Topoi (2023) 42:1001–1016
https://doi.org/10.1007/s11245-023-09910-9

Metalinguistic Negotiation, Speaker Error, and Charity

Pedro Abreu[1]

Accepted: 17 March 2023 / Published online: 25 April 2023

Abstract
This paper raises a new form of speaker error objection to the analysis of disputes as metalinguistic negotiations in cases in which disputants reject that analysis. It focuses on an obvious but underexplored form of speaker error: speakers' misattribution of contents both to others and to themselves. It argues that the analyses of disputes that posit this type of speaker error are uncharitable in three different ways: first, by portraying speakers as mistaken interpreters of their interlocutors; second, by portraying speakers as uncharitable interpreters of their interlocutors; third, by portraying speakers who retract their claims as mistaken interpreters of their own prior utterances. Taken together, these unfavorable consequences weigh significantly against the plausibility of this type of analysis for the cases in question.

Keywords Metalinguistic negotiation · Speaker error · The principle of charity · Disagreement over meaning · Disagreement

1 Introduction

In various papers (Plunkett and Sundell 2013a, b, 2021a, b, forthcoming; Plunkett 2015), David Plunkett and Timothy Sundell (P&S) have proposed a new category in which to fit, and with which to make sense of, an important number of disputes, that of *metalinguistic negotiation*. Allowing for the successful reconciliation of two apparently conflicting intuitions—the presence of both linguistic divergences and of genuine disagreement between disputants—their influential approach maintains that many substantive controversies about non-linguistic and non-representational matters are often mediated through conflict over how certain linguistic expressions ought to be used. These proxy conflicts, however, are not explicitly articulated but carried out metalinguistically—that is, the opposing proposals for linguistic uses are not described but deployed, rendered manifest in conflicting uses of the same expressions; hence the misleading appearance of immediate object-level disagreement.

The following dispute is a paradigmatic candidate to this category of metalinguistic negotiation:

A: Waterboarding is torture.

B: No, waterboarding is not torture.

According to this line of analysis, the two parties in the dispute genuinely disagree, but their disagreement is not about the object-level matters described by the utterances (let us assume that the speakers happen to agree on all the basic details of the waterboarding procedure and its effect on victims). Instead, their disagreement concerns the meaning and appropriate use of the central term in the exchange, 'torture'.

This disagreement over meaning is expressed not by explicit reference to the term or to its disputed meaning but by the interlocutors' divergent use of that term, each use conforming to a different preferred meaning[1]. Following P&S (2013a), we can take the relevant alternative definitions to be something like:

$\text{torture}_1 =_{\text{def}}$ any act inflicting severe suffering, physical or mental, in order to obtain information or to punish.

✉ Pedro Abreu
pedroabreu@fcsh.unl.pt

[1] ArgLab, NOVA Institute of Philosophy (IFILNOVA), Faculty of Social and Human Sciences, NOVA University of Lisbon, Lisbon, Portugal

[1] There are different available proposals concerning the mechanisms involved in metalinguistic disputes; this is a minimal description that is intended to be neutral on this matter.

torture$_2$=$_{def}$ any act inflicting severe suffering, physical or mental, in order to obtain information or to punish, rising to the level of death, organ failure, or the permanent impairment of a significant body function.

By ascribing these alternative meanings to A and B, we can explain the apparent conflict in their uses even in the face of agreement on the relevant facts. At the same time, we're not forced to relinquish the intuition that there is indeed some form of genuine and important disagreement at work in the example. We can take this disagreement to concern which definition or pattern of use should be adopted.

P&S's analysis of disputes as metalinguistic negotiations—or P&S's *Metalinguistic Negotiation Analysis* (MNA), for short—belongs to an identifiable tradition of reinterpretative analysis. Various authors have rejected the literalness of certain disputes and in one way or another have tried to account for apparent conflicts in terms of linguistic divergence. Often, emphasis is placed on the *verbalness* of the dispute—that is, on the fact that speakers fail to align regarding the meaning associated with some problematic term or expression used in their exchange (e.g. Locke 1979, James 1907, Carnap 1963, Sidelle 2007, Hirsch 2005, 2009, Chalmers 2011). Other times, the significance of different semantic options is recognized and elaborated on (e.g. Carnap 1950, Schiappa 2003, Plunkett and Sundell 2013a, Ludlow 2014, Thomason 2017, Belleri 2017). P&S's MNA constitutes a particularly complete and sophisticated elaboration of this reinterpretative strategy. Open in its scope, non-dismissive of intuitions of genuine disagreement, nuanced about the possible motives underlying meaning negotiations, and as theoretically neutral as possible about the nature of the relevant linguistic facts and mechanisms, the MNA lends itself to many fruitful applications and promises to make sense of many disputes.

Despite its attractive attributes, however, the MNA has met obstacles and resistance on various fronts. One prominent challenge to the approach questions its applicability to cases on the grounds of ordinary speakers' resistance to how the metalinguistic account portrays the disputes in which they take part. Let us generally refer to this family of challenges as '*speaker error objections*'. This potential source of trouble has been elaborated and discussed by both advocates and critics of the merits of the MNA from the very beginning (Plunkett and Sundell 2013a, pp 23–4; see also Cappelen 2018, pp 173–9, and Odrowaz-Sypniewska's contribution to this collection, forthcoming). More recently, P&S (2021a) have assembled an article-length reflection on the topic, a defense against what they "take to be the best versions of an objection that [their] view involves an unacceptable attribution of false beliefs to ordinary speakers" (Plunkett and Sundell 2021a, p. 142). In this paper, I explore a particular version of the speaker error objection. It poses new and important difficulties for the MNA—difficulties for which I have found no response or remedies in the literature.

In Sect. 2, I introduce the topic of speakers' resistance to the MNA. Sect. 2.1 presents the type of speaker error that is central to this paper, *the Crucial Type of Speaker Error*. Sects. 2.2 and 2.3 distinguish this type of speaker error from two other types of error considered in detail by P&S (2021a).

Section 3 explains how the Crucial Type of Speaker Error constitutes an obstacle to the MNA of disputes. A necessary condition for the adequacy of the MNA is that the disputants in the relevant dispute associate different meanings with some problematic term that plays a pivotal role in that dispute. Section 3.1. delineates a method for determining whether speakers' divergent uses of some term should indeed be accounted for in terms of different associated meanings: charitable interpretation. The following subsections present three challenges to the charitable character of MNA in contexts where the speakers reject such an analysis of their dispute: Sect. 3.2 denounces the counting of speakers as mistaken interpreters of their interlocutors; Sect. 3.3, denounces the counting of speakers as uncharitable interpreters of their interlocutors; Sect. 3.4, denounces the counting of speakers as mistaken interpreters of their own prior thought and utterances.

2 Speaker Error

It is widely accepted (either as a fact or as highly plausible) that, in many cases the speakers involved in (alleged) metalinguistic negotiations reject the MNA as an analysis of their case—or would reject it if such an analysis were proposed to them. Presumably, and returning to our initial example, speakers (would) express their protest with something along the lines of:

> That's not a correct analysis of what is going on in our discussion. We're not misunderstanding each other; we speak the same language, and our words carry the same meaning. We simply disagree about what is and what is not torture. We disagree about whether waterboarding is torture, not about how the word 'torture' should be applied. We care and are concerned with torture and things in the world, not with 'torture', language, meanings, or concepts. We're trying to figure out what is and what is not torture, not how our words should be used; how our words should be used follows from our object-level discoveries, not the other way around.

This seems to me to be a credible reaction, one that could be expected in many cases, but I'll not be arguing for the

likelihood of such a response here. In what follows (with the exception of Sect. 4), I will take this for granted as an unquestioned premise. Let us call speakers who (are disposed to) express resistance to the MNA '*reluctant speakers*'.

Any conflict of the kind just illustrated between the MNA and speakers' own views of the exchanges in which they're engaged constitutes a potential basis on which to raise a speaker error objection. At its most generic, this is the "objection that [the MNA] involves an unacceptable attribution of false beliefs to ordinary speakers" (Plunkett and Sundell 2021a, p. 144). Beyond that, different types of error and different explorations of their problematic potential introduce important distinctions between more specific challenges. The next three subsections present three different types of speaker error.

2.1 The Crucial Type of Speaker Error

I start with the type of speaker belief/error[2] (rather tendentiously named) that takes center stage in this paper:

> *The Crucial Type of Speaker Belief/Error*: Speakers take[3] their words to mean the same thing in disputes where, according to the MNA, they actually mean different things.

Two other important errors follow from this one:

> *First Consequence*: Speakers take their utterances to literally express conflicting object-level propositions.
>
> *Second Consequence*: Speakers take themselves to disagree about those propositions.

Speakers believe that their utterances express conflicting object-level propositions because they take their shared words to mean the same thing. The second consequence follows from the first, together with its being a shared assumption among speakers that speakers believe what they affirm. These two consequences are rooted in the same fundamental error. They follow from it so closely that they are also at stake in the various subsequent discussions and arguments about the Crucial Type of Speaker Error.

The identification of *types* of beliefs/errors demands the kind of generic and abstract formulation used in the three characterizations just offered. To be clear, however, the *actual instances* that they are meant to capture are not generic and abstract beliefs/errors about unspecified words, propositions, and disagreements, but rather specific beliefs about specific words, propositions, and disagreements. To illustrate this, returning to the waterboarding example from above, here are the three beliefs/errors held by each disputant.

Cr_A. A takes both A and B to mean $torture_A$ by 'torture'.

Some further clarifications are in order. '$torture_A$'[4] refers to whatever meaning A happens to associate with 'torture'. At this stage, no further characterization of the meaning in question is advanced. In particular, $torture_A$ is not being identified with $torture_1$, $torture_2$, or $torture_0$, as defined in various places of the text. The only fixed determination is that it is the meaning that A associates with the term; '$torture_A$' allows us to pick that meaning and keep hold of it in subsequent ascriptions, but nothing more is said about it.

Of course, A herself has a much firmer and fuller grasp of what that meaning is. She grasps it in whatever way speakers grasp the meaning of their own words. This is the meaning that she takes her own uses of the term to carry; this is the meaning that she believes herself to mean by 'torture'. Accordingly, part of what Cr_A affirms is simply that A believes her own uses of 'torture' to carry this meaning, $torture_A$, that A herself associates with the term. Besides this very trivial bit, Cr_A adds that A also believes B's uses of 'torture' to carry that same meaning. That is, A interprets her and B's uses of that term in the same way; she takes both to mean $torture_A$ by 'torture'.

[2] Sometimes, I need to refer to the relevant beliefs without assuming that they are mistaken. Other times, I need to describe those beliefs as errors, or alleged errors. Other times still, it is useful to render both properties—*being a belief* and *being an (alleged) error*—simultaneously salient, and in such cases I use 'belief/error'.

[3] I use 'take' in this formulation because of its versatility and ambivalence in the context of interpretations or meaning ascriptions. The beliefs that matter for the Crucial Type of Speaker Error are what we might describe as interpretation beliefs, beliefs about what some speaker means by (her use of) a certain term or sentence:

-X believes Y's term/sentence to mean *such and such*.

These are beliefs of a very common and mostly silent type. In particular, I don't have in mind (only) explicit, articulated beliefs involving the formulation of definition or rule for the application of the term. Most often, the relevant beliefs are just an indispensable but inconspicuous counterpart of ordinary episodes of interpretation or understanding. Without wishing to get any further into the details of the relation between interpretations and corresponding beliefs, I'm assuming the following:

-If X interprets/understands Y's term/sentence to mean *such and such*, then X believes Y's term/sentence to mean *such and such.*

Having clarified the implicit, lightweight nature of the beliefs in question, I'm taking this assumption to be quite uncontroversial.

By using 'take', I intend to preserve the ambiguity between 'believe' and 'interprets', thus ascribing both a belief and an interpretation with a single ascription:

-X takes Y's term/sentence to mean *such and such.*

[4] In general, for any speaker X, $torture_X$ corresponds to whatever meaning X happens to associate with 'torture'.

Now, taken for granted that A also takes them to share the meanings of the other terms exchanged in the dispute—divergence over these meanings is not in question in the story—, from Cr_A it follows:

$1C_A$. A takes her utterance to express the (object-level) proposition *waterboarding is torture*$_A$ and takes B's utterance to express the proposition *waterboarding is not torture*$_A$.

Finally, since A takes both utterances to be the honest expression of their authors' thoughts, from $1C_A$ it follows:

$2C_A$. A takes A and B to disagree over whether waterboarding is torture$_A$.

Naturally, the symmetric is also true of B:

Cr_B. B takes both A and B to mean *torture*$_B$ by 'torture'.

$1C_B$. B takes her utterance to express the (object-level) proposition waterboarding is torture$_B$ and takes A's utterance to express the proposition waterboarding is not torture$_B$.

$2C_B$. B takes A and B to disagree over whether waterboarding is torture$_B$.

We need two sets of ascriptions because, according to the MNA, A's and B's utterances express different propositions. According to the MNA, all six ascribed beliefs/interpretations are mistaken: Cr_A is mistaken because 'torture' means *torture*$_A$ when uttered by A and *torture*$_B$ when uttered by B and these are different meanings; they correspond, respectively, to *torture*$_1$ and *torture*$_2$. $1C_A$ is mistaken because B's utterance expresses a different and non-conflicting (object-level) proposition, namely that *waterboarding is not torture*$_B$. $2C_A$ is mistaken because there is actually no disagreement over whether waterboarding is torture$_A$, just a misleading appearance of disagreement due to the speakers' different uses of the same term. The same explanations, with the obvious adaptations, account for the alleged incorrectness of Cr_B–$2C_B$.

These, I believe, are the especially problematic ascriptions of speaker error to which the MNA is committed. In Sect. 3, I elaborate on the threatening potential of this Crucial Type of Speaker Error and its immediate consequences; that is, I explain why I take it to raise significant difficulties for the MNA of disputes involving reluctant speakers.

First, however, in the following subsections, I consider the two types of speaker error addressed by P&S (2021a), along with their responses to potential objections built upon them. One important goal is to establish that the Crucial Type of Speaker Error is not a (clear or direct) target of P&S's attention. In preparation, and for the sake of brevity and clarity, let us use 'Type I Speaker Belief/Error' and 'Type II Speaker Belief/Error' to refer, respectively, to the first and the second type of belief/error addressed by P&S (2021a; Sects. 3.1 and 3.2).

2.2 P&S's Type I Speaker Error

According to P&S's, Type I Speaker Belief/Error consists in speakers' believing "that their utterances semantically express conflicting propositions" (Plunkett and Sundell 2021a, p. 152). In other words, it consists in speakers believing "that the claims over which they disagree are the literal semantic content of their utterances" (Plunkett and Sundell 2021a, p. 151).

There is some ambiguity in P&S's formulations. A very natural reading of these two characterizations delivers something very close to the Crucial Type of Speaker Belief/Error. They seem to express its First and Second Consequences, respectively.

> *First Reading*: Type I Speaker Belief/Error consists in speakers' believing "that their utterances semantically express conflicting [*object-level*] propositions". In other words, it consists in speakers believing "that [*what each speaker takes to be*] the claims over which speakers disagree are the literal semantic content of their utterances".

However, the adequacy of this very natural reading of passage is compromised by P&S's response to the problem in question. The response they offer to Type I Speaker Belief/Error doesn't fit the type of belief/error identified via this first reading.

P&S's response to Type I Speaker Error can be divided into three steps:

(i) First, P&S argue that the MNA is not strictly committed to the falsity of Type I Speaker Beliefs. Type I Speaker Beliefs are compatible with some versions of the MNA: those that take "the mechanism for metalinguistic usage" (P&S 2021a, p. 153) to be semantic rather than pragmatic.

(ii) Second, given (i), P&S hold that any intention to adjust our account to accommodate reluctant speakers' Type I Speaker Beliefs should lead us not to reject the MNA but simply to privilege versions of the MNA that affirm a semantic rather than a pragmatic mechanism of metalinguistic usage.

(iii) Third, P&S question the likelihood of ordinary speakers' effectively entertaining Type I Speaker Beliefs, given their highly theoretical nature; further-

> more, they argue that even if they do entertain such beliefs, there is no special reason to expect them to be correct—the ascription of error on such remote matters constitutes no great burden for the MNA.

According to the First Reading, the relevant belief/error consists in speakers mistakenly taking their utterances to express conflicting *object-level* propositions. If this reading is the right one, in their response to this type of belief/error, P&S should be addressing the speakers' alleged misidentification of conflict in the content of the propositions expressed. But that's not what they do. From the first step of their response, P&S make clear that they take Type I Speaker Belief/Error to concern not some potentially mistaken identification of object-level conflict between disputants—as the First Reading would have it—, but rather the nature of the mechanism through which that conflict is expressed. They make clear that they take Type I Speaker Belief/Error to concern "the mechanism of metalinguistic usage".

Furthermore, they also claim that it is not necessarily the case that there is a conflict between Type I Speaker Belief and MNA of the dispute. However, if the First Reading is the correct one, there is no avoiding the conclusion that there is indeed a conflict. The MNA could not but reject as mistaken speakers' identification of an object-level conflict, between disputants, on the worldly matters described by their utterances.

The two last paragraphs contain more than enough to make us hesitate about the adequacy of our First Reading of P&S's characterization of Type I Speaker Belief/Error. The challenge now is to come up with an alternative reading for P&S's characterization, one that renders their response to the problem a well suited one.

Here's that second attempt:

> *Second Reading*: Type I Speaker Belief/Error consists in speakers' believing "that their utterances semantically express conflicting [*meta-level*] propositions". In other words, it consists in speakers believing "that [*what the MNA takes to be*] the claims over which speakers disagree are the literal semantic content of their utterances".

According to this Second Reading, the disagreement and conflict mentioned in P&S's characterization of Type I Speaker Error/Belief turns out to be not a disagreement about object-level matters of the kind that is involved in The Crucial Type of Speaker Belief/Error and its consequences—a disagreement that disputants believe to be the case in virtue of taking each other's words to mean the same thing. According to this Second Reading, the disagreement and conflict in question turns out to be the very disagreement that the MNA itself affirms to be the case among disputants—i.e., a disagreement about how some central term in the dispute ought to be used. In such a reading, the potential conflict between speakers and MNA concerns not the existence of some disagreement—MNA agrees that there is a disagreement, albeit a meta-level one—but simply the way in which that disagreement is expressed, semantically or otherwise.

This is not the most natural reading of P&S's characterizations of Type I Speaker Belief/Error, but it is perhaps the one that makes best sense of their response to this type of belief/error. If the worry is that speakers might be mistaken in their supposed belief that their disagreement in conceptual ethics is semantically (rather than pragmatically) expressed, then it is perfectly opportune to note that the MNA is not committed to the pragmatic nature of the mechanism of metalinguistic negotiation; the possibility of a semantic mechanism is still on the table for P&S, and such a version of the MNA would actually render true (the second reading of) Type I Speaker Belief. Furthermore, if that is indeed the worry, P&S are again right in noting that it would be too rash to assume that such a potential error could threaten the whole project of the MNA instead of simply tipping the balance in favor of those versions that involve a semantic mechanism of metalinguistic negotiation. Finally, I also agree with P&S in rejecting the likelihood of ordinary speakers' holding such 'theoretically sophisticated' beliefs about metalinguistic mechanisms, as well as any presumption of correctness in the unlikely event of their having views on the subject.

The only (but very significant) objection to P&S's take on Type I Speaker Belief/Error is that, understood along the lines of reading two, it constitutes an unexpected choice of speaker error that is difficult to justify. There is no initial plausibility to the idea of basing a threatening objection to the MNA on it. Finally, in any case, what principally matters for the general argument of the paper is that P&S's response to Type I Speaker Belief/Error neither addresses the Crucial Type of Belief/Error nor contributes to dispelling the worries that issue from it.[5]

2.3 P&S's Type II Speaker Error

Type II Speaker Belief/Error consists in speakers mistakenly "believing that their disagreement is not, in the first instance,

[5] Alternatively, we can put it in the form of a dilemma for P&S. Either P&S's Type I Speaker Belief/Error is meant to capture something closely resembling the Crucial Type of Speaker Belief/error and its consequences (the first reading is the correct one), or it is meant to capture something else (and perhaps the second reading is the intended one). If the first is the case, then P&S's subsequent exploration of the potential threats to the MNA coming from this type of error ascription, along with their response to such threats, are inadequate. If the second is the case, P&S's have simply failed to address a very prominent and compromising, third form of speaker error.

about language and thought" (Plunkett and Sundell 2021a, p. 151). In other words, it is "the mistake of believing that their disputes do not, in the first instance, concern representational-level matters of language and thought, but rather directly address the object-level matters that are intuitively at issue" (Plunkett and Sundell 2021a, p. 158).

An important part of P&S's treatment of this form of speaker error is concerned with distinguishing it from other forms of speaker error that the MNA does not ascribe to reluctant speakers. P&S are careful to make clear that the potential conflict between the MNA and reluctant speakers' views of the disputes in this thematic vicinity is much more circumscribed than it first appears.

P&S explain that the MNA is compatible with reluctant speakers' claim that, in the relevant disputes, their disagreement goes beyond matters of language and thought and extends to object-level matters. It is even compatible with the claim that language and thought are not the most important objects of disagreement; P&S are more than ready to acknowledge that what speakers mainly care and disagree about is indeed the world and what goes on in it, not how we do or ought to think or talk about it. They stress that metalinguistic negotiations over some term typically reflect background normative disagreements on the topic of the dispute. Using the waterboarding example, and in response to Cappelen's challenge that speakers take "their debate, and their disagreement, [to be] independent of how particular words are used"— to be "about torture, not 'torture'" (Cappelen 2018, p. 175)—P&S convincingly reply that

> ... in many cases, the debate that really matters is not about the word 'torture' or about torture. It's about waterboarding, and whether we should be doing it. (And how we should treat those who engage in it, etc.) Those are the fundamental normative issues ultimately at stake, at least in many contexts of arguing "about torture". (Plunkett and Sundell 2021a, p. 161)

Because how we use language is so consequential to how we live (think, talk, reason, argue, behave, interact, shape the world), it makes perfect sense to expect important changes to be effected through changes in language. Still, the fundamental and ultimate concern in such attempted interventions is the world and how to shape it through our action. Accordingly, there is also agreement between reluctant speakers and the MNA on the driving force and final purpose of the relevant disputes.

Where the two conceptions ultimately fail to coincide is on the *immediacy* of the relevant interventions. The MNA holds that reluctant speakers are mistaken in maintaining that they "are expressing … disagreements about object-level issues directly" (Plunkett and Sundell 2021a, p. 158); these disagreements, the MNA maintains, are expressed "via the intermediate step of issues in conceptual ethics that are closely tied to (and perhaps run directly in parallel with) the first-order matters that are intuitively at issue" (Plunkett and Sundell 2021a, p. 158).

P&S don't take the ascription of this specific type of speaker error to be particularly troublesome for their MNA. They provide two reasons for this. First, they claim not to find it "terribly plausible" that "ordinary speakers of natural language possess the conceptual toolkit necessary to have beliefs … about how (object-level only, or object-level via representation-level disputes) they express their disagreements" (Plunkett and Sundell 2021a, p. 18). That is, the MNA is committed to viewing such beliefs as mistaken if they happen to be entertained, but P&S express doubts about ordinary speakers' ability to actually form such beliefs. Second, P&S affirm that, even if speakers do happen to entertain such conflicting beliefs, we can simply allow them to be mistaken. Such mistakes, they argue, are not particularly problematic, as they do not challenge speakers' competent use of their own language but simply their "folk-linguistic theory"; regarding the latter, speakers should not be presumed to be accurate.

I'm not convinced by their first point. P&S offer no clear reason to deny ordinary speakers the representational resources, or "conceptual toolkit", needed to understand and distinguish the relevant kinds of direct and indirect disagreement. The conceptual sophistication involved does not appear to be all that demanding; arguably, it requires little more than what speakers must already possess to be able to engage in routine acts of communicative repair following malapropisms, slips of the tongue, or even ambiguous usages. Furthermore, the ability to form the beliefs in question is precisely what allows speakers (correctly or incorrectly) to resist the MNA and to count as reluctant speakers. As explained above, this is my starting point, the claim that reluctant speakers do in fact exist and are perhaps even common. For now, I am simply taking this to be a shared assumption in the discussion. In Sect. 4, I briefly consider the possibility of rejecting this premise.

I also have doubts about their second point. There is more than speakers' "folk-linguistic theory" at stake in the present kind of speaker error. Speakers' fundamental competencies qua interpreters are also in question. Here, the ascription of error raises more serious concerns. We can make room for the occasional misinterpretation, but too much error compromises the plausibility of the account that ascribes it. More details about the scope of application of the MNA and about ordinary speakers' resistance to it must be factored in before we can reach a more conclusive assessment of how compromising MNA's commitment to Type II Speaker Error actually is. In Sect. 3, I return to this topic and elaborate on forms of *interpreter error* that bear more directly on the Crucial Type of Speaker Error.

Wrapping up, Sect. 2 delineated three types of speaker error. P&S divide the potential conflict between reluctant speakers and the MNA into Type I and Type II Speaker Error but in doing so fail to address (at least directly and fully) that important 'portion' of divergence that I have dubbed the Crucial Type of Speaker Error. In the next section, I elaborate on the difficulties for the MNA that issue specifically from this third type of error.

3 The Problem with the Crucial Type of Speaker Error

Let us now return to the Crucial Type of Speaker Error and the danger it poses for the MNA. Why is this type of speaker error a problem? That is, why is its ascription supposed to be a problem for any theory that incurs it? The present section answers this question.

An obvious necessary condition for the MNA's correctly describing some dispute is that the disputants in question must mean different things by the same relevant term or terms—let us call this the '*Different Meaning Hypothesis*'. Only if this hypothesis is true can it be the case that the disagreement expressed is not an object-level one but a meta-level one about which of the two meanings ought to prevail.

Opposed to this first hypothesis is a second one: the default hypothesis, the *Same Meaning Hypothesis*, which blocks the viability of the MNA and instead affirms the standard object-level character of the conflict.

It is here, when justifying choosing the Different Meaning Hypothesis over the Same Meaning Hypothesis, that the MNA reveals its vulnerability to the Crucial Type of Speaker Error. Commitment to this kind of speaker error puts significant pressure on the plausibility of the Different Meaning Hypothesis.

In what follows, I reveal how the conflict between the MNA and speakers' ordinary intuitions makes it difficult to accept that reluctant speakers are associating different meanings with the same problematic term. This argument develops as follows. First, in Sect. 3.1, charitable interpretation is presented as the most obvious choice of method for deciding between the Different Meaning Hypothesis and the Same Meaning Hypothesis. Then, in Sects. 3.2, 3.3, 3.4, three arguments are offered that show how, in three different ways, the ascription of different meanings to disputants in the face of their honest and reflected resistance to reinterpretation should be taken to weigh very significantly—perhaps decisively—against the charitable character of that ascription.

3.1 From Use to Meaning Via Charity

How to choose between the Different and the Same Meaning Hypothesis? I start with P&S's indication about what, in the first place, can start by suggesting that disputants associate different meanings with the same term. They take it to be uncontroversial that

> … at least one crucial type of data for figuring out what a speaker means by a term T are facts about the speaker's usage of T—patterns of usage that reflect her disposition to apply that term one way or another … (Plunkett and Sundell 2013a, p. 16)

Accordingly, they claim that systematically different uses of the same word, "at the very least, provide *prima facie* reason for thinking that the speakers mean different things by the word" (Plunkett and Sundell 2013a, p. 16). But, how, more exactly, is this supposed to be so? At the risk of belaboring the obvious, let us consider a few important points related to how *use* can be explored as good evidence for interpretation.

First, for use to be a valuable indication of meaning, we must assume *competent use*. Use is the complex product of at least two factors: (i) the meaning attributed to the term in question, and (ii) the thought expressed by the relevant utterance (typically, that of a longer expression in which the term is embedded). Accordingly, competent use, in the particularly important example of honest assertion, requires (i) that the speaker possess an adequate grasp of the meaning of the term in the relevant language or idiolect and (ii) that the belief that she intends to express by the utterance be true, or at least an understandable mistake. If any of these fails to hold, use may cease to constitute good evidence for interpretation.

Second, uses must be considered collectively. The interpretation of some term requires a collection of numerous uses of it, together with a background of similar collections for the various other expressions that are used together with that term. Interpretation must, to some appropriate and healthy extent, proceed holistically. Such wide and complex patterns of use cannot be expected to be exclusively competent or incompetent; actual patterns of use will unfailingly involve mistakes. These might be linguistic mistakes or epistemic mistakes. They might involve the interpretation of the central term or other terms associated in use. They might be occasional, surface, quickly retractable mistakes or deeply ingrained, systematic mistakes that are much harder to revise. It is to be expected that no actual history of uses will be entirely consistent, pulling in a single interpretative direction.

The only (rudimentary) method that I know of for navigating these complex and messy landscapes is a method of best fit: being open to alternative interpretative hypotheses, comparing them holistically, and choosing the best (or the least damaging) to the speaker's overall rationality—that is, choosing that which would render the speaker's conduct, words, and thoughts the most rational and truthful; in short, interpreting in conformity with *the Principle of Charity*.

Donald Davidson is undoubtedly the central reference when thinking about the role of Charity in interpretation. Throughout his long and rich philosophical trajectory, we encounter a thorough and intense exploration of this principle and method.[6] Davidson's interest in this notion was sparked by Quine (1960), who in turn borrowed the term 'charity' and the core idea from Wilson (1959). More recently, investigations into the Principle of Charity and charitable or rationalizing interpretation that are particularly relevant to the matters explored here include, among others, Jackman (2003, 2020), Hirsch (2005, 2009), and Schroeter and Schroeter (2014, 2015). Jackman's work on the compatibility between Charity and metasemantic externalism is especially interesting for our present case because it goes a long way towards correcting the pervasive misapprehension that Charity forces interpretation to the most obvious and immediate kind of conformity to use.

Jackman convincingly argues that, as long as we keep in mind "both that the principle's application is holistic and that many of our most central commitments are implicitly held" (2003, p. 153), we can see how charitable interpretation can naturally converge with the kinds of intuitions of stable, externally anchored, and shared contents that are usually invoked in support of metasemantic externalism. He calls our attention to central and deeply entrenched commitments held by ordinary speakers and interpreters concerning the functioning of language and linguistic communication that match this externalist picture. Even when not explicitly formulated, such commitments are nonetheless manifest in our content attributions, deference behavior, readiness to retract, and so on. Arguably, these commitments do not constitute a marginal or negligible folk-linguistics but are rather structuring and essential to our mental and cognitive lives, creating necessary conditions for tracking sameness of topic and, consequently, for learning, for the accumulation and refinement of knowledge, reasoning, conversation, and argumentation (cf. Schroeter 2012 and Schroeter & Schroeter 2014).

The centrality and importance of such commitments means that, in an important number of cases, considerations of Charity can determine that it is better (in terms of the overall rationality of the speaker) to ascribe local inaccuracy or partial ignorance or obscurity of meaning than to count the first as mistakes. That is, it would be preferable to choose an interpretation that locally fails to fit a number of uses of some term, revealing them as to some extent incompetent and inaccurate, but does not sacrifice the truthfulness of these other, more general and consequential, commitments.

Any possible increment of rationality or truthfulness must be measured globally, not locally. Once it is recognized that the comparison of interpretative hypotheses must proceed in a holistic fashion, paying due attention to the distinct weight and centrality of each belief or commitment rendered true or false by each interpretation, the possibility of alignment between charitable interpretation and metasemantic theories that run closer to reluctant speakers' intuitions of shared meanings—such as metasemantic externalism—becomes much more plausible.

Summing up, and returning to our central concern, the current proposal is that the adequacy of the Different Meaning Hypothesis, which is essential to the viability of the MNA, is to be tested by means of charitable interpretation of the two disputants. The important moral to take from the brief digression of the last few paragraphs is that Charity does not commit us to a myopic faithfulness to any partial fragment of uses, however salient or systematic. It can be perfectly in line with Charity to count a fragment of uses as mistakes so as to preserve the truthfulness and rationality of some other aspects of the speaker's mind and conduct. No breach of rationality or truthfulness should be avoided at the expense of another more serious breach of rationality or truthfulness.

With this settled, it is now time to compare the gains and losses, in terms of the overall rationality and truthfulness of the speakers, associated with each of the two alternative interpretative hypotheses:

Option I. The Different Meaning Hypothesis, which paves the way for the MNA and, *prima facie,* sticks closer to a salient set of speakers' divergent uses of some term.

Option II. The Same Meaning Hypothesis, which blocks the MNA and saves speakers from the Crucial Type of Speaker Error.

Reluctant speakers believe that their words have the same meaning. Adopting the Different Meaning Hypothesis commits us to counting such beliefs, together with the other consequences of the Crucial Type of Speaker Belief, as errors. The following subsections argue that this is so detrimental to reluctant speakers' overall rationality and competence as linguistic agents that it significantly tips the balance between

[6] The classical place to start would be essays 9–11 in Davidson (1984).

interpretations in favor of the Same Meaning Hypothesis. In other words, the next three subsections uncover three challenges to the charitableness of adopting the MNA and the Different Meaning Hypothesis in disputes involving reluctant speakers.

3.2 First Challenge: Errors in the Interpretation of Others

Our first challenge concerns the attribution of *simple interpretative error* to speakers. My claim is that endorsing the MNA against the disputants' own understanding of the dispute is in tension with the basic principle that, *ceteris paribus*, interpretations that reveal speakers to be good interpreters of their interlocutors are to be preferred to interpretations that render them bad interpreters. To illustrate this point, let us return to our waterboarding example and consider the two available interpretative options.

Option I. The Different Meaning Hypothesis.

The MNA involves the ascription to disputants of something like the alternative meanings introduced above (and repeated here for ease of access):

$\text{torture}_1 =_{\text{def}}$ any act inflicting severe suffering, physical or mental, in order to obtain information or to punish.

$\text{torture}_2 =_{\text{def}}$ any act inflicting severe suffering, physical or mental, in order to obtain information or to punish, rising to the level of death, organ failure, or the permanent impairment of a significant body function.

This option *sticks closer* to the salient portion of speakers' divergent uses of 'torture'. That is, it provides us with a good explanation of their conflicting dispositions to apply the term in the waterboarding example despite ample agreement on all the most obvious relevant facts pertaining to the practice in question.

What speaks against it—the flip side of the Different Meaning Hypothesis—is its commitment to the ascription of interpretative error to the speakers. Ordinary speakers are often good at detecting and correcting linguistic confusions, innovations, and other forms of momentary lack of coordination. The kind of persistent misunderstanding involved in the Crucial Type of Speaker Error—reluctant speakers insist on ascribing (allegedly) incorrect meanings to their interlocutors' words, even when alerted to the (allegedly) correct alternative interpretation by proponents of the MNA—is in stark contrast to other cases in which speakers are quite ready to charitably reinterpret and be reinterpreted. Such mistaken ascriptions would make a non-negligible dent in the speakers' overall truthfulness and competence as interpreters.

Option II. The Same Meaning Hypothesis.

The alternative interpretation is just the default approach: treating the disputants as meaning the same thing by 'torture' and taking the dispute at face value, as literally expressing conflicting object-level views.

The difficulty with this second option lies in finding a meaning for disputants to share that still allows us, the interpreters, to make adequate sense of their dispositions to use the term in conflicting ways despite their consensus on the basic facts of the case. This is not as difficult as it may first appear, however. Here's a first approach to a deliberately loose characterization of the relevant meaning:

$\text{torture}_0 =_{\text{def}}$ any act inflicting severe suffering, physical or mental, in order to obtain information or to punish, involving a distinctive kind of moral wrongness that is *difficult* to further analyze or articulate and *difficult* to detect.

Once we take into account factors such as the difficulty of some inquiries and of the rules and principles underlying the application of some concepts,[7] as well as speakers' frequently less than flawless, complete and transparent grasp of their own meanings and contents, the possibility of conflicting applications of the same term, even if associated with the same meaning, becomes quite acceptable.

Of course, Option II entails that at least one of the speakers is wrong. Accordingly, and unsurprisingly, the choice between Option I and Option II is a choice between interpretations that fail to deliver perfect rationality and truthfulness. Different errors must be weighed against each other, and

[7] In the present case, the difficulty might reside, for instance, in figuring out the right equilibrium between concurring and conflicting beliefs, general principles and intuitions. The judgment in question connects with numerous other topics, arguments, judgments, and concepts; mistakes (positions that fall short of optimal coherence) are to be expected within such a dense inferential network. The method of reflective equilibrium seems especially suited for such cases, and with it the expectation that, before a stable conclusion is reached, there will be occasion for the temporary endorsement of imperfect positions and flawed commitments. Alternatively, the difficulty and margin of understandable error might instead have to do with speakers' simply lacking the right kind of sensibility, an adequate 'moral vision'. Perhaps they have yet to acquire or refine this capacity through accumulated experience, discussion, and reflection. This explanation would fit, for instance, McDowell's (1998) theory of moral judgment. (To be sure, these two illustrations are not meant to exhaust all the relevant potential sources of trouble; other explanations for difficulties and error in the application of the term are of course possible.)

the considerations adduced thus far are clearly insufficient to determine which of the two mistakes—the disputants' mutual misunderstanding of one another or the misapplication of the term by at least one of them—is the most plausible and the least damaging. Fortunately, our three-pronged argument doesn't end here. In the following subsections, there is more to take into account when comparing the two interpretative options.

As inconclusive as this first challenge to the charitableness of the Different Meaning Hypothesis may be, there are two important points to take from it: first, ascribing interpretative error to ordinary speakers is an added burden to any account committed to it; second, there are other ways—ways that do not involve the ascription of different meanings—of reconciling speakers' conflicting linguistic behavior with their general agreement on the basic facts of the matter. Complexity, subtlety and obscurity of topic, concept and judgment allow for alternative explanations of such disputes—explanations that, *prima facie*, appear to do at least as good a job at rehabilitating the overall rationality and truthfulness of the disputants as the MNA account does.

3.3 Second Challenge: Speakers as Uncharitable Interpreters

The first challenge objects to counting speakers as *mistaken* interpreters of their interlocutors. The second challenge objects to counting speakers as *uncharitable* interpreters of their interlocutors.

If we accept all of the following:

(i) the Different Meaning Hypothesis, along with.
(ii) the Crucial Type of Speaker Belief, and.
(iii) common knowledge, on the part of the disputants, of the details that make it the case that the problematic term is obviously applicable according to one of the definitions and obviously non-applicable according to the other,

then we are led to conclude that the speakers are guilty of a very extreme lack of Charity in their interpretation of their interlocutors.

The MNA of some dispute commits us to the Different Meaning Hypothesis; that's how we get (i). Reluctant speakers' resistance to the MNA and reinterpretation gives us (ii). Condition (iii), common knowledge among the disputants of the basic facts of the case, is a typical part of the story in every dispute presented as a good candidate for the MNA; otherwise, were disputants to ignore important details of the case, object-level disagreement based on ignorance would be a much more plausible explanation of the dispute, rendering the MNA a much less attractive account.

To illustrate, let us apply the above to our waterboarding example. First, by accepting (i)—i.e., by accepting the MNA and the Different Meaning Hypothesis as developed in P&S's analysis of the case—we get:

1_A. A means $torture_1$ by 'torture'

1_B. B means $torture_2$ by 'torture'

Assuming that both A and B interpret themselves correctly[8], from (1_A) and (1_B) we get:

2_A. A takes A to mean $torture_1$ by 'torture'.

2_B. B takes B to mean $torture_2$ by 'torture'.

As explained above, the Crucial Type of Speaker Belief consists in each interlocutor taking both interlocutors to mean the same meaning by the relevant expression. Accordingly, accepting (ii) takes us from (2_A) and (2_B) to:

3_A. A takes B to mean $torture_1$ by 'torture'.

3_B. B takes A to mean $torture_2$ by 'torture'.

(3_A) and (3_B), together with the fact that both A and B take the other's utterance to offer a sincere expression of her views, gives us:

4_A. A takes B to claim and to believe that waterboarding is not $torture_1$. (i.e., A takes B to be claiming that the belief and claim that waterboarding in not an act inflicting severe suffering, physical or mental, in order to obtain information or to punish.)

[8] Can this assumption be questioned? In particular, can the proponent of the MNA reject that assumption? In § 3.4., I elaborate on the possibility of some speaker misinterpreting her *past* uses of some term (in cases in which there is a change of meaning between that prior moment and now) and argue that such a mistake would, to an important extent, compromise the possibility of counting that speaker's usage of the term as evidence for what she means by that term.

Presently, the question is even more serious. If an advocate of the MNA were to reject the assumption that A correctly interprets her own uses of the term, how could those uses still count as evidence for what that speaker means by the term? This move, rejecting this assumption, would discredit the very evidence that the MNA invokes to support the ascription of different meanings to the two speakers.

Furthermore, even apart from such dialectical considerations, the assumption itself seems more than compelling. It is even hard to understand what the misunderstanding of one's own present words could consist of, or how it could come about.

4_B. B takes A to claim and to believe that waterboarding is torture$_2$. (i.e., B takes A to be claiming that the belief and claim that waterboarding in an act inflicting severe suffering, physical or mental, in order to obtain information or to punish, rising to the level of death, organ failure, or the permanent impairment of a significant body function.)

However, if we also accept (iii)—i.e., if we accept that the disputants possess common knowledge of all the relevant details that make it so obviously the case that 'torture' is presently applicable according to the *torture$_1$* definition and non-applicable according to the *torture$_2$* definition—the various attributions attributed above[9] become extremely uncharitable. These attributed attributions would, in such a scenario of common knowledge of the relevant facts, render both A and B extremely uncharitable interpreters of each other.

This *uncharitableness*, in turn, would constitute a serious breach of their own rationality and truthfulness, as it makes no sense for speakers to choose an interpretation that ascribes blatant and unexplainable errors to their interlocutors. Speakers owe each other the mutual assumption of rationality and truthfulness, and our interpretation of them must reflect that; it must conform to *Charity to Charity*:[10]

> Charity to Charity: interpret in such a way as to avoid attributing to speakers (who are also interpreters) the attribution of obvious and unexplainable mistakes to others.

The same arguments (Davidson 1984, 1999; Ludwig 2004, Hirsch 2005) and the same intuitive pull behind the Principle of Charity ought also to support Charity to Charity. A speaker who recognizes herself, *qua* interpreter, as rationally compelled to seek that interpretation that best preserves the rationality and truthfulness of her interpretee must also project that obligation onto every other interpreter. This includes, of course, her interpretees. Accordingly, to presume that one's interpretees are rational is also to presume that they are charitable interpreters when they are playing that role.

The abstract considerations of the previous paragraph align well with our natural interpretative dispositions in such cases. When we think about it, it is indeed quite hard to accept that A would be willing to ascribe to B the belief and claim attributed in 4_A and that B would be willing to ascribe to A the belief and claim reported in 4_B. After all, that waterboarding inflicts severe physical or mental suffering and that it is performed to obtain information or to punish follows trivially from what is assumed to be common knowledge between A and B; that it does not (typically) rise to the level of death, organ failure, or the permanent impairment of a significant body function follows just as uncontroversially. In these circumstances, to deny the first or to affirm the second would be nothing short of absurd. But if this is so, how could the speakers, in good faith, ascribe such absurd positions to each other?[11]

Furthermore, this analysis, involving such blatant errors, fails to fit our intuitive understanding of the example and, more generally, of this type of dispute. We don't take reluctant speakers who deploy homophonic interpretations of each other to arrive at absurd claims; we take them to arrive at what they view as unexpected, intensely contestable claims, but not absurd claims. Contrast our dispute, for instance, with Grice & Strawson's (1956) famous example in which Y asserts "My neighbor's three-year-old child is an adult." Here, we have a case where a standard interpretation of Y's words effectively delivers a hopelessly absurd claim. This is not what is going on between A and B in our dispute; they're startled not by the unintelligibility of their interlocutor's claim in their shared circumstances but by what they take to be its objectionable but intelligible character.

We are thus strongly compelled to conclude that the two alternative definitions, *torture$_A$* and *torture$_B$*, fail to capture the meaning that A and B associate with 'torture'. Once this is accepted, a more standard analysis along the lines of the Same Meaning Hypothesis provided above becomes much more attractive, and indeed much harder to resist, for cases involving reluctant speakers.[12]

[9] Note that the repetitive phrasing is deliberate. The intention is to make salient that I am referring not to *the attributions of attributions* listed above—(2_A)-(4_B)—but to the *attributed attributions* mentioned in them, the attributions by A to B and by B to A.

[10] The name of the principle is meant to be in line with Hirsch's (2005) labels for the various aspects of Charity: 'charity to perception', 'charity to understanding' and 'charity to retraction'.

[11] Balcerak Jackson (2014, pp 46–7) raises a similar point, although in the slightly different context of an analysis of metaphysical discussions as merely verbal disputes. Balcerak Jackson also notes the implausibility of reinterpretative proposals that depict disputants as attributing to each other the rejection of trivial truths and the assertion of trivial falsehoods.

[12] Although to be sure, the inadequacy of *torture$_1$* and *torture$_2$* as fitting definitions for A and B does not imply that no alternative analysis involving the attribution of different meanings to speakers could be correct. However, first, it is not clear what else could work as an alternative pair of definitions. Second, any alternative pair of definitions that, to some extent, manages to avoid the present challenge will, by the same token, provide a less compelling case for reinterpretation. What, in the initial moment—i.e. before the news of speakers' reluctance—renders the original pair of alternative definitions so appealing is the obvious acceptability and unacceptability of the two utterances interpreted according to the two alternatives. This makes it the case that, in that initial moment, reinterpretation is much preferable than the ascription of obvious mistakes. However, the moment we start searching for less contrasting definitions (such that it could still count as *reasonable enough* for disputants to ascribe to their interlocutors claims that are directly opposed to theirs), we also start to attenuate the original pressure to reinterpret. If some claim is reasonable enough for some speaker to charitably ascribe it to her interlocutor, then it is also reasonable enough for any external observer to ascribe it, thus choosing the Same Meaning Hypothesis.

Importantly, analogous conclusions are equally warranted for cases of analogous disputes that are offered as exemplary candidates for the MNA. Think, for instance, of the famous Secretariat example (Ludlow 2014, Plunkett and Sundell 2013b) in which A and B, despite agreeing on all the relevant details of Secretariat's extraordinary record as a racehorse, still debate whether to apply 'athlete' to him. Once again, the alternative definitions proposed by the P&S's MNA of the case are just as antagonistically laid out. Just as in the present case, if the disputants in the Secretariat story were to resist reinterpretation and insist that they mean the same by 'athlete', a parallel argument to the one articulated in this section could be deployed to establish that P&S's MNA of that case is just as guilty of violating Charity to Charity as their MNA of the waterboarding case.

The same kind of marked discrepancy between the two alternative interpretations made available is an almost inescapable element of the setup of such disputes. That is, the two proposed alternative definitions are such as to render each of the two apparently conflicting utterances[13] either obviously true or obviously false. Such stories *need* this. What renders the Different Meaning Hypothesis initially so attractive is that the alternative on display is an obvious and unexplainable mistake. However, once we factor in the speakers' resistance to reinterpretation, the possibility of alternative charitable interpretations involving shared meanings (Sect. 3.2), and the violation of Charity to Charity (Sect. 3.3), this initial allure fades considerably.

3.4 Third Challenge: Speakers Fail to Distinguish Their Own Meanings

The third challenge objects to counting speakers as *mistaken interpreters of their own prior utterances*. At stake here is the possibility of *retraction*. I argue that adopting the MNA in cases of disputes involving reluctant speakers *in which one of the disputants retracts her initial position* commits us to taking speakers to be wrong about (the identity of) their own meanings and thoughts.

Every so often, speakers change their minds. Every so often, disputes of the kind we have been considering result in one of the speakers' actually budging and revising her attitudes towards the claims at the heart of the dispute. In our example, for instance, we can imagine that B might change her mind and converge with A in accepting and affirming 'Waterboarding is torture'.

Proponents of the MNA will naturally understand B's change of heart as a change of meaning. Each party starts by advocating a different meaning for 'torture', and A ultimately persuades B to adopt her more liberal concept. Such a turn of events must surely be counted as a possibility by any advocate of the MNA. For it to make sense for speakers to engage in metalinguistic negotiations, it must be possible, every so often and as a result of such disputes, for disputants to be moved to revise their positions in conformity with their interlocutors' proposals.

As a reluctant speaker, B will naturally take herself to have changed her beliefs about torture and waterboarding. Having initially thought that only extremely consequential actions, those "rising to the level of death, organ failure, or the permanent impairment of a significant body function", could constitute torture, she now sees that she was wrong. At first, she was convinced that it was essential to torture that such consequences be at stake, but she now realizes that this should not be taken as a necessary condition for torture. To make the example more vivid, we might imagine that B changes her position after discussing the topics of torture, justice, coercion, autonomy, punishment, and so on with A. Perhaps A manages to expose some incoherence in B's thought, and the present change corresponds to B's attempt to correct that problem and reach reflective equilibrium. Or we might imagine that A, perhaps by means of some fine rhetorical strategy, manages to convey the suffering and terror of those subjected to waterboarding so effectively as to *open B's eyes* to the presence, in waterboarding, of the same kind of moral wrongness that is to be found in other practices that B paradigmatically identifies as torture.[14]

After her change of ways, at t_2,[15] B takes herself to mean the same thing as she did before by 'torture', to be now asserting a proposition that she previously denied, and to have revised one of her prior beliefs—that is, she holds a belief of the Crucial Type with respect to her prior self, together with its two consequences:

Cr_{Bt2}. B_{t2} takes both B_{t2} and B_{t1} to mean $torture_{Bt2}$ by 'torture'.

$1C_{Bt2}$. B_{t2} takes her utterance to express the proposition *waterboarding is* $torture_{Bt2}$ and takes B_{t1}'s utterance to express the proposition *waterboarding is not* $torture_{Bt2}$.

$2C_{Bt2}$. B_{t2} takes $_{t2}$ and B_{t1} to disagree over whether waterboarding is $torture_{Bt2}$. (i.e., B takes herself to have changed her mind about whether waterboarding is $torture_{Bt2}$.)

[13] I mean the utterances that constitute the *core* of the dispute, 'p' and 'not-p'.

[14] See fn. 6 above.

[15] Here, 't_2' refers to time 2, 'B_{t2}' refers to B at t_2, and '$torture_{Bt2}$' refers to whatever meaning B happens to associate with 'torture' at t_2.

According to proponents of the MNA, these beliefs are mistaken:

Cr_{Bt2}! At t_2, B adopts a different meaning; B_{t1} and B_{t2} do not mean the same thing by 'torture'.

$1C_{Bt2}$! B_{t1}'s utterance does not express the proposition *waterboarding is not torture*$_{Bt2}$.

$2C_{Bt2}$! B has not changed her mind—she is not disagreeing with her prior self—about whether waterboarding is torture$_{Bt2}$, because what she believed at t_1 was not that waterboarding is not torture$_{Bt2}$ but rather that waterboarding is not torture$_{Bt1}$.

The trouble with this position is that, in counting Cr_{B2}–$2C_{B2}$ as mistakes on the part of the speaker, the proponent of the MNA is ascribing to B a misinterpretation of her own words, along with a misidentification of her own contents and thoughts. This option is problematically uncharitable.[16]

First, in this type of case, the MNA proponent's strategy appears to have something self-defeating about it. The whole point of reinterpreting, of ascribing different meanings to the two disputants, is to allow such ascriptions to be more faithful to the speakers' divergent uses of the same term. As explained above (Sect. 3.1), for the speakers' usage to even be counted as good evidence for an interpretation, we must know or assume that they are competent users of the terms in question and in possession of adequate knowledge of their meanings. Here, however, in taking B to be unable to distinguish her meanings and concepts at t_2 from her corresponding but supposedly different meanings at t_1, the proponent of the MNA is to some extent undermining that very competence and knowledge.

Proponents of the MNA would perhaps be willing to respond that the errors ascribed do not diminish *the kind* of linguistic competence and knowledge that matters for validating a speaker's use of the language as adequate evidence for interpretation. By differentiating competencies, they may hold on to the claim that speakers can be counted as competent users of some term even if, at a later time, they misinterpret those prior uses; they may affirm those speakers' inability to detect changes in their own meanings but insist that this only speaks against their accuracy as interpreters or folk linguists, not against their competence as speakers. P&S have responded to similar challenges in a similar way, that is, by distinguishing "linguistic judgments about how to correctly use an ordinary language term" from "folk-linguistic beliefs about the nature and extent of linguistic variation and concept sharing" that, according to P&S, "are no more likely to be accurate … than folk-scientific beliefs in any other domain" (P&S 2021a, p. 159).

I have doubts about the possibility of drawing a neat distinction between competent and meaning-revealing linguistic uses, on the one hand, and accurate beliefs about such uses and meanings, on the other, especially in the case of self-interpretation and judgments about the identity vs. the alterity of one's own meanings. In general, interpreters have no option but to rely on speakers' implicit judgment that they mean the same thing each time they repeat a term; that's what allows interpreters to count various uses of the same term, by the same speaker, as accumulated evidence for interpretation. Accordingly, speakers' beliefs about their own meanings—in particular, those concerning identity and difference of meaning, stability and change over time—cannot be set aside as a distinct and independent aspect of their complex relation to language, largely irrelevant to determining what they mean by their words. It follows that, even if not straightforwardly incoherent, the (alleged) discovery of error in a speaker's self-interpretation does indeed put pressure on the very assumptions upon which that discovery is (allegedly) made.

Second, even leaving aside the idea of internal tensions within the MNA, the error in self-interpretation that the MNA ascribes to reluctant speakers in situations of retraction appears to be sufficiently damaging to speakers' overall rationality and truthfulness to compromise the charitableness of this interpretative hypothesis. It is speakers' ability to correctly identify the sameness and difference of their own meanings and contents that is at stake. The MNA and its Different Meaning Hypothesis hold that, after retracting the relevant claim, the speaker has moved on to talk about something else, however closely related, despite the reluctant speaker's claim to the contrary.

Far from being constricted to a peripheral, unimportant, well-defined portion of our thought—what P&S refer to as our 'folk-linguistics'—beliefs of the kind in question (those exemplified by Cr_{Bt2}–$2C_{Bt2}$ above and counted as mistakes by the MNA) are absolutely central and structural to all our thought and practices. It is our ability to correctly discern identity and alterity, continuity and change of meaning and content, that allows us to know that we are thinking and talking about the same thing as before, to continue thinking and talking about it, intermittently, across time, to accumulate information and learn more about it, to reflect upon our views about it, and to revise or maintain our thoughts about it. It is also this ability to identify sameness and difference of content that allows us to reason and argue about something

[16] Jackman (2003) and Hirsch (2005) also view retraction following a dispute as a very powerful indication that the two disputants mean the same thing by a given term, that is, that considerations of Charity require a common interpretation of their words. They explore reasons other than those offered in my own analysis in support of this claim.

and to detect and explore logical relations—consistency, inconsistency, contradiction, entailment—between contents (cf. Schroeter 2012, and Schroeter & Schroeter 2014).

One way to try to reconcile the stability affirmed by the speaker (and necessary for making sense of continuity of thought and talk across contexts) with the fluidity and instability of meanings and contents identified by the MNA is to accept various notions of content with different granularities. For instance, a coarse-grained category like Cappelen's (2018) notion of *topic* or Sawyer's (2018, Sawyer 2020a, b, 2021) notion of *concept* could perhaps afford us the stability needed to allow the same agents, in different circumstances, to persevere in the same investigation or discussion, despite associating different *meanings* with the same terms.

However, there are important obstacles to such a solution. On the one hand, worries have been raised (e.g. Shields 2020) about the theoretically "inflationary" character of appealing to additional coarse-grained notions and about the potential *ad hocness* and "explanatory vacuity" of such a solution in accounting for stability across meaning changes. These are sensible worries, and a fully convincing response to them is still lacking. On the other hand, such a solution appears to make no progress on the problem of logical relations and logical reasoning across differences in meanings. Even if sameness of topic allows for some form of continuity in thought and talk, it is still insufficient to restore logical nexuses. Imagine that, at t_1, B also happens to hold (i) to be true:

(i) Torture is morally wrong.

Then, at t_2, following her discussion with A, B changes her position on the original dispute and now accepts (ii):

(ii) Waterboarding is torture.

After that, any ordinary observer of the affair would take B to also be committed to accepting (iii):

(iii) Waterboarding is morally wrong.

B herself shares this view. Being a reluctant speaker, she insists that she has revised one of her prior beliefs, *that waterboarding is not torture*, and preserved the other, *that torture is morally wrong*. Accordingly, she is now ready to accept *that waterboarding is morally wrong*. B believes that her commitment to (iii) follows logically from her prior commitment to (i) and her recent commitment to (ii).

The proponent of the MNA, on the other hand, detects an equivocation in the inference. 'torture' means different things in (i) and (ii). After having revised her meaning at t_2, it is again open to B whether to hold (i) true—after all, she now associates the term 'torture' with a different and (furthermore) more inclusive meaning; her prior belief (according to the MNA, that *torture$_2$ is morally wrong*) is different from, and does not entail, the new belief (*that torture$_1$ is morally wrong*). This form of indeterminacy about B's attitude towards the content that, after retraction, she expresses by 'Torture is morally wrong' is at odds with our ordinary understanding of the situation. Appeal to *topics* offers no help in this regard[17].

In sum, to wrap up this last challenge, in the case of disputes involving reluctant speakers and retraction, the MNA portrays speakers as mistaken interpreters of their own utterances and as unreliable judges of the identity of their own meanings and thoughts. Given the importance and centrality of the relevant beliefs in agents' thought, the problematically uncharitable character of the MNA is thus again made clear.

4 Conclusion

In this paper, I have shown that the threat posed by speaker error objections to the MNA has been neither exhausted nor neutralized in recent discussions of the subject. I have focused on an obvious but underexplored form of speaker error: speakers' misattribution of contents both to others and to themselves. I have shown how a commitment to such misattributions weighs in three different ways against the charitableness of the MNA in the case of disputes involving reluctant speakers: first, by portraying reluctant speakers as *mistaken* interpreters of their interlocutors; second, by portraying reluctant speakers as *uncharitable* interpreters of their interlocutors; and third, by portraying reluctant speakers who retract their claims as *mistaken interpreters of their own prior utterances*. Cumulatively, such discoveries tip the interpretative balance very significantly in favor of default, face-value analyses of disputes involving reluctant speakers.

What to conclude from all of this? The above seems to recommend the rejection of the MNA in the case of disputes involving reluctant speakers. Of course, given the holistic nature of Charity, the impreciseness of its guiding ideal of rationality, and the enormity and complexity of the evidential basis for interpretation, we're bound not to arrive at a definitive and undisputable verdict. This paper does not advocate the sweeping rejection of the MNA for all and any

[17] Of course, there might be other, more favourable, options available to the proponent of MNA. An anonymous referee suggested that a more sophisticated semantics, of the kind explored by Barker (2002, 2013) and Kocurek, Jerzak and Rudolph (2020), could perhaps be more suited to the preservation of the relevant logical relations by affording us stable semantic values across changes in more specific modulations.

This is a suggestive line of response to the challenges raised here, but here is not the place to pursue it, nor to exhaust other possible responses. At this point of the dialectic, it is up to the proponent of MNA to test such options and come up with worked out solutions to the problems here identified.

disputes involving reluctant speakers. It merely claims that speakers' reluctance renders the adequacy of such an analysis much more unlikely. Overall, any analysis that depicts disputants as incompetent interpreters who misunderstand their interlocutors, as implacably insisting on ascribing flagrant and unexplainable mistakes to them, and as poor interpreters of themselves, disposed to misidentify their own meanings and thoughts in (actual or potential) cases of retraction, will hardly be convincing.

Much surely depends on the degree and quality of the reluctance exhibited by speakers. As made clear above, I have been working on the assumption that there are reluctant speakers of the right kind, lucid and honest in their rejection of the analysis, and that they are not especially rare. As hinted above, this assumption can, of course, be challenged. More than that, this paper's central argument is susceptible to being turned against the hypothesis of there being reluctant speakers of the right kind. The bulk of what we have covered is aimed at establishing something like the following conditional:

> If speakers instantiate the right kind of reluctance, then considerations of Charity undermine the adequacy of the MNA.

Once we allow that the truth of its antecedent—our prior assumption—can be challenged, a *modus tollens* inference becomes just as much a possibility as a *modus ponens* one. That is, strong independent reasons in favor of the adequacy of the MNA can be used to raise doubts about the nature of the reluctance of the speakers in question.

It is well beyond the scope of this paper to try to settle whether there are indeed reluctant speakers of the right kind. It may well be that speakers' resistance to the MNA is not as pronounced as participants in the speaker error discussion appear to accept. Or it might be that there is indeed resistance, or a disposition to resist, but of the wrong kind, for the *wrong* reasons: for reasons that do not support our Charity-based argument against the adequacy of the MNA. One salient example of resistance of the wrong kind is that of speakers' rejecting the MNA for reasons of what we might call *rhetorical efficacy*: speakers might sensibly reject the MNA precisely because accepting it—accepting that it is meanings that are, in the first instance, being discussed—may invite deflationary attitudes towards the importance of the dispute or could even be detrimental to that disputant's cause (cf. Plunkett and Sundell 2013a, p. 24 and 2021a, pp. 162–4). Ultimately, the question of the quality and pervasiveness of speakers' reluctance is at least partially an empirical one, demanding close and careful case-by-case scrutiny.

As hinted above in more than one place, observations and arguments supporting the expectation that speakers exhibit the reluctance of the right type can be found in various texts in the semantic externalism tradition. I highlight two especially important examples, Burge's (1979, pp. 116 – 32) exploration of ordinary speakers' natural resistance to reinterpretation, and Schroeter and Schroeter's (2014) elucidation of powerful reasons underlying that natural resistance and its correlate, speakers' default assumption of sameness of meaning for shared terms. Here, however, is not the place to explore any further the details of any argument for or against expectations of speakers' reluctance. In this paper, I limit myself to the conceptual exercise of elucidating what follows, in terms of the adequacy of the MNA, if disputants do instantiate the right kind of reluctance. It was shown that the fulfillment of this condition renders much less plausible the analysis of the dispute as a case of metalinguistic negotiation.

Acknowledgements Many thanks to Esteban Céspedes, Cameron Domenico Kirk-Giannini, Alexander Kocurek, Marcin Lewiński, Mark Pinder, Rachel Rudolph, Andrés Soria-Ruiz and other participants and audiences at the Metalinguistic Disagreements and Semantic Externalism conference (ArgLab, IFILNOVA, Nova University of Lisbon 2022) and the VII Conference of the Brazilian Society for Analytic Philosophy (Rio de Janeiro 2022) for their attention, questions, and comments. Special thanks to David Plunkett for illuminating discussion on these topics, and to two anonymous reviewers for their careful reading and insightful suggestions.

Funding Open access funding provided by FCT|FCCN (b-on). This work is funded by portuguese national funds through the FCT—the Portuguese Foundation for Science and Technology under the Norma Transitória—DL 57/2016/CP1453/CT0087.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Balcerak Jackson B (2014) Verbal disputes and substantiveness. Erkenntnis 79(S1):31–54. https://doi.org/10.1007/s10670-013-9444-5

Barker C (2002) The dynamics of vagueness. Linguist Philos 25(1):1–36. https://doi.org/10.1023/A:1014346114955

Barker C (2013) Negotiating taste. Inquiry 56(2–3):240–257. https://doi.org/10.1080/0020174X.2013.784482

Belleri D (2017) Verbalism and metalinguistic negotiation in ontological disputes. Philos Stud 174(9):2211–2226. https://doi.org/10.1007/s11098-016-0795-z
Burge T (1979) Individualism and the mental. foundations of mind: philosophical essays, Volume 2 (1 edition). Clarendon Press, Oxford, pp 100–150
Cappelen H (2018) Fixing Language: an essay on conceptual Engineering. Oxford University Press, Oxford
Carnap R (1950) Empiricism, semantics, and ontology. Revue Int de Philosophie 4(11):20–40
Carnap R (1963) Intellectual autobiography. In: Schilpp PA (ed) The philosophy of Rudolf Carnap (First Edition). Cambridge University Press, Cambridge
Chalmers DJ (2011) Verbal disputes. Philosophical Rev 120(4):515–566. https://doi.org/10.1215/00318108-1334478
Davidson D (1984) Inquiries into truth and interpretation. Clarendon Press, Oxford
Davidson D (1999) Reply to Cutrofello. In: Hahn F (ed) The library of living philosophers: Donald Davidson. Open Court
Grice HP, Strawson PF (1956) In defense of a Dogma. Philosophical Rev 65(2):141–158. https://doi.org/10.2307/2182828
Hirsch E (2005) Physical-object ontology, verbal disputes, and common sense. Philos Phenomenol Res 70(1):67–97. https://doi.org/10.1111/j.1933-1592.2005.tb00506.x
Hirsch E (2009) Ontology and alternative languages. In: Chalmers DJ, Manley D, Wasserman R (eds) Metametaphysics: new essays on the foundations of ontology. Oxford University Press, Oxford, pp 231–258
Jackman H (1999) We live forwards but understand backwards: linguistic practices and future behavior. Pac Philos Q 80(2):157–177. https://doi.org/10.1111/1468-0114.00078
Jackman H (2003) Charity, self-interpretation, and belief. J Philosophical Res 28:143–168. https://doi.org/10.5840/jpr_2003_20
Jackman H (2005) Temporal externalism and our ordinary linguistic practices. Pac Philos Q 86(3):365–380. https://doi.org/10.1111/j.1468-0114.2005.00232.x
Jackman H (2020) Temporal externalism, conceptual continuity, meaning, and use. Inquiry 63(9–10):959–973. https://doi.org/10.1080/0020174x.2020.1805706
James W (1907) Pragmatism a new name for some old ways of thinking. Duke University Press, Durham
Kocurek AW, Jerzak E, Rudolph RE (2020) Against conventional wisdom. Philosophers' Impr 20(22):1–27
Locke J (1979) The Clarendon Edition of the Works of John Locke: An Essay concerning Human Understanding. Oxford University Press. (Original work published 1689)
Ludlow P (2014) Living words: meaning underdetermination and the dynamic lexicon. Oxford University Press, Oxford
Ludwig K (2004) Rationality, language and the principle of charity. In: Mele AR, Rawling P (eds) The Oxford handbook of rationality (1 edition). Oxford University Press, Oxford
Odrowaz-Sypniewska J (Forthcoming) Spicy, tall, and metalinguistic negotiations. Topoi
Plunkett D (2015) Which concepts should we use? Metalinguistic negotiations and the methodology of philosophy. Inquiry 58(7–8):828–874
Plunkett D, Sundell T (2013a) Disagreement and the semantics of normative and evaluative terms. Philosopher's Imprint, 13(23). http://hdl.handle.net/2027/spo.3521354.0013.023
Plunkett D, Sundell T (2013b) Dworkin's interpretivism and the pragmatics of legal disputes. Leg Theory 19(3):242–281. https://doi.org/10.1017/S1352325213000165
Plunkett D, Sundell T (2021a) Metalinguistic negotiation and speaker error. Inquiry 64(1–2):142–167. https://doi.org/10.1080/0020174X.2019.1610055
Plunkett D, Sundell T (2021b) Metalinguistic negotiation and matters of language: a response to Cappelen. Inquiry 0(0):1–25. https://doi.org/10.1080/0020174X.2021.1983456
Plunkett D, Sundell T (Forthcoming) Reflections on some varieties of metalinguistic negotiation. Topoi
Quine WVO (1960) Word and object (new edition, 2013). The MIT Press, Cambridge
Sawyer S (2018) The importance of concepts. Proc Aristot Soc 118(2):127–147. https://doi.org/10.1093/arisoc/aoy008
Sawyer S (2020) Talk and thought. In: Cappelen H, Plunkett D, Burgess A (eds) Conceptual engineering and conceptual ethics. Oxford University Press, Oxford
Sawyer S (2020) Truth and objectivity in conceptual engineering. Inquiry 63(9–10):1001–1022. https://doi.org/10.1080/0020174X.2020.1805708
Sawyer S (2021) Concept pluralism in conceptual engineering. Inquiry 0(0):1–26. https://doi.org/10.1080/0020174X.2021.1986424
Schiappa E (2003) Defining reality: definitions and the politics of meaning, 1st edn. Southern Illinois University Press, Carbondale
Schroeter L (2012) Bootstrapping our way to samesaying. Synthese 189(1):177–197. https://doi.org/10.1007/s11229-012-0099-6
Schroeter L, Schroeter F (2014) Normative concepts: a connectedness model. Philosopher's Imprint, 14(25). http://hdl.handle.net/2027/spo.3521354.0014.025
Schroeter L, Schroeter F (2015) Rationalizing self-interpretation. In: Daly C (ed) The Palgrave handbook of philosophical methods (1st edition). Palgrave Macmillan, London
Shields M (2020) Conceptual change in perspective. Inquiry 63(9–10):930–958. https://doi.org/10.1080/0020174X.2020.1805705
Sidelle A (2007) The method of verbal dispute. Philosophical Top 35(1/2):83–113. https://doi.org/10.5840/philtopics2007351/25
Thomasson AL (2017) Metaphysical disputes and metalinguistic negotiation. Analytic Philos 58(1):1–28. https://doi.org/10.1111/phib.12087
Wilson NL (1959) Substances without substrata. Rev Metaphys 12(4):521–539