Canadian Journal of Philosophy (2025), 55: 4, 363–378
doi:10.1017/can.2026.10063

ARTICLE

Confusion and Convention

Atheer Al-Khalfa

Australian Catholic University, Sydney, Australia
Email: atheer.alkhalfa@gmail.com

Abstract
Cases of identity confusion raise a challenge for convention-based meta-semantics. David Lewis's influential account of convention, as it applies to language use, fails to distinguish, among two candidate languages, the one used by the population in question. The problem extends to Paul Grice's influential account of sentence meaning.

Keywords: linguistic convention; metasemantics; interpretation; confusion; sentence meaning

1. Introduction

Some names semantically corefer. Frege (1948) popularized the case involving the names "Hesperus" and "Phosphorus," which both semantically refer to the same object in the sky, yet they seem to have different meanings, or a difference in *cognitive value*. Cases like this raise problems for various views of meaning such as Millianism according to which the meaning of a name is exhausted by its semantic referent. A kind of reverse problem arises when we think of cases that involve only *one* name with two speaker referents, where the challenge is to give the intuitively correct verdicts about semantic reference in such cases. Cases of identity confusion make this challenge particularly vivid. Two influential convention-based meta-semantic accounts fail to meet this challenge. The first is Lewis's (1983) influential account of convention as it applies to language use.[1] In Section 2, I briefly outline the elements of this account. In Section 3, I go through the kind of cases I have in mind and say why they pose a challenge for the account. In Section 4, I examine a potential refinement of Lewis's account based on *counterfactual regularities* and argue that this strategy ultimately fails. In Section 5, I show how, under certain principles that enjoy prima facie plausibility, a similar challenge can be raised for Grice's (1989) account. I conclude in section 6 pointing out why the problems raised will likely confront other similar accounts.

2. Lewis's Account of Convention

According to Lewis (1983, p. 163), a language is a function that maps strings of types of sounds or of marks to meanings (for Lewis, meanings are sets of possible worlds).[2] Lewis (1983, pp. 166–167)

[1]This remains a popular and much discussed view. See for example, Armstrong (2016), Gilbert (2008), Geurts (2018), Keiser (2023), Ravenscroft (2015), and Yalcin (2014). However, this literature has not engaged with the quite foundational problem presented here.

[2]Lewis is aware that this is too simple a formulation and ends up making certain tweaks to it that accomodate indexicality, ambiguity, and non-indicatives. These tweaks do not target the kind of issues I will be raising.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Canadian Journal of Philosophy, Inc. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

claimed that a population *P* uses a particular language by virtue of the *convention* of language prevailing in *P*. According to Lewis, a convention (with respect to the use of language £ in population *P*) is a regularity in a specific sort of action or in a specific sort of action and belief. The regularity that matters for language, Lewis says, is a regularity of *truthfulness* and *trust*. Regarding these two notions he says:

> To be truthful in £ is to act in a certain way: to try never to utter any sentences of £ that are not true in £. Thus it is to avoid uttering any sentence of £ unless one believes it to be true in £. To be trusting in £ is to form beliefs in a certain way: to impute truthfulness in £ to others, and thus to tend to respond to another's utterance of any sentence of £ by coming to believe that the uttered sentence is true in £.[3] (Lewis, 1983, p. 167)

One point is worth noting here. Consider the following propositions:

(I) *"snow is white" is true in English*

(II) *snow is white*

Lewis says that propositions are sets of possible worlds and that languages are functions from sentences to sets of possible worlds. He adds that "a sentence σ is true in a language £ at a world *w* if and only if *w* belongs to the set of worlds £(σ)" (Lewis, 1983, p. 163). Given this, propositions (I) and (II) are, for Lewis, identical. That is (I) and (II) are one and the same set of possible worlds. Thus, for Lewis, the following two ways of speaking are interchangeable:

- To be truthful in £ is to avoid uttering any sentence of £ unless one believes that that sentence is true in £.
- To be truthful in £ is to avoid uttering any sentence of £ unless one believes the proposition assigned to that sentence by £.

Similarly, for Lewis, the following two ways of speaking are interchangeable:

- To be trusting in £ is to tend to respond to another's utterance of any sentence of £ by coming to believe that that uttered sentence is true in £.
- To be trusting in £ is to tend to respond to another's utterance of any sentence of £ by coming to believe the proposition assigned to that uttered sentence by £.

According to Lewis (1983, pp. 167–168), for this regularity of truthfulness and trust (with respect to the use of language £ in population *P*) to count as a convention, the following six conditions must hold:

(i) There prevails in *P* at least a regularity of truthfulness and trust in £.

(ii) The members of *P* believe that this regularity of truthfulness and trust in £ prevails among them. Each believes this because of his experience of others' past truthfulness and trust in £.

[3]Note the logical structure of Lewis's claims. They can be understood as the following conditionals. 1) if a speaker utters σ then they believe σ is true in language £ and 2) if a hearer hears σ they will come to believe σ is true in language £. Note also that in another place Lewis says 'and the hearer (or reader) responds by coming to share that belief of the speaker's *(unless he already had it)*, and adjusting his other beliefs accordingly' (1983, p. 167, emphasis mine). Lewis therefore is open to hearers *already* having the belief in question *before* the speaker's utterance. Moreover, note that given Lewis's Bayesian orientation one could also rewrite what it is to be truthful and trusting in the language of *subjective credence*. So written, a regularity of truthfulness in £ would become a regularity of uttering a sentence of £ only when the speaker has high credence in the proposition £ assigns to that sentence. And a regularity of trust in £ would become a regularity of the hearer increasing their credence in the proposition £ assigns to the sentence heard.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

(iii) The expectation of conformity ordinarily gives everyone a good reason why he himself should conform.
(iv) There is in *P* a general preference for general conformity to the regularity of truthfulness and trust in £. Given that most conform, the members of *P* want all to conform.
(v) The regularity of truthfulness and trust in £ has alternatives.
(vi) Facts (i) to (v) are common knowledge in *P*. Everyone knows them, everyone knows that everyone knows them, and so on.

Before moving on to the critical component of this paper, I should emphasise that Lewis's above account is an account of how the content of sentences in a language is fixed. It is *not* an account of how the content of beliefs are fixed. Lewis's account just assumes or helps itself to the agents' actual beliefs.

3. Two Cases of Common Confusion

3.a. Assassination

Recall Kripke's (1977, p. 263) leaf raking case in which two people see Smith in a distance and mistake him for Jones. They say things like "Jones is raking leaves." Kripke points out that the semantic value of "Jones" in this case is just Jones (not the person the conversational participants are looking at). However, Kripke maintains that there is a sense in which the speaker and audience are talking about Smith—a merely pragmatic sense. Strictly speaking, Kripke would say, the speaker's utterance of (say) "Jones is raking leaves" is false if (for example) Jones is asleep, regardless of what Smith is doing. Take the following communal variant of the leaf raking case.

> **Assassination.** Jones (from now on X) is known by all the inhabitants of a village; he is the local butcher. The villagers see him daily, speak with him, and call him 'Jones'. One day, some villager sees an individual Y from a distance and, due to Y's striking resemblance to X, they mistake Y for X. While watching Y (believing him to be X), the villager discovers that Y is secretly plotting with an outsider to assassinate the village leader. Alarmed, the villager calls the rest of the village to join in spying on Y—who they all similarly mistake for X. The villagers collectively observe Y for some time, learning new details about his life, personality, habits, and plans for assassination. As they disperse, they say things like, 'I cannot understand why Jones is doing this'. They also use perceptual demonstratives in utterances like 'That guy [pointing at Y] is an assassin' and 'That guy [pointing at Y] is Jones'. They then gather to discuss, using the name 'Jones', a counter-plan to foil the assassination.[4]

Assassination involves speakers who make an identity mistake and use the name "Jones" to refer.[5] But there are at least two kinds of reference. There is *semantic reference*: a three-place relation between term, context, and object (as semantic value). And there is *speaker reference*: a four-place relation between speaker, term, object, and audience. A speaker referent is an object that a speaker uses a term to get their audience to think about.[6] (I will use the phrases "semantic referent" and "semantic value" interchangeably.) Given the distinction between these two kinds of reference and

[4]"X" and "Y" are not terms in the language the villagers are using. They are terms in the meta-language used to discuss these cases.

[5]Trickier cases of confusion received a systematic study in Camp (2002).

[6]Here I am helping myself to an approximate version of *speaker reference* as explicated by Bach who says "Speaker reference is a four-place relation, between a speaker, an expression, an audience, and a referent: you use an expression to refer someone to something" (Bach, 2006, p. 520). Other explications of speaker reference are available, see notably Kripke (1977). Although there are interesting questions about how speaker reference—and more generally, pragmatically conveyed propositions—are fixed, these questions are beyond the scope of this paper. This paper is about how semantic content is fixed.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

given that the mistake made by the spying villagers is in thinking that two different individuals are identical, I find it natural to say that when those spying villagers use the name "Jones" they speaker refer to at least two different individuals: X and Y.[7]

What proposition are the villagers expressing? There are certain delicate cases of confusion under which it's not clear what the semantic referent is and therefore what is being expressed, but the case of **Assassination** is not such a case; it's a common and mundane case of mistaken identity. It's similar to the following case in which intuitions might be more clear.

Suppose Joe Biden is in the library and mistakes the guy sitting across from him (call him "V") for Barack Obama. As a result, Biden says things like "that guy [pointing at V] is Obama" and "Obama is sitting across from me." There is an overwhelming intuition that these sentences expresses a false proposition because the perceptual demonstrative "that guy [pointing at V]" semantically refers to V and "Obama" semantically refers to Obama; and V is not Obama. Suppose Biden invites his friends to the library and starts reporting to them what he believes. When his friends start uttering to each other things like "Obama is in the library; he is sitting over there" they all utter a falsehood—Obama is not in the library. Before the confusion, "Obama" names Obama, and it is not clear what it is about this kind of confusion that would block the name from continuing to name Obama after the confusion. Again, this is a mundane case of mistaken identity much like the case of **Assassination**.

In the case of **Assassination**, what X's name is before the confusion sets in is not supposed to be at issue. The villagers know X's name is "Jones"; X is the semantic value of "Jones." The question is what the semantic value is when the villagers start holding the false identity belief. But just as it is overwhelmingly obvious that "Obama" continues to name the same person after the confusion sets in, so too "Jones" continues to name X. The case can be set up so that "Jones" is a well-known name (like "Obama") and X a well-known individual (like Obama).

Return to **Assassination** and suppose that Y is the only assassin. When those villagers utter "Jones is an assassin," they are uttering a falsehood. This is because the semantic value of "Jones" is still X. And naturally, when a villager, while confused, is looking at Y and they utter "That guy [pointing at Y] is Jones' they utter a falsehood; Y is not X. This is the mistake in this case of mistaken identity. We may imagine that X himself had been keeping track of the villagers' movements and knows about their confusion. It seems completely appropriate that X reacts to their utterances of "Jones is an assassin" by uttering "No I am not." This is the kind of reaction one would expect if X was being accused of something false. I take this as evidence that the proposition semantically expressed by "Jones is an assassin" is *X is an assassin.*[8]

What happens if the villagers are cleared of their confusion? They would withdraw utterances like "Jones is an assassin" because they think it was false. And given that Y is in fact an assassin and they know him to be, they think it was false because "Jones" is a name for X not Y. In contrast, they will not withdraw sentences like "That guy [pointing at Y] is an assassin" because they would still think it expressed a truth about Y. These observations line up well with how an omniscient being is disposed to reporting their beliefs before their confusion is cleared. Such a creature would report them *truly* by uttering sentences like the following.

[7]Natural but not uncontroversial. However, crucially, the problems I raise in this paper do not turn on this. The problems go through even if one thinks that there is no speaker reference to any individual. What is needed for the problems to go through (as will soon become clear) is that a certain pattern of belief formation holds with respect to certain sentences.

[8]One might say, as one referee suggested, that the reason why X reacts by uttering "No I am not" is because *X is an assassin* is the more historically dominant or entrenched meaning of the sentence "Jones is an assassin" and not because that is what it semantically expresses in the villagers' mouths. I believe the phenomenon of semantic reference change tells against this kind of suggestion. Take a case of semantic change that does not involve confusion, but involves semantic change that is induced by community wide stipulations of the form "From now on 'Jones' is to mean that guy [pointing at Y]." After a short time it may very well be that the only proposition semantically expressed by "Jones is home" is *Y is home* even though it is the less historically dominant/entrenched meaning relative to *X is home.* Crucially, how the villagers are disposed to react will be inline with the less historically dominant/entrenched meaning. Parallel problems will arise when attempting to always privilege earlier rather than later regularities.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

a. "They wrongly think that Jones is an assassin"

Given that Y is the only assassin, the most natural explanation for why this belief report is true is that the semantic value of "Jones" is X. (It follows also from the truth of these reports that the confused villagers do have the belief that X is an assassin and that this belief is false.) This omniscient creature would persist in making these reports even when the confusion continues for a while. Intuitions are sharper when we imagine that Y is posing as X. An omniscient being is disposed to reporting the villagers truly as follows. (And again, the semantic values I've been arguing for predict this result.)

b. "This guy [pointing at Y] has everyone fooled. He is not Jones"

But an assignment of Y to "Jones" in this report predicts that this omniscient creature is uttering a falsehood.

What do the villagers believe when uttering or hearing a sentence like "Jones is an assassin"? I have just argued that "Jones is an assassin" semantically expresses that *X is an assassin.* Naturally, given that the speakers are sincere that's what the speakers believe and that's what the hearers come to believe (given they have no reason to doubt the villagers' sincerity). Moreover, assuming that the villagers have a fairly modest degree of deductive competence (i.e., they believe the obvious deductive consequences of their beliefs) they also, via the below reasoning, come to believe that *Y is an assassin.*[9]

c. *X is an assassin*.
d. *X is Y*.
e. Therefore: *Y is an assassin*.

Naturally, the above sequence is expressing a train of thought for each of the villagers that culminates in another belief.[10] Though the above sequence is valid, my point is not undermined if the reader thinks it's not. What is important is that the villagers arrive at the conclusion after some reasoning. The above considerations then show that the villagers believe both that *X is an assassin* and that *Y is an assassin* when uttering or hearing "Jones is an assassin."[11]

We can run the above considerations without names and intuitions might be sharper for some. For example, take the following sentence (uttered by the confused villagers using a perceptual demonstrative while looking at Y) and candidate propositions.

[9]This is the same kind of deductive competence exhibited in the following case that does not involve confusion. I am sitting across from Sarah, and I know her name is "Sarah." I believe that *Sarah is sitting there*, and I also believe that *Sarah is this person in front of me.* Therefore: I believe that *this person in front of me is sitting there.*

[10]Parallel remarks will apply if we do not assume Millianism. Suppose the semantic value of "Jones" is the *the butcher of the northern village.* If so then upon hearing 'Jones is an assassin' the villagers will believe *the butcher of the northern village is an assassin.* And given that they also believe that *the butcher of the northern village is identical to the person being perceived*, they will, via the same modest deductive competence, conclude that *the person being perceived is an assassin.*

[11]News outlets frequently report alleged sightings of well-known figures. For example, after Saddam Hussein went into hiding, a man appeared in a video posing as Saddam. Viewers might have fallen for the hoax and reported "Saddam is in this video." There is an overwhelming intuition here that utterers of this sentence (despite their identity confusion) express a false belief about Saddam which is passed on to other members of the public via similar utterances. In other words, this kind of identity confusion (which is similar to the one in **Assassination**) does not make it the case that the beliefs expressed are empty or "partial" (or that "Saddam" stops being a name for Saddam). Such nonstandard views about the content of their beliefs and utterances would have trouble explaining the community's subsequent actions. See https://www.wired.com/2003/07/is-it-saddam-on-arab-tv-or-hoax/.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

f. "This guy [pointing at Y] is standing there"

g. *Y is standing there*

h. *X is standing there*

It is hard to deny that (f) semantically expresses (g). After all, the speakers' visual system is intact, and they are indeed pointing at Y and the relevant location when making that utterance. Naturally, speakers of this sentence and hearers will believe (g) when (f) is uttered. But because the villagers believe that *X is Y* they will also, via the same very modest degree of deductive competence, arrive at (h).

In what follows however, I return to using names to illustrate the challenge. Names, unlike visual demonstratives, have semantic continuity (regardless of confusion) which make them appropriate for illustrating the challenge for Lewis's account.

What language are the villagers using? Let us summarize the above as follows. We have the sentence (1) and the propositions (2) and (3):[12]

(1) "Jones is an assassin"

(2) *X is an assassin*

(3) *Y is an assassin*

(2) and (3) are two candidates for the proposition that is semantically expressed by speakers of the sentence in (1). As I have just tried to emphasize, it is plausible that (at least at an early stage of the confusion) the proposition that is semantically expressed by (1) is (2) and not (3). This kind of situation raises a problem for Lewis's account given that when a speaker utters (1) they believe (2) *and* (3) and hearers in the population respond to (1) by coming to believe both (2) *and* (3). Suppose further that almost all of the villagers do not find out about their confusion and none of the apparent facts give them any reason to doubt their beliefs and so the speakers continue to utter (1) believing (2) and (3) and the hearers in the population respond to (1) by coming to believe both (2) and (3).[13]

Let us now distinguish between two possible languages: £$_1$ and £$_2$. Suppose that out of £$_1$ and £$_2$ it is only £$_1$ that tracks our above semantic intuitions. That is, suppose it is only £$_1$ that assigns *X is an assassin* to "Jones is an assassin" and *X is executing their plan* to "Jones is executing their plan" and so on. As a rough way of describing £$_1$ let us say that it is the language that assigns the X meanings to the "Jones" sentences. And suppose that £$_2$ is the language that assigns *Y is an assassin* to "Jones is an assassin" and *Y is executing their plan* to "Jones is executing their plan" and so on. Similarly, let us say that £$_2$ is the language that assigns the Y meanings to the "Jones" sentences. This is the only

[12]Recall that Lewis says that meanings are functions from sentences to sets of possible worlds or simply, sets of possible worlds. If so, (2) is the set of possible worlds in which X is an assassin and (3) is the set of possible worlds in which Y is an assassin. These sets are not identical.

[13]I have assumed that the candidate propositions expressed by (1) are (2) and (3). These are constrained by the orthodox view of the semantic value of proper names namely, Millianism. Lewis however, seems to have held a view according to which names may semantically function as descriptions of a causal sort. He says: "The descriptive sense associated with a name might for instance be 'the place I have heard of under the name "Toromeo"' or maybe 'the causal source of this token: Toromeo', and for the account of the relation being invoked here, just consult the writings of causal theorists of reference" (Lewis, 1999, p. 353, n22). But if one thinks that the causal descriptive belief corresponding to the sentence in (1) (whatever that is) is held by speakers of (1) in *P* and comes to be believed by hearers of (1) in *P* then our problem is just made worse, for all this means is that *P* are truthful and trusting in another language simultaneously: £$_3$. £$_3$ is the language that assigns the causal descriptive belief (whatever that is) to (1) and is otherwise just like English. So long as we cannot deny that (2) *and* (3) are *also* both believed by speakers uttering (1) and come to be believed by hearers of (1) then our problem will still go through; how does Lewis's account privilege the intuitively correct language being used by *P* out of the (now) three candidates?

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

difference between the two languages. That is, £$_1$ assigns sentences not containing the name "Jones" the same meaning that £$_2$ assigns. This way we end up with the following challenge for Lewis:

> Given that there is a regularity of truthfulness and trust in both, by virtue of what is *P* using £$_1$ and not £$_2$?

Lewis's condition (i) in Section 2 is met in relation to both £$_1$ and £$_2$.[14] Moreover, most members in *P* believe that both regularities associated with £$_1$ and £$_2$ prevails among them—they just think it's the same regularity because they think that $X = Y$. They believe that the regularity in £$_1$ prevails among them because of their experience of others' past truthfulness and trust in £$_1$ and they believe that the regularity in £$_2$ also prevails among them because of their experience of others' past truthfulness and trust in £$_2$. So condition (ii) is also met. And conditions (iii), (iv), (v), and (vi) are met too in relation to both £$_1$ and £$_2$.[15]

Lewis's account must privilege £$_1$ as the only language being used by *P* for it to make the intuitively correct semantic predictions. But it seems that his account can offer no answer to the above question. The prediction Lewis's account makes about **Assassination** is that there are two languages that are used by *P* simultaneously. What this means is that members of *P* semantically express (2) and (3) with their utterances of (1). I've argued that this is implausible.

One might think that after some time the semantic value of "Jones" is going to change to Y and only Y and so one might think that after a while the proposition expressed by (1) is (3) only (and not (2) only anymore). In other words, one might claim that after a while the language that the villagers are using is £$_2$ only. That might be so, but that does not change Lewis's predictions *before* the alleged semantic value change. And the problem remains *after* the alleged semantic value change given that when a speaker utters (1) they would still, even then, believe (2) and (3) and the hearers in the population, even then, would still respond to (1) by coming to believe both (2) and (3). The problem remains so long as this belief pattern persists. And this belief pattern persists so long as members of the community believe that $X = Y$. Both before and after the alleged semantic value change Lewis's account has no way of privileging, among two possible languages, the one that would be intuitively used; whatever that *one* is (be it £$_1$ only or £$_2$ only).[16]

3.b. Madagascar

Take another case of common confusion. Evans (1973, pp. 195–196) reports (from Isaac Taylor's (1898) book *Names and their History*) that "Madagascar" was a name for a section of the African mainland. Upon hearing it (perhaps from some Arab or Malay sailors), Marco Polo took it to designate that island we today call "Madagascar" and started using it this way. A fairly widespread intuition is that the semantic referent of "Madagascar" shifted sometime afterwards. Utterances of "Madagascar is an island" are, intuitively, true today even though they were not before Marco Polo used the term.[17] That the semantic value of "Madagascar" has shifted is one way to explain these

[14]Met or *almost* met given that X and Y are not themselves confused. Lewis (1983) says that a few exceptions can be tolerated (p. 165). The next case of **Madagascar** removes this complication.

[15]Kölbel (1998) identifies some regularities in the vicinity of the one Lewis proposes. However these do not seem helpful in overcoming the challenge.

[16]One might try saying, on behalf of Lewis, that in this case it is vague or indeterminate which language is being used by *P*. The problem with this suggestion is that it is in tension with the truth-value judgments surveyed earlier regarding what the relevant sentences semantically express. In other words, intuitively, it is not indeterminate or vague which proposition is semantically expressed. Some might also take up a suggestion in Lewis (1983) and claim something along the following lines. What makes it the case that members of *P* use £$_1$ and not £$_2$ is some causal/historical asymmetry between X and Y. This kind of suggestion has its own familiar problems and, in any case, it risks robbing Lewis's account of its distinctiveness and interest. Recall that Lewis's account is supposed to be a *convention*-based meta-semantic account and not a causal-historical one.

[17]Burgess (2014) gives a more detailed and accurate account of the history of the name "Madagascar". However, the problems presented here will not turn on these historical complications. Much like how **Assassination** and variants of it are

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

truth-value observations. I want to discuss a variant of this case which focuses on how things might have gone very shortly after Marco Polo overhears the name "Madagascar" being used by those Arab sailors.

> **Madagascar.** Upon hearing the name 'Madagascar' Marco Polo thinks that the Arab sailors are talking about an island. (But he did not think of any one particular island that he was already familiar with, that it was identical to the thing those Arab sailors were talking about.) Marco Polo then introduces the name to his friends by saying things like "I heard some Arab sailors talking about an island called 'Madagascar', it sounds exciting" and so on. Marco Polo continues to use the name by saying things like 'Madagascar has very few inhabitants', 'Madagascar has plenty of natural resources' and so on. After that, Marco Polo and his friends get on a ship and discover an island (call it 'Y'). For some reason or other they think that Y is the island that is identical to what those Arab sailors were talking about. Marco Polo and his friends land on Y, unpack and set themselves up for permanent stay. They also design a flag and elect a ruler forming a country with a distinctive identity. Suppose all the above happens within a couple of days of Marco Polo's introduction to the name 'Madagascar'. Throughout, Marco Polo and his friends continue to use the name 'Madagascar' uttering sentences like 'Madagascar is a wonderful island'.

When Marco Polo first picks up the name and introduces it to his friends and they use it, they all use it with that section of mainland Africa as semantic value. Let us call that section of mainland Africa "X."[18] At this point, it seems plausible to say that Marco Polo and his friends express beliefs only about X; they are just saying lots of false things about it. But once they discover island Y, they start to believe that $X = Y$. Still, on the day of discovery of that island, it seems that the semantic value of "Madagascar" in their mouths continues to be X only (recall the situation with Kripke's leaf raking case and **Assassination** which seems relevantly similar to **Madagascar** at this point of the story). In what follows, I say more about the beliefs held by Marco Polo and his friends, as well as what they express when making certain utterances. Given the similarities between **Madagascar** and **Assassination,** my discussion will be quite brief.

What proposition are Marco Polo and his friends expressing? Upon arrival on Y (the island), when Marco Polo utters "Madagascar is a wonderful island," he semantically expresses the proposition that *X is a wonderful island.* Naturally, when Marco Polo utters "This island [pointing at Y] is Madagascar" this utterance expresses the false proposition that *Y is X.* Suppose a trusted outsider arrives on Y very shortly after Marco Polo and his friends arrive there. This outsider informs Marco Polo and his friends of their confusion. When Marco Polo responds to this information by uttering "Oh okay, so this island [pointing at Y] is not Madagascar," that seems like the appropriate response and it's true. Interpreting his utterance by assigning X to "Madagascar" and Y to "this island" predicts this judgment. However, on a view that says that "Madagascar" has, for example, Y as semantic value then Marco Polo semantically expressed the falsehood that *oh ok, so Y is not Y.* This is the wrong result. Once Marco Polo is cleared of confusion he expresses a truth not a falsehood.

Moreover, an omniscient being is disposed to reporting the situation before the confusion is cleared as follows.

imaginary, **Madagascar** and variants of it may be imaginary. What is distinctive about the case of **Madagascar** is that a) it involves *total* community-wide confusion (unlike **Assassination**), and b) it involves long-term "misuse," which many writers on the topic believe induces semantic change (this naturally lends itself to the discussion in Section 4 regarding counterfactual regularities).

[18]As one referee emphasizes, the piece of the African mainland that Marco Polo confused with the island off the coast of Africa is, according to most scholars, Mogadishu—see Oliver (1977).

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

i. "They are all wrong. This place [pointing at Y] is not Madagascar"

But an assignment of Y to "Madagascar" in this report predicts that this omniscient creature is uttering a falsehood. An omniscient being will continue reporting them in the above away even if the confusion continues for a while. Again, the case of **Madagascar** as I have staged it is a mundane case of mistaken identity much like the case of **Assassination**.

What do Marco Polo and his friends believe? I have just argued that "Madagascar is a wonderful island" semantically expresses that *X is a wonderful island.* Naturally, given that the speakers are sincere that's what the speakers believe and that's what the hearers come to believe (given they have no reason to doubt the sincerity of the speakers). Moreover, assuming that Marco Polo and his friends have a fairly modest degree of deductive competence (i.e., they believe the obvious deductive consequences of their beliefs), they also, via the below reasoning, come to believe that *Y is a wonderful island.*

j. *X is a wonderful island.*

k. *X is Y.*

l. Therefore: *Y is a wonderful island.*

Again, the above sequence plausibly expresses a piece of reasoning that culminates in another belief for Marco Polo and his friends.

What language are Marco Polo and his friends using? Take now the sentence (4) and the propositions (5) and (6) below:

(4) "Madagascar is a wonderful island"

(5) *X is a wonderful island*

(6) *Y is a wonderful island*

As in the previous case, we are now in a situation in which the above propositions are at least two candidates for the proposition that is semantically expressed by speakers of (4). I claimed that (at least when *P* first arrive on Y) their utterances of (4) semantically express (5) only. This case poses a challenge for Lewis's account given that when a speaker utters (4) they believe (5) *and* (6) and the hearers in the population respond to (4) by coming to believe both (5) *and* (6). Suppose further that no one finds out about this mistake in the population in question and none of the apparent facts give them any reason to doubt their beliefs, and so the speakers continue to utter (4) believing (5) and (6), and the hearers in the population respond to (4) by coming to believe both (5) and (6).

As with the previous case, let us now distinguish between two possible languages: £$_1$ and £$_2$. Suppose that between £$_1$ and £$_2$ it is only £$_1$ that tracks our above semantic intuitions. That is, suppose it is only £$_1$ that assigns *X is a wonderful island* to "Madagascar is a wonderful island" and *X is safe* to "Madagascar is safe," and so on. As a rough way of describing £$_1$ let us say that it is the language that assigns the X meanings to the "Madagascar" sentences. And suppose that £$_2$ is a language that assigns *Y is a wonderful island* to "Madagascar is a wonderful island" and *Y is safe* to "Madagascar is safe" and so on. Similarly, let us say that £$_2$ is a language that assigns the Y meanings to the "Madagascar" sentences. This is the only difference between both languages. This way we end up with the same challenge for Lewis:

> Given that there is a regularity of truthfulness and trust in both, by virtue of what is *P* using £$_1$ and not £$_2$?

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

As with the previous case, Lewis's account cannot privilege the language that is (intuitively) being used by the population in question. The problem remains (as it does in the previous case) even if we think there is semantic value change later on.

4. Counterfactual Regularities

Several counterfactual strategies might be proposed to deal with the above challenge. In this section, I survey a few and argue that they are implausible.

4.a. Stability

A natural move one might try to block the problem I have been illustrating is to consider not just whether there prevails among P a regularity of truthfulness and trust in £ but also whether that regularity *would continue* to prevail among P at all nearby worlds at which P held arbitrarily different beliefs. If it does we will say that that regularity is *stable*. We will say then that P uses £ iff:

(i′) There prevails in P at least a stable regularity of truthfulness and trust in £,

(ii′) Conditions (ii)–(vi) in §2 hold in a stable way.[19]

To illustrate why this alteration is an improvement, take again **Assassination**. Consider w_1 at which P do not hold the belief that $X = Y$ (so they are not confused) but do see X planning an assassination. In w_1 when a speaker utters (1) "Jones is an assassin," they would believe (2) *X is an assassin* but not (3) *Y is an assassin.* And hearers in the population would respond to (1) by coming to believe (2) but not (3). Given that (2) is assigned to that sentence by $£_1$ then the regularity of truthfulness and trust that prevailed in w_0 (the world of the case as presented) continues to prevail among P in w_1. If this is the case across all nearby worlds with respect to $£_1$ but not $£_2$ (and condition (ii′) is met) then the language used by P is correctly predicted to be $£_1$.

However, it is doubtful whether such stability across all nearby worlds can be secured. To see why, take **Madagascar**. Since it involves semantic change it will be helpful to relativise (i′) to a time parameter. We will say that P uses £ at t iff:

(i″) There prevails in P at t at least a stable regularity of truthfulness and trust in £,

(ii′) Conditions (ii) - (vi) in §2 hold at t in a stable way.

If a regularity of truthfulness and trust in £ prevails among P at t and continues to prevail among P at t in each nearby world at which P held arbitrarily different beliefs, we will say that that regularity is *stable* at t. But now consider the sentence (4) "Madagascar is a wonderful island." Suppose w_2 is a world at which P never held the false identity belief $X = Y$ that they hold in w_0. Because of this there never arises at w_2 a future regularity of truthfulness and trust in $£_2$ which is the language that assigns (6) *Y is a wonderful island* to (4). However, in order to fix the language used in w_0 at t' (the time at which P's language intuitively changed) we want Lewis's account to predict that the regularity of truthfulness and trust in $£_2$ that prevails in w_0 at t' at least continues to prevail in w_2 at t'. But of course we are not going to get that result at w_2. In w_2, P never got confused and their language never changed. In w_2, they never come to believe "Y is a wonderful island" when hearing "Madagascar is a wonderful island." So it's not the case that the $£_2$ regularity prevailing in w_0 at t' is stable. This gives the wrong semantic prediction.

[19]That is, they hold and would continue to hold at all nearby worlds at which P held arbitrarily different beliefs.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

4.b. Stability*

Here is another variation on *stability* that might be an improvement. Let us now say that P uses £ at t iff:

(i‴) There prevails in P at t at least a *stable** regularity of truthfulness and trust in £,

(ii′) Conditions (ii) - (vi) in §2 hold at t in a stable* way.

If a regularity of truthfulness and trust in £ prevails among P at t and continues to prevail at t at all nearby worlds at which P held arbitrarily different beliefs (*shortly before t*), we will say that that regularity is *stable** at t. The idea with this variant is to go to all nearby worlds that had the same history as w_0 all the way until shortly before t. In those worlds, shortly before t, P's beliefs are changed in some arbitrary way and, as it were, we check which regularity of truthfulness and trust (that prevails among P at w_0) continues to prevail at all those worlds. w_2 (with respect to **Madagascar**) is now excluded from our set of nearby worlds because its history up until shortly before t' is different to w_0 (i.e., P at w_2 never held the belief that $X = Y$). In the case of **Madagascar**, it is also plausible that at some worlds which had w_0's history up until shortly before t' the following is true: when P discover (shortly before t') that the identity belief $X = Y$ is false then at t' P would only be truthful and trusting in $£_2$ which is the language that assigns the Y meanings to the relevant sentences.

However, it is easy to flesh out the case of **Madagascar** in such a way that at w_5 (a more distant but still nearby world) when P discover (shortly before t') that the identity belief $X = Y$ is false, then at t' P would only be truthful and trusting in $£_1$; the language that assigns the X meanings to the "Madagascar" sentences. In other words, P in w_5 decide (shortly before t') that a) they want their usage (from that point on) to go somewhat the way it would have gone had they never held that false identity belief to begin with and b) invent a new name for Y. Therefore, in w_5 at t', they hear "Madagascar is on the African mainland," for example, and they believe *X is on the African mainland* (and not *Y is on the African mainland*). However, as before, we wanted Lewis's account to predict that P's language has changed at t' in w_0. But of course we are not going to get that result if we go to w_5 and see what regularities continue to prevail there at t'. Or take w_6. When P at w_6 discover (shortly before t') that the identity belief $X = Y$ is false they decide (shortly before t') that it is all too confusing and stop using the word "Madagascar" altogether. So, it is not the case that both of the above refined conditions are met with respect to the intuitively correct language.

4.c. Conditional Regularity

Stability and *Stability** look at what regularities prevail at *all* nearby worlds. Some of these worlds will be relatively distant to w_0 than others, but they are all close to w_0. Instead, one might try a different proposal, one that incorporates considerations having to do with what would prevail under various counterfactual proposals. One such proposal (suggested by a referee) is instructive to examine in detail. Let us say that P uses £ at t iff:

(i†) There prevails in P a regularity of truthfulness and trust in £ at t,

(i‡) If P learnt of their confusion shortly before t, a regularity of truthfulness and trust in £ would continue to prevail among them at t,

(ii″) Conditions (ii) - (vi) in §2 hold and would continue to hold at t had P learnt of their confusion shortly before t.

I present two cases that show this proposal is implausible. In each case, the proposal does not deliver the intuitively correct verdicts with respect to which language P is using at t.[20]

[20]To make vivid why Conditional Regularity is not equivalent to Stability or Stability* take the following toy case involving a coin that I have not tossed. We can easily set up a case where it's true that if I had tossed it, it would've landed heads (so the claim

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

Foolish trapeze artist. It's the world of Gotham city. The hero of Gotham city goes by the name "Batman" (as in the DC comics) but also has various nicknames like "Batsy" (so "Batman" and "Batsy" are names for X). There is a trapeze artist (Y) that poses as Batman every now and then. And because he poses as Batman people mistake him for Batman (they think X is Y). We can flesh out the case so that if the population discovers the truth (that X is not Y) and so are cleared of confusion they would reserve the name "Batman" for X and, as a way of embarrassing Y, start using "Batsy" for Y only (that's the trapeze artist).

Now take the following sentence which when uttered under confusion expresses the following proposition only.

m. "Batsy is laughing"

n. *X is laughing*

The account under consideration, however, cannot predict that this sentence has this meaning. After all, we have set things up so that the condition in (i‡) is false with respect to the intuitively correct language (the one that assigns [*n*] to [*m*]). If *P* learns of their confusion moments before *t*, they would not continue using "Batsy" as a name for X. What that means is that they would not (in that counterfactual scenario) upon uttering "Batsy is laughing" continue to believe *X is laughing*. They would instead believe *Y is laughing*.

The mask. What about other kinds of confusion? Consider a case in which the community is not confused by thinking two individuals are identical but are confused in a different way. They think two distinct individuals (X and Y) look alike when in fact they do not look alike.

Suppose "Sarah" is a name for X. One day the community sees another individual Y (originally from a different town). Y, unbeknownst to the community, is in fact wearing an impressive mask that makes her look like X. Members of *P* naturally come to believe that X looks like Y. Due to Y's apparent facial similarity to X they decide to name Y "Sarah 2." This way "Sarah" semantically refers to X and "Sarah 2" semantically refers to Y. Naturally, members of *P* come to utter the sentence in (o) which semantically expresses (p) only.

o. "Sarah looks like Sarah 2"

p. *X looks like Y*

But suppose that if *P* had learnt about their confusion (that Y is wearing a mask and does not really look like X), they would stop uttering "Sarah 2" all together. In fact, however, they are still confused in the relevant way and utter (o) to express (p). The account under consideration cannot predict that this sentence has this meaning. After all, it's not the case that (i‡) and (ii$''$) are satisfied with respect to the intuitively correct language, the one that assigns (p) to (o).[21]

So learning of a confusion (or varying the beliefs in some way) might shift the language a community is using. This is no surprise given that Lewis's account of what it is to use a language exploits the communities' pattern of beliefs. And the general issue with all the above counterfactual

if I had tossed the coin it would've landed heads *is true).* Still, it does not follow from that that in all nearby worlds in which I had tossed the coin it would've come up heads. There will be a nearby world in which it lands tails.

[21]Many more problematic cases can be constructed. Take the following two. The first is a variant of **The Mask**. Suppose there is a third person (Z) who also appears to look like X. We can setup the case so that if *P* had found that Y is not a real look-alike they would shift "Sarah 2" as a name for Z instead of Y. In the second case suppose there is a person in some village who claims to be divine or God incarnate. *P* name him "Divinus." We could easily flesh out the case so that if the community finds out that this person is a poser they would kill him and ban the name. In both cases, if the community learns of their confusion they would shift their language. But the language has not shifted. And so the convention associated with the intuitively correct language does not persist.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

proposals is that there are going to be nearby worlds in which *P* use a slightly different language and we do not want the meaning of *P*'*s* sentences to depend on the regularities that prevail in those worlds. After all, there is a sense in which the noise semantically associated with a proposition is an accident.

4.d. A Compositional Constraint

Cases of common confusion like **Assassination** and **Madagascar** are special cases of the more general problem of *entailment* which confronts Lewis's account. This more general and basic problem can be illustrated as follows. Intuitively, the language we English speakers actually use associates the proposition *Jones is home* with the sentence "Jones is home" and not the weaker proposition *someone is home* which is entailed by *Jones is home*. Still, a function from strings to meanings that associates *someone is home* with "Jones is home" and was otherwise just like English would still be one in which there prevailed among us a regularity of truthfulness and trust in it. So why are we only using the former language and not the latter? The answer is that the latter language is not compositional. That is, it does not (for example) assign the proposition *it's not the case that someone is home* to "It's not the case that Jones is home" and a function that did would not be one in which there prevailed among us a regularity of truthfulness and trust in it since speakers of English may very well assert "It's not the case that Jones is home" when they believe someone *is* home.

So quite apart from the requirement that a function from strings to meanings be a function for which there prevailed among *P* a regularity of truthfulness and trust in it, there needs to be some requirement of compositionality (although I will not pursue here exactly how that requirement is to be spelled out).[22] But while a compositional requirement blocks an entailment of *Jones is home* from being associated with "Jones is home," it does not provide a good basis for discriminating between a function that assigns the X meanings to the "Jones" sentences and a function that assigns the Y meanings to the "Jones" sentences. After all, if you believe $X = Y$, then if you believe *X is not F*, then you will also believe that *Y is not F*, and if the belief that $X = Y$ is pervasive across *P*, then if there is a regularity of truthfulness and trust in a function that associated the proposition *X is not F* with the sentence "Jones is not F," then it seems that no trouble will obviously loom on the score of truthfulness and trust for a function that associated *Y is not F* with that sentence instead.

That rules of compositionality were required for his account was already recognized by Lewis (1992), but there he wanted to maintain that these rules were only needed to privilege the correct meanings for the fragment of language that we did *not* use (because they were, say, too long or complicated—see Hawthorne (1990)) and, moreover, that these rules can simply be read-off or extrapolated from the regularities of truthfulness and trust of the small fragment of language that we do use. Lewis might then be justified in thinking that conventional meaning is determined by conventions of truthfulness and trust alone for the *whole* set of used and unused sentences. But if (as entailments like *someone is home* show) we need compositional rules to privilege the intuitively correct meanings for even the fragment that we do use, then conventional meaning (despite what Lewis says) cannot be determined by conventions of truthfulness and trust alone. Moreover, I have shown in this paper that the problem arising from cases of common confusion—special cases of the more general problem of entailments, remains even if we assume actual compositionality constraints for the whole set of used and unused sentences.

5. Gricean Procedures

I have raised a problem for Lewis, but Lewis's was an admirably clear and spelled out theory which has made it especially easy to apply in a variety of cases. However, the general theme of the problem

[22]For a relevant discussion see Lewis (1992).

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

can apply to other theories. To illustrate, I will take a look at the Gricean meta-semantic tradition, originating with Grice (1957, 1969) and further developed by authors such as Davis (2002), Neale (1992), and Schiffer (1972, 1987). As is well known, Grice (1957, 1969) gave an analysis of utterer meaning that goes roughly as follows.

> An utterer *U* means *P* by uttering *S* *iff* *U* intends in uttering *S* that:[23]
>
> (A) their audience *A* come to believe *P*,
> (B) *A* recognize this intention, and
> (C) (A) occurs on the basis of (B).

This is an analysis of *utterer meaning*. Grice however distinguished between utterer meaning and sentence meaning. And, as recognized in Neale (1992, pp. 552–553), a) Grice still aimed at an analysis of sentence meaning in terms of utterer's intentions and b) sentence meaning for Grice is a species of complete utterance-type meaning. Grice (1989, p. 129) also calls this sort of meaning *timeless meaning in an utterers idiolect* ("TM" for short). Below is a population-level version of that account which is found in Grice (1989, p. 127).

> **TM (Population Level)**. For group *G*, sentence type *S* means *P* iff at least some (many) members of group *G* have in their repertoires the procedure of uttering a token of *S* if, for some *A*, they want *A* to believe that *P*, the retention of this procedure being for them conditional on the assumption that at least some (other) members of *G* have, or have had, this procedure in their repertoires.[24]

The rough idea behind this definition seems to be that conventional meaning corresponds to an internal rule that utterers conform to and continue to conform to because they think other utterers in the community conform to it as well. The rule is roughly: If *U* wants *A* to believe *P* then utter a token of *S*.

5.a. A challenge for Gricean procedures

A similar problem can be raised for Grice, though the problem does not appear quite as clearly, since we must rely on supplementary principles governing intentions—principles that may call for further debate. Still, here is one way the challenge can be developed using the case of **Assassination**. Consider the following principle governing intentions, which carries some prima facie plausibility.

> **Intention Transfer**. If *U* intends *A* to *F*, and *U* believes that necessarily: *A Fs* iff *A Gs*, then *U* intends *A* to *G*.

Take again the following two propositions and sentence:

(1) "Jones is an assassin"
(2) *X is an assassin*
(3) *Y is an assassin*

Recall that in the case of **Assassination** the confused utterers (and audience) believe that *X* = *Y*. They also intend, by uttering (1), *A* to believe (2). If utterers also believe that necessarily: *A* believes

[23]Here, "*P*" is a variable for propositions and not populations.
[24]Grice (1989) seems to have wanted to stay neutral on what attitude the utterer wants the audience to have toward *P*. The issue I am about to discuss persists even if we move to seemingly weaker attitudes like *think* or *entertain*.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

(2) iff *A* believes (3) then, according to **Intention Transfer**, the utterers also intend, by uttering (1), *A* to believe (3).

So, it seems that members of the confused group *G* have in their repertoires the procedure of uttering a token of "Jones is an assassin" if, for some *A*, they want *A* to believe (2). But they also have in their repertoires the procedure of uttering a token of "Jones is an assassin" if, for some *A*, they want *A* to believe (3). Moreover, the retention of these two procedures is for them conditional on the assumption that at least some (other) members of *G* have, or have had, these procedures in their repertoires.

Unfortunately then, **TM–Population Level** does not allow us to privilege a unique procedure that we can associate with the unique proposition semantically expressed by (1), namely (2) only.

6. Conclusion

What is common to the cases discussed in this paper is that members of the linguistic community are confused in a particular way: they believe that two individuals are identical. Identity is a strong relation, and believing that it holds between two individuals creates some serious difficulties. For if members of a community take two individuals to be identical, then whatever they believe of one, they will believe of the other—thereby threatening Lewis's version of a conventional-based meta-semantics. And if they think two individuals are identical, then whatever they intend with respect to one, they will intend with respect to the other—thereby threatening Grice's version of a conventional-based meta-semantics. A meta-semantic theory that aims to explain why one individual, rather than the other, is the semantic value of a term will struggle in such cases, since the beliefs and intentions of the community do not distinguish between these individuals. In general, if a meta-semantic account turns on some attitude and this attitude transmits over identities, then it is likely that that meta-semantic account will be subject to similar challenges as those presented in this paper.

Acknowledgments. For helpful discussions and feedback, I would like to thank John Hawthorne, Kyle Blumberg, Antony Eagle, Gillian Russell, Dmitri Gallow as well as audiences at the 2024 AAP in Perth. Thanks also to two anonymous reviewers for helpful comments on an earlier version of this paper.

Atheer Al-Khalfa is a lecturer at the Australian Catholic University. His work lies at the intersection of the philosophy of language and the philosophy of mind.

References

Armstrong, J. (2016). Coordination, triangulation, and language use. *Inquiry: An Interdisciplinary Journal of Philosophy*, 59(1), 80–112.

Bach, K. (2006). What does it take to refer?. In E. Lepore & B. Smith (Eds.), *The Oxford handbook of philosophy of language* (pp. 516–554). Oxford University Press.

Burgess, J. P. (2014). Madagascar revisited. *Analysis*, 74(2), 195–201.

Camp, J. L. (2002). *Confusion: A study in the theory of knowledge*. Harvard University Press.

Davis, W. A. (2002). *Meaning, expression, and thought*. Cambridge University Press.

Evans, G. (1973). The causal theory of names. *Aristotelian Society Supplementary Volume*, 47(1), 187–208.

Frege, G. (1948). Sense and reference. *The Philosophical Review*, 57(3), 209–230.

Geurts, B. (2018). Convention and common ground. *Mind and Language*, 33(2), 115–129.

Gilbert, M. (2008). Social convention revisited. *Topoi*, 5–16.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66(3), 377–388.

Grice, H. P. (1969). Utterer's meaning and intentions. *The Philosophical Review*, 78(2), 147–177.

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Hawthorne, J. (1990). A note on 'languages and language'. *Australasian Journal of Philosophy*, 68(1), 116–118.

Keiser, J. (2023). Languages and language use. *Philosophy and Phenomenological Research*, 107(2), 357–376.

Kölbel, M. (1998). Lewis, language, lust and lies. *Inquiry: An Interdisciplinary Journal of Philosophy*, 41(3), 301–315.

Kripke, S. A. (1977). Speaker's reference and semantic reference. In A. Peter Jr., T. E. Uehling, & H. K. Wettstein (Eds.), *Studies in the philosophy of language* (pp. 255–296). University of Minnesota Press.

Lewis, D. (1983). *Philosophical papers volume I*. Oxford University Press.

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press

Lewis, D. K. (1992). Meaning without use: Reply to Hawthorne. *Australasian Journal of Philosophy*, 70(1), 106–110.
Lewis, D. K. (1999). *Papers in metaphysics and epistemology: Volume 2*. Cambridge University Press.
Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy*, 15(5), 509–559.
Oliver, R. (1977). *The Cambridge history of Africa: Volume 3*. Cambridge University Press.
Ravenscroft, I. (2015). The natural origins of convention. *Philosophia*, 43(3), 731–739.
Schiffer, S. R. (1972). *Meaning*. Oxford University Press.
Schiffer, S. R. (1987). *Remnants of meaning*. MIT Press.
Yalcin, S. (2014). Semantics and Metasemantics in the context of generative grammar. In A. Burgess & B. Sherman (Eds.), *Metasemantics: New essays on the foundations of meaning* (pp. 17–54). Oxford University Press.

Cite this article: Al-Khalfa, A. 2025. Confusion and Convention. *Canadian Journal of Philosophy* 55: 363–378, doi:10.1017/can.2026.10063

https://doi.org/10.1017/can.2026.10063 Published online by Cambridge University Press