fail.” In L. McNair, K. Singer, L. Dolbrin, and M. Aucon (Eds.), *Parasession on Theory and Data in Linguistics*, vol. 32, pp. 77–106. Chicago: Chicago Linguistics Society.

Lewis, D. (1999). “Many, but Almost One.” In *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Liao, S.-Y., and A. Meskin (2015). “Aesthetic Adjectives: Experimental Semantics and Context-Sensitivity.” *Philosophy and Phenomenological Research*, early view.

Ludwig, K. (2007). “The Epistemology of Thought Experiments.” *Midwest Studies in Philosophy 31*, 128–159.

Ludwig, K. (2010). “Intuitions and Relativity.” *Philosophical Psychology* 23(4), 427–445.

Machery, E. (2012). “Expertise and Intuitions about Reference.” *Theoria* 73, 37–54.

Machery, E., M. Deutsch, R. Mallon, S. Nichols, J. Sytsma, and S. Stich (2010). “Semantic Intuitions: Reply to Lam.” *Cognition* 117(3), 363–366.

Machery, E., M. Deutsch, and J. Sytsma (2015). “Speaker’s Reference and Cross-Cultural Semantics.” In A. Bianchi (Ed.), *On Reference*. Oxford: Oxford University Press.

Machery, E., R. Mallon, S. Nichols, and S. P. Stich (2004). “Semantics, Cross-Cultural Style.” *Cognition* 92, B1–B12.

Machery, E., C. Y. Olivola, and M. de Blanc (2009). “Linguistic and Metalinguistic Intuitions in the Philosophy of Language.” *Analysis* 69(4), 689–694.

Mallon, R., E. Machery, S. Nichols, and S. Stich (2009). “Against Arguments from Reference.” *Philosophy and Phenomenological Research* 79(2), 332–356.

Martí, G. (2009). “Against Semantic Multi-Culturalism.” *Analysis* 69(1), 42–48.

Martí, G. (2015). “General Terms, Hybrid Theories and Ambiguity: A Discussion of Some Experimental Results.” In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language*, pp. 157–172. London: Bloomsbury.

Maynes, J., and S. Gross (2013). “Linguistic Intuitions.” *Philosophy Compass* 8(8), 714–730.

Nado, J. (2014). “Philosophical Expertise.” *Philosophy Compass* 9(9), 631–641.

Nichols, S., N. Ángel Pinillos, and R. Mallon (2015). “Ambiguous Reference.” Forthcoming in *Mind*.

Nisbett, R. E., K. Peng, I. Choi, and A. Norenzayan (2001). “Culture and Systems of Thought: Holistic vs. Analytic Cognition.” *Psychological Review* 108, 291–310.

Phelan, M. (2010). “The Inadequacy of Paraphrase Is the Dogma of Metaphor.” *Pacific Philosophical Quarterly* 91(4), 481–506.

Phelan, M. (2014). “Experimental Pragmatics: An Introduction for Philosophers.” *Philosophy Compass* 9(1), 66–79.

Pietroski, P., J. Lidz, T. Hunter, and J. Halberda (2009). “The Meaning of ‘Most’: Semantics, Numerosity and Psychology.” *Mind & Language* 24(5), 554–585.

Pinillos, A. (2015). “Ambiguity and Referential Machinery.” In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language*, pp. 139–156. London: Bloomsbury.

Putnam, H. (1975a). “Is Semantics Possible?” In *Mind, Language and Reality: Philosophical Papers*, vol. 2, pp. 139–152. Cambridge: Cambridge University Press.

Putnam, H. (1975b). “The Meaning of ‘Meaning.”’ In *Mind, Language and Reality: Philosophical Papers*, vol. 2., pp. 215–271. Cambridge: Cambridge University Press.

Raffman, D. (2014). *Unruly Words.* New York: Oxford University Press.

Ripley, D. (2015). “Contradictions at the Borders.” In R. Nouwen, R. van Rooij, U. Sauerland, and H.-C. Schmitz (Eds.), *Vagueness in Communication,* pp. 169–188. Dordrecht: Springer.

Schütze, C. T. (1996). *The Empirical Base of Linguistics.* Chicago: University of Chicago Press.

Schütze, C. T. (2011). “Linguistic Evidence and Grammatical Theory.” *WIREs Cognitive Science* 2(2), 206–221.

Sprouse, J. (2013). “Acceptability Judgments.” *Oxford Bibliographies* (**http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0097.xml**).

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.

Sytsma, J., and J. Livengood (2011). “A New Perspective Concerning Experiments on Semantic Intuitions.” *Australasian Journal of Philosophy* 89(2), 315–332.

Sytsma, J., J. Livengood, R. Sato, and M. Oguchi (2015). “Reference in the Land of the Rising Sun: A Cross-Cultural Study on the Reference of Proper Names.” *Review of Philosophy and Psychology* 6(2), 213–230.

Notes:

([1]) This description is based on the account of experimental philosophy in general given in Knobe (2012).

([2]) “Intuition” is a hotly contested term in contemporary philosophy. For present purposes, “intuition” should be understood to pick out a judgment that is “not based on conscious reasoning, past or present, one’s own or another’s” (Maynes and Gross, 2013, p. 716). Maynes and Gross (2013) provide a very useful survey of different conceptions of linguistic intuitions and how they figure in debates in linguistics and philosophy.

([3]) For more detailed discussion of issues related to this assumption, see the discussion of “meta-internalism” and “meta-externalism” in Cohnitz and Haukioja (2013).

([4]) See also Cappelen (2012).

([5]) Machery et al. report large standard deviations in responses to the Gödel stories, which suggests that there is a great deal of variation *within* each of the two cultural groups (p. 8). The finding of intra-cultural (but not cross-cultural) variation was replicated in Machery et al. (2009), and the finding of cross-cultural variation was replicated in Machery et al. (2010), using a Chinese translation of the Gödel story for Chinese participants.

([6]) Genone (2012) is a useful survey of the debate surrounding experimental investigations of theories of reference for proper names. See Genone (2012, p. 156) for roughly this categorization of responses to Machery et al. (2004).

([7]) While there are linguistic tasks that don’t ask participants to respond to metalinguistic stimuli (elicited production, for example, and tasks that request actions in response to commands), the default type of experimental task (acceptability judgments in syntax, truth value and entailment judgments in semantics and pragmatics) involves metalinguistic judgments. Birdsong (1989, p. 2) observes that acceptability judgments are the “prototypical metalinguistic performance in the language sciences.” See Schütze (1996, 2011) for surveys of different types of evidence for linguistic theories, metalinguistic and otherwise.

([8]) Schütze’s observation was brought to my attention by Cohnitz and Haukioja (2014), which includes a very helpful and detailed examination of the difference between metalinguistic intuitions and linguistic use.

([9]) When Martí herself recommends an improved prompt for collecting evidence of theories of reference based on use, it too involves a metalinguistic judgment (p. 47).

([10]) See also Kripke (1980, p. 85, n. 36), Ichikawa et al. (2012, pp. 59–60), Ludwig (2007, p. 150), and Sytsma and Livengood (2011, §2.2).

([11]) See Kripke (1977) for the canonical statement of the distinction.

([12]) Ichikawa et al. (2012, p. 59) and Deutsch (2009, p. 454, n. 7) both say that it is easier to hear the prompt as asking about speaker reference. Machery et al. (2015) argue that because the prompt does not ask about any particular use of “Gödel,” it isn’t possible to figure out what intentions the speaker might have had in using it, and so the semantic reference interpretation is the only one available.

([13]) Devitt (2010, p. 836) explicitly compares his and Martí's terminology.

([14]) Devitt's views about expertise extend only to (meta)linguistic intuitions, not to other forms of linguistic or nonlinguistic behavior, which he considers to be "more direct" evidence of linguistic reality (Devitt, 2010, p. 500; 2011, p. 425).

([15]) For a survey of the "expertise defense" as it arises in areas beyond philosophy of language, see Nado (2014).

([16]) See Devitt (2010) and Gross and Culbertson (2011) for further discussion.

([17]) Note that this experiment used a prompt ("when Ivy uses the name 'Tsu Ch'ung Chih,' who do you think she is actually talking about?") that doesn't distinguish semantic and speaker reference as clearly as the disambiguated prompt in Machery et al. (2015). It is possible that different experts have different levels of sensitivity to the semantic/speaker reference distinction, which could account for the different rates of Kripkean responses. Thanks to a referee for this observation.

([18]) See Sprouse (2013) for a survey of the debate in linguistics.

([19]) The traditional view is typically attributed to Kant. Katz (1975) defends a version of the traditional view against Putnam and Kripke's criticisms.

([20]) For criticism of Jylkkä et al.'s scenarios on the grounds that they prime essentialist responses, see Haggqvist & Wikforss (2015).

([21]) Machery et al. (2004, p. 8) recognize that it is a "very live possibility" "that the variability exists even at the individual level, so that a given individual might have causal-historical intuitions on some occasions and descriptivist intuitions on other occasions," but they conclude (contrary to Genone and Lombrozo) that that possibility shows that "the assumption of universality is just spectacularly misguided."

([22]) Nichols et al. (2015) seem to assume that the causal-historical theory wouldn't treat "catoblepas" as having no reference at all, because of the existence of a causal-historical connection to wildebeests.

([23]) A referee suggested that a subject might hear the sentence "'Catoblepas' refers to wildebeests" as pragmatically conveying something like "the notion of the mythical catoblepas came about because of encounters with wildebeests." If that's the case, someone who agrees with statement 1 because they agree with what it pragmatically conveys might at the same time deny that catoblepas exist, and that wouldn't pose a

problem for the causal-historical theory. Pinillos (2015) discusses some related worries about how the experiments in Nichols et al. (2015) are interpreted.

([24]) If “catoblepas” and “wildebeests” are not co-referential, then it is not puzzling why responses to 1 and 2 would differ: Participants might be treating “catoblepas” as a subset of “wildebeests,” which would lead them to agree with 1 and disagree with 2. For example, one should agree with “dogs are animals” but disagree with “animals are dogs.” Thanks to Daniel Cohnitz for discussion of this point.

([25]) See Martí (2015) for critical discussion of these results.

([26]) Genone (2012, p. 153), for example, says that Machery et al. (2004) has “become an exemplar for a particular way of understanding the goals, methods, and prospects for experimental philosophy.”

([27]) While this article was going to press, Josh Knobe independently posted a similar list surveying recent work in experimental semantics at the Experimental Philosophy blog: http://philosophycommons.typepad.com/xphi/2015/04/formal-semantics-and-experimental-philosophy.html.

([28]) For surveys, see Chemla and Singh (2014a, 2014b) and Phelan (2014).

Nathaniel Hansen

Nathaniel Hansen, University of Reading, UK