Forthcoming in *Mind & Language*

Penultimate Draft

Inner Speech and Metacognition: In Search of a Connection

Peter Langland-Hassan

Abstract: Many theorists claim that inner speech is importantly linked to human metacognition (thinking about one's own thinking). However, their proposals all rely upon unworkable conceptions of the content and structure of inner speech episodes. The core problem is that they require inner speech episodes to have *both* auditory-phonological contents and propositional/semantic content. Difficulties for the views emerge when we look closely at how such contents might be integrated into one or more states or processes. The result is that, if inner speech is especially valuable to metacognition, we do not currently understand *why* it is. The paper concludes with two positive proposals for understanding the content and structure of inner speech episodes, which should serve as constraints on future accounts of the metacognitive value of inner speech.

1. Introduction

Formal estimates have it that inner speech occurs during 25% of waking life, on average (Heavey & Hurlburt, 2008)[1]. These are the internal monologues and inward remarks, the silent self-reminders and covert speech rehearsals, the 'little voice in the head' that says what it likes throughout the day.

Special thanks to Jacob Beck, Christopher Gauker, Peter Carruthers, Richard Samuels, Timothy Schroeder, Max Seeger, and Santiago Amaya for helpful discussions of this material. I am also grateful for the advice I received from audiences at the University of Cincinnati and Ohio State University, where some of this material was presented, and from two referees for this journal.

Address for correspondence: Department of Philosophy, University of Cincinnati, 2700 Campus Way, Cincinnati, OH 45221.
Email: Langland-Hassan@uc.edu

[1] Subjects tested using Hulburt's DES paradigm wear a device that beeps at random intervals throughout the day. At the sound of a beep, subjects are to write down notes about the nature of their subjective experience immediately prior to the beep (in the hopes of garnering reports that are minimally influenced by prior expectations). Other studies by Hurlburt (1993) and Klinger & Cox (1987-1988) place the average frequency of inner speech as high as 80% of waking life.

Why do we go on in this way? What cognitive role does inner speech play? Inner speech has long been implicated in research on short term memory (Baddeley, 2007), yet it is only during a fraction of its uses that inner speech serves to 'refresh' a recently encountered stimulus. To account for the many other instances of inner speech, the following sort of answer has recently gained popularity: inner or 'self-directed' speech facilitates important forms of second-order or 'metacognitive' thought. The key idea is that, by attending to our own inner speech, we are able to 'bring thoughts to consciousness,' and thereby 'objectify and contemplate our own thoughts' in some new way (Martínez-Manrique & Vicente, 2010, p. 143). This idea is sometimes advanced by those who view inner speech episodes primarily as inner *expressions* of propositional thoughts that themselves occur in some other format (Jackendoff, 1996). For the many who hold that humans do not literally think *in* a natural or 'public' language (such as English or French), the metacognitive value of inner speech promises to explain why we devote to it so much cognitive energy—why we continually express our thoughts to ourselves. For even if we do not literally think *in* a natural language, it might still be the case that we become better aware of our own thoughts when they are expressed into (what appear to be) internal utterances in a natural language.

Yet the metacognitive value of inner speech is also endorsed by some who equate it with (at least some) propositional thought itself (Frankish, 2004). Either way, it is held that there is something about the format or causal properties of inner speech that enables one to become aware of one's own propositional thoughts in a way one otherwise could not, when those thoughts either take place in, or are expressed by, episodes of inner speech. Variations on this general view have been proposed by a wide variety of theorists, including Carruthers (2011), Bermúdez (2003), Clark (1998), Jackendoff (1996), Frankish (2004), Morin (2005), Martínez-Manrique and Vicente (2010).

The feature of inner speech that makes it suitable to play this metacognitive role is typically thought to be one of three things (where these sometimes overlap). First, inner speech has sensory character and involves auditory verbal imagery; having inner speech is somewhat like hearing speech. For theorists such as Carruthers (2011), who hold that there is a special link between states with sensory character and states that are conscious (or 'globally broadcast' and 'available to mindreading'), the sensory character of inner

speech episodes forms part of the explanation of why these states—unlike 'amodal' representations—are well-suited to play a role in metacognition.

A second, closely related, view is that inner speech episodes are the only really clear cases of conscious propositional thought (where this view can be held independent of any theory about *why* they are conscious) (Frankish, 2004; Jackendoff, 1996; Bermúdez, 2003). If a state's being conscious is just a matter of one's being aware of it in some particularly direct and reliable manner, then inner speech would seem to offer the best opportunity we have to be aware of our own propositional thought processes.

Third, some have argued the symbols and sentences of natural languages have special features (e.g., they are 'context-resistant' and 'modality transcending' (Clark, 1998, p.178; see also Bermúdez (2005, p. 295)) that make them especially well-suited to serve as the objects of second-order, metacognitive thought. If inner speech involves the inner tokening of sentences in a natural language, it follows that inner speech episodes are better suited than other forms of cognition to serve as the objects of second-order, metacognitive thought.

This paper argues that these accounts of why and how inner speech aids metacognition cannot be correct as they stand, because they rely on unworkable conceptions of the representational contents and structure of inner speech episodes. With respect to the first idea—that inner speech has sensory character—it is argued that, to the extent this is true, it prevents inner speech from carrying the sort of propositional contents it must in order for it to play the envisioned metacognitive roles. With respect to the second—that inner speech is the only really obvious case of conscious propositional thought—the response is that, once we properly understand the relationship between the sensory and propositional contents of inner speech, there is no reason to think that propositional thoughts are only conscious when occurring in the form of inner speech. With respect to the third—that inner speech has important structural characteristics of representations that occur in a natural language—it is argued that the theorists in question offer no reason to think that the propositional contents associated with inner speech themselves occur in a natural language.[2] Thus, we lack grounds for holding that, when

[2] This argument echoes, while expanding upon, some points made by Machery (2005).

thinking *about* such internal representations (and their contents), we are thinking about representations that occur in a natural language. Their occurring in a natural language cannot be what explains their (supposed) role in metacognition.

These critical remarks follow from a broader positive thesis to be defended here, which is that there are only two promising ways of understanding the contents of inner speech. On the first, inner speech episodes only represent the sounds of words, and thus do not have the right kinds of contents to constitute instances or expressions of one's propositional thoughts. Call this the 'impoverished view' of the contents of inner speech. On the second viable approach, what we might intuitively mark as a single inner speech episode is in fact (at least) two distinct, co-occurring states. One of these states has a propositional semantic content, while the other of represents the sounds of words (or of a sentence). Call this the 'semantically rich' view.[3] It is argued that, once inner speech is split into two separate mental states in this way, problems arise for the above explanations of why it plays a special role in metacognition. A key part of the argument consists in showing why (on the semantically rich view) these two contents cannot be carried by a single mental state or neural vehicle—why they cannot be 'attached'.

Here is the paper's plan: section two argues that all inner speech episodes represent word sounds, at a minimum. Many of those targeted by this discussion already grant as much. However, given the importance of this point to the later argument, and my wish to make some very general points about the possible contents of inner speech, it is necessary to see *why* we should accept that all inner speech at least represents word sounds. In Section Three, I argue that there is no way to vindicate the claim that inner speech episodes *also* have propositional semantic contents, without dividing inner speech episodes into two separate mental states, in a way that causes problems for the views under discussion. Finally, in Section Four, I summarize what I see as the two (very general) viable options for understanding the content and structure of inner speech episodes, and weigh some reasons for accepting one over the other (without endorsing either). The

[3] By a "semantic" content I will simply mean the sort of content had by words (and concepts, as traditionally conceived). By a "propositional semantic" content, I will mean the sort of content had by paradigmatic propositional attitudes, such as beliefs and desires.

paper concludes that any theory that wishes to give inner speech a special metacognitive role must do so in a way that is consistent with understanding its content and structure in one of these two ways.

2. What inner speech is always about

Before beginning, it will help to introduce a piece of terminology. It is natural to describe episodes of inner speech as involving *saying* or *uttering* things. However, to describe inner speech in that way is already to suggest that episodes of inner speech are indeed cases of *speech,* and therefore carry the same propositional contents as spoken utterances. However, that is a substantive theoretical claim—one at odds with the impoverished view. So, going forward, what are typically called 'inner speech utterances,' I will call 'inner speech *episodes,*' or 'ISEs,' for short. When referring to specific ISEs, I will use the formulation, 'the ISE 'X'' in place of 'the inner speech utterance 'X''. And I will use the form, 'when one has the ISE "X"' in place of the less neutral, 'when one says, in inner speech, "X"'. So, for instance, I will use the phrase 'the ISE "Oscar is a good dog"' to refer to the mental episode one might normally, pre-theoretically, refer to as, 'saying "Oscar is a good dog" in inner speech.'

2.1. Inner speech and silent rhyme judgments

There is an obvious difference between visualizing the words 'jet engine' as written on a page (in a particular color and font) and visualizing a jet engine itself. It is a difference in what each visualization represents: some words in the first instance, or a kind of engine in the second. In both cases, shapes are represented in a fine-grained way that is distinctive of the visual modality. But, in the first, they are shapes of letters and words; and, in the second, it is the shape of a jet engine. We can think of the visualization of words as a small subset of all possible visual imagery. We can coin a term—*visuolexicalization*—to refer to this subset of all visualizations.

By analogy, there is a difference between auditorily imagining the sound of the words 'jet engine' as spoken in English, and auditorily imagining a jet engine by imagining the sound it makes. In both cases, one is representing sounds in a fine-grained way that is distinctive of the auditory modality. In the first, they are sounds of spoken words (or

phonemes); in the second, they are sounds of a jet engine. Inner speech can be thought of as using a subset of all possible auditory imagery—it involves auditory imagery of spoken sentences. It is the auditory equivalent of visuolexicalization. To hold that the auditory qualities represented by inner speech exhaust its content is to accept what I have called the impoverished view of the contents of inner speech. To hold that ISEs carry propositional semantic contents *in addition to* these auditory-phonetic contents is to hold the semantically rich view. In this section I wish only to establish that all ISEs have auditory-phonetic contents *at a minimum.*

As already noted, several of those targeted by this discussion accept this minimal claim about inner speech. Yet, while it seems intuitive to say that the 'little voice in the head' typically has some kind of sensory component to it, one might reasonably ask whether it *always* does. If it does not, then accounts of inner speech's metacognitive role need not always take such contents into account. Here I want to offer some (non-phenomenological) reasons for thinking that inner speech invariably carries such contents, and that they must therefore be countenanced by any theory of its metacognitive role.

First, inner speech is very useful in judging rhymes, and in similar auditory-phonetic tasks (such as judging homophones). Rhyming is a similarity relation between phonemes, where phonemes are the basic units of sound by which words are aurally discriminated. A standard test among psychologists for whether someone's inner speech is intact is to present them with visually dissimilar word pairs—e.g., 'ate' and 'eight'—and ask them to (silently) judge whether they rhyme (Feinberg, Gonzalez Rothi, & Heliman, 1986; Geva, Bennett, Warburton, & Patterson, 2011). Such questions are easily answered through the use of inner speech, even when the pairs involve non-words that the subject has never before seen. Some psychologists go so far as to *define* inner speech operationally as the capacity to silently make such judgments (Levine, 1982; Feinberg et al. 1985, p. 592). While that sort of definition begs important questions, it nevertheless reveals the very close pre-theoretical link that exists between the capacity to silently judge rhymes and inner speech. On the assumption that a mental state allows one to think about what it represents, the facilitation of rhyme judgments by inner speech can be explained by the hypothesis that ISEs represent word sounds.

However, one might object that the usefulness of ISEs in making judgments of these kinds does not yet weigh in favor of ISEs *representing* sounds. Compare: two written words on a page—'pen' and 'filibuster'—can be used, in some sense of 'used', to help one judge which word is longer. It does not follow that either written word represents its own length. 'Pen' represents a pen, and 'filibuster' a filibuster.

But this misunderstands the form of argument. In the case of 'pen' and 'filibuster', positing that each written word represents its own length does not help explain how the determination of relative length is made. For the representational properties of written words are not things we literally see. So, even if those words did represent their own lengths, that would not explain how we determine their relative lengths just by looking at them. Further, there is another obvious explanation available: vision provides information about the relative sizes of nearby objects, and we can easily *see* the words (and can see that they are written using fonts of the same type and size).

On the other hand, it is a foundational tenet of the representational theory of mind—and much of cognitive science—that being in a mental state with a particular representational content allows one to think about the objects and properties represented. Thus, if ISEs represented sounds, this *would* explain their facilitation of rhyme judgments, which are judgments *about* sounds. At the same time, there is no other obvious explanation for how such judgments are being made.

Nevertheless, a related objection might seem to show that the argument over-generalizes. Inner speech can clearly be used, in *some* sense of 'used', to think and reason about all kinds of non-phonetic matters. Baddeley (2007) places inner speech (via the 'phonological loop') among the core components of working memory. For instance, to take an intuitive example, one might recite in inner speech a shopping list—'bananas, coffee, bread'—in order to remember to get bananas, coffee, and bread. Supposing we successfully *use* this inner speech to remember these items, should we not, by parity of reasoning, posit that the ISE 'bananas, coffee, bread,' represents bananas, coffee, and bread? Or consider the example of silently preparing a question or objection while attending academic talk; surely those episodes of inner speech help guide one's behavior with respect to the issue at hand. Are they not, for that reason, plausibly *about* those issues, and not just about the sounds of sentences?

All else equal, this is a reasonable challenge. If a mental state X helps one successfully engage with Ys, then we have at least *some* reason to think that X represents or is 'about' Ys. But note that this does not give us reason to doubt that inner speech always represents word sounds, which is all that this section seeks to establish. Rather, it gives some reason to think that inner speech sometimes represents things *in addition to* word sounds.

Moreover, consider a comparable case involving visuolexicalization. Suppose that you had written a shopping list and dropped it somewhere along your route through the store. You might try to visualize the list in order to remember what was written on it, forming a mental image of the written word 'bananas.' This might help you remember to get bananas. Yet, clearly, in visualizing the word 'bananas' you were representing the shapes (and, likely, colors) of certain English letters. Were you, with that visualization, *also* representing bananas? Here the natural thing to say is that the visual representation of the word 'bananas' automatically triggers the *distinct* activation of the concept BANANAS (realized in some distinct mental structure), which is about bananas. In that case, the representing of bananas is not part of the visualization itself. Nevertheless, we have a clear understanding of why something (the visualization of 'bananas') that does not itself represent bananas would, in some sense, helps us engage with bananas. So the important role that inner speech plays in working memory is compatible with the impoverished view of its contents. Like visuolexicalizations, inner speech episodes may trigger conceptual thoughts without themselves bearing conceptual semantic contents.

Nevertheless, the door can be left open for the view that the activation of the concept BANANAS *together with* the visual representation the word 'bananas', is what constitutes the visualization of 'bananas'. This would make it the case that the visualization 'bananas' is itself about bananas (because the visualization is composed of these two distinct contents). Then a similar proposal could be put forward with respect to ISEs. All the same, the analogy to visuolexicalization still helps bolster the claim that all ISEs represent word sounds *at a minimum*.

2.2 Does all Inner Speech Represent Word Sounds?

Granting that *some* ISEs represent word sounds, one might nevertheless object that we still lack reason to think that the ISEs that occur when we are *not* making auditory/phonetic judgments represent word sounds. Yet there is ample evidence that, even outside of auditory phonological tasks, ISEs represent word sounds. Much of that evidence stems from work on short-term memory. In a seminal study, Conrad and Hull (1964) found that sequences of phonologically similar letters—B D T G C P—are recalled less accurately than dissimilar sequences, such as F K Y W R Q. Baddeley (1966) also found that phonetically similar sequences of words such as *man cat cap map can* are recalled correctly only about 10% of the time, whereas dissimilar sequences such as *pit day cow pen sup* are recalled correctly about 80% of the time. Interestingly, performance is only mildly affected when the sequences of words are semantically (but not phonetically) related (e.g., *huge big long wide tall*). The task instructions for these studies do not ask subjects to remember anything about the sounds of the stimulus words or letters. Yet the sounds of the words strongly affect performance. This suggests that the ISEs that facilitate such memory tasks represent word sounds even when making judgments about such sounds is not part of the task.

Further, positing ISEs with exclusively semantic contents would commit one to the existence of a class of ISEs that would *not* be useful in judging rhymes. This in turn raises a serious methodological question concerning how we might reliably pick them out, both from a first and third person perspective. How would we know when someone was engaging in inner speech of this kind? While the question is difficult, there seems at least one case where we would have good (non-question-begging) evidence for inner speech with non-phonetic characteristics. It would most likely occur in the context of a brain lesion or pathology where a person loses the capacity to silently complete the kinds of rhyme judgment tasks typically associated with inner speech (see, e.g., the people with aphasia studied by Geva *et al.* (2011)). If such people still reported being able to speak silently to themselves, or reported having 'the little voice in the head,' and so on, we would have some reason to think that inner speech lacking in phonetic content was preserved. However, there are no known reports of this kind. To the contrary, in cases of aphasia where patients lose the ability to complete the silent rhyme judgment tasks, they also

describe themselves as having lost inner speech (Levine, Calvano, & Popovics, 1982). So, while the existence of ISEs that do not (at least in part) represent sounds is possible *in principle*, the available evidence points in the opposite direction.

Nevertheless, a recent study by Oppenheim and Dell (2008) might seem to weigh against the idea that ISEs represent word sounds. Overt speech has been shown prone to two kinds of bias concerning the kinds of errors or 'slips' one makes when enunciating 'tongue twisters.' First there is a lexical bias, which is the tendency for phonological errors to create words over non-words. For instance utterances of REEF LEECH are more likely to 'slip' to LEAF REACH (both words), than utterances of WREATH LEAGUE are likely to slip to LEATH REEG (non-words) (Oppenheim & Dell, 2008, p. 529). But there is also what is known as a phonetic bias, which is the tendency to slip to a similar phoneme over a non-similar phoneme. For instance, the likelihood that REEF LEECH slips to LEAF REACH is greater than REEF BEECH slipping to BEEF REACH, because the *r* sound is more similar to the *l* sound than it is to the *b* sound (*ibid.*). In asking participants to repeat such tongue twisters in inner speech, Oppenheim and Dell sought to assess whether these same biases are present in inner speech (relying on subjective report to determine what kind of slip occurred). They found that the lexical bias was replicated in inner speech, but not the phonetic bias, and concluded that such findings 'are contrary to the hypothesis that inner speech is the product of an articulatory/acoustic system with no contact with lexical information' (p. 534).

However, this conclusion does not challenge the claim that ISEs represent phonemes. (Nor does it weigh particularly in favor of ISEs having semantic contents). First, the fact that phonetically similar words create slips *at all* in inner speech still strongly suggests that phonemes are represented by inner speech. While the phonemic bias was not found, the total number of slips reported in inner speech remained comparable to the number that occurs in outer speech (p. 534).[4] And, indeed, Oppenheim and Dell only

[4] But *why* didn't the phonemic slip occur in inner speech? Given that the number of total "inner slips" was comparable to the number of outer slips, the difference in the inner speech is that there was no *bias* in the slips such that phonemes more likely slipped to ones that were *very* similar to them. In essence, the phonetic slips were more pronounced in inner speech, in that the slips were less constrained to a narrow set of very similar phonemes. To consider just one possibility, it could be that the added articulatory steps of overt speech, involving lips and tongue, serve to reduce the severity of the phonetic slips that occur in inner speech, by adding another step where more radical errors are weeded out.

conclude that acoustic/articulatory representations 'are weakened' with respect to 'lexical representations.' Second, it is part of the present hypothesis that in representing phonemes one is, in the process, representing words. With ISEs, one thinks about words by thinking about some of their properties—namely, their phonetic properties. If that is all it is to carry 'lexical information'—if lexical information is just information about words—then the lexical bias in inner speech is consistent even with the 'impoverished' view that ISEs *only* represent word sounds. On the other hand, Oppenheim and Dell's results would weigh against the impoverished view (where, to be clear, this is an account of the *complete* contents of inner speech) if lexical information had to be *semantic* information—that is, if it were information specifically about the *meanings* of the words. But the lexical bias is not a bias to slip to a semantically related words; it is just a bias to slip to some word or other, as opposed to non-words. This very general bias is predicted by any account that, like the impoverished view, holds that it is the function of ISEs to represent words.[5]

2.3 A point about mental state individuation

Before leaving the topic of whether there are some purely semantic ISEs, it is worth reflecting on a further *theoretical* problem with that idea that there are. Someone might be tempted by the following analogy: written words, such as 'bank', may mean different things in different contexts. It seems, then, that the same thing could go for inner speech. Perhaps there is a *type* of ISE—the ISE 'justice', say—some tokens of which represent the sound of the spoken word 'justice', others of which represent justice.

Setting aside the empirical reasons already given for thinking there are no purely semantic (token) ISEs, it is important to see that the analogy to written words fails. For we can only speak of two different tokens of 'bank' meaning different things to the extent that the orthographic type 'bank' is not individuated by its representational content or functional or inferential role. 'Bank' is, we must assume, individuated as a particular kind of letter sequence (or, indeed, a set of *shapes*). By contrast, mental states *are* typically typed by their representational content or functional/inferential role (this is especially true

[5] It bears noting that there are a number of other recent results that weigh in favor of inner speech being "phonetically rich," including Oppenheim & Dell (2010), Ozdemir *et al.* (2007), and Postma and Noordanus (1996). These studies have not been singled out for discussion because they do not explicitly treat the question of whether *all* inner speech is phonetically rich.

of mental states with propositional or conceptual content). So one cannot coherently claim that two tokens of the same type of ISE have *different* contents. At least, the only way to do this would be to individuate by functional or inferential role, and hold that two tokens of this type (with their distinctive contents) have the very same functional or inferential role. But this would be to claim that radical differences in representational content make no difference in functional or inferential role.[6]

Would it help if instead we individuated mental states by their neurobiological 'vehicular' properties? Could the same neurobiological state carry two different contents, depending on the context? Yes, but the possibility offers no comfort. The neurobiological properties by which contentful mental states might be typed are not going to be *arbitrarily* related to the representational contents they carry, any more than the functional or inferential role of a mental state will be arbitrarily related to its content. By contrast, the semantic properties of natural language words are (more or less) arbitrarily related to their 'vehicular' letter sequences. That is why there is no difficulty in two tokens of the letter sequence b-a-n-k referring to different things. Yet it would be a miraculous bit of coincidence if every ISE-token that represented the sound of a word 'X' was realized by a neurological state that *also* sometimes realized a state that represented Xs (substitute 'dog' 'sky' 'heat' or whatever you like for 'X').[7]

To conclude this section, there is strong empirical support for the idea that all ISEs represent word sounds, and there are serious problems with at least one tempting, broadly

[6] While there are cases where a difference in (wide) content may make no difference in functional or inferential role (e.g., Twin Earth cases), this is not one of them. The functional role of representations of word sounds will not even approximate that of representations of things like justice—especially not within a single individual. There is no reason to think that the *narrow* content of a mental state representing justice and that of a state representing the sound of the word 'justice' are remotely similar—much less that such roles would line up for all the many ISEs one entertains.

[7] A referee raises the further possibility of appealing to relationships of polysemy. A symbol is polysemous if it has two related meanings. So, 'mole' is polysemous to the extent that it can refer to a small animal typically hidden in the ground and to a spy (who is typically hidden from view in some relevant sense). Could it be that the ISE 'justice' has two related meanings, in roughly the same way—one relating to the sound of the word 'justice', the other relating to justice itself? Here it seems the two meanings would not be closely related in the way typical of polysemy. There are not important properties that the two referents share. So it remains hard to see how a single type of neurological state could somehow play the right functional role, or have the right causal profile, for representing both.

theoretical proposal for how it might be that some tokens of a single type of ISE do not represent word sounds.

3. In Search of a Place for Semantic Content

Recall that the views to be challenged hold that inner speech is well-suited to facilitating metacognition for one of three possible reasons: the first is that ISEs are states with sensory character, which by their nature are conscious or globally broadcast (unlike amodal mental states); the second is that ISEs are the only obvious cases of conscious propositional thought; and the third is that ISEs occur in a natural language, with natural language representations having features that make them well-suited to facilitate metacognition. All three require that ISEs have propositional contents, and so are incompatible with the impoverished view (which has it that the content of ISEs is exhausted by their auditory phonological component). Considering it established (in Section 2) that all token ISEs represent word sounds, these views must find a way to hold that at least some ISEs *also* have propositional semantic content.

There are two general ways in which some token ISEs could have both kinds of content. On an 'attachment view,' it could be held that ISEs are states where propositional semantic contents are 'attached' or 'bound into' the representations of word sounds; on this sort of view—explicitly endorsed by both Carruthers (2010, 2011) and Frankish (2004)—a token ISE is a single mental state that carries two importantly different kinds of content. Alternatively, one might adopt a 'co-occurrence view,' holding that token ISEs are in fact best thought of as two (or more) distinct contemporaneous mental states that co-occur: one of the co-occurring states represents a set of word sounds, while the other has a propositional or semantic content. On this view, what we intuitively mark as a single ISE—e.g., saying 'Oscar is a good dog,' in inner speech—is in fact the co-occurrence of two (or more) *distinct* mental states, one of which represents the sound of the sentence 'Oscar is a good dog' as spoken in English, while the other represents Oscar himself, predicating of him the property of being a good dog.

The next sub-section argues that the 'attachment' approach is unworkable. There is no plausible case to be made for the idea that propositional contents are 'attached' to

representations of sounds in the way necessary for the two contents to be carried by one and the same mental state. By contrast, the co-occurrence view is a viable option. However, Sections 3.2-3.4 argue that, if the co-occurrence view is true, then the reasons theorists have offered for *why* inner speech is especially linked to metacognition cannot be correct. We still lack an understanding of why inner speech aids metacognition (if it does).

3.1 The Attachment View: Theoretical Difficulties

Carruthers (2010, p. 104) and Frankish (2004, p. 57) explicitly endorse the attachment view. They hold that ordinary ISEs are complex states possessing both phonological content and semantic content, attached or 'bound' together. Problems first arise for this view when we think again about the ways in which mental states are to be individuated.

Consider the token ISE, 'Jones deserves justice.' The attachment view has it that this ISE has two different contents: one representing the sound of a sentence, the other representing Jones, ascribing to him the property of deserving justice. If we individuate mental states by their content, it seems we are left with a very strange representation—a single state with contents that pull in different directions. An analogy would be a symbol of a language that, *whenever it occurs*, means two different things. It is akin to holding that the English phrase 'Jones deserves justice,' in all of its utterances, means both that the sentence 'Jones deserves justice' has a certain sound, and that Jones deserves justice. It is very hard to see how such a representation could figure in truth-preserving inferences. Languages, and representational (and computational) systems generally, cannot function with symbols of that kind.

Suppose instead that we individuate ISEs by functional (or inferential) role, and then ascribe representational contents to those states based on their playing a role appropriate for such contents. This would require there to be a mental state with a functional or inferential role of the kind that warrants simultaneously attributing to it two very different sets of representational properties: some relating to the sound of the sentence 'Jones deserves justice,' others relating to Jones and his deserving of justice. Yet surely there is no *single* inferential or functional role to be associated with reasoning both about the sound of that sentence and Jones's deserving justice.

The same point applies if we individuate an ISE by the neural properties of its representational vehicle. The neurobiological properties by which we individuate that vehicle will not have causal properties appropriate to identifying it *both* with thoughts about the sound of a sentence and with Jones's deserving justice. Unlike some conventional symbols (e.g. written words on a page), the intrinsic (neural) properties of the vehicle of a mental representation are not arbitrarily related to its content. Quite the opposite: the intrinsic properties of a neural vehicle determine, by their causes and effects, which ascription of content to the vehicle is appropriate. And a single type of neurobiological state will not have a causal profile appropriate for ascribing to it two very different sets of representational properties—some relating to fine-grained phonological properties, others to things like justice—simultaneously.[8]

Given these problems, why do Carruthers and Frankish think it is clearly possible for such contents to be attached? Unfortunately, neither offers many details, aside from asserting that the contents are indeed attached or 'bound into' each other. However, a style of answer not yet considered can be extracted from remarks that Carruthers makes about the nature of visual perceptual states. We can call it the 'subject-predicate view' of attachment.

3.1.1 The Subject-Predicate view of attachment Carruthers holds that by the time a visual perceptual state becomes conscious and 'globally broadcast,' various 'higher level' conceptual contents have already become attached to it. For instance, as one watches a dog chasing a ball, one enters into a visual perceptual state that 'has been partially conceptualized…coming to the mindreading system [and being globally broadcast] with the concepts *dog, chasing,* and *ball* already attached' (2010, p. 81). Whereas some would argue that visual perceptual states represent only more superficial 'perceptible' properties, such as colors, shapes, and relative positions, Carruthers offers Kosslyn's (1994) research as support for the idea that more complex conceptual contents can form 'part of the

[8] This is true even if one conceives of the auditory-phonological representation as being grounded in a physical isomorphism between the representational vehicle and its content. In that case, the structure of the neural vehicle will place significant constraints on which sensory content can be ascribed to it—constraints which will conflict with its having causal properties suitable for ascribing it an arbitrarily related semantic/propositional content as well.

perceptual state itself' and are 'globally broadcast as part of the perceptual state itself' (2010, p. 104). He then extends this idea to inner speech:

> Likewise in the case of speech (both overt and inner): the language comprehension system gets to work on the auditory input, interpreting it and attaching a content. The latter is globally broadcast along with the representation of the sounds heard. Hence we can introspect, not just the phonology of inner speech, but also its content or meaning (2010, p. 104)

Let us grant, for the sake of argument, that visual perceptual states can indeed represent properties over and above ordinary 'perceptible' or 'surface' properties—they can carry 'conceptual contents,' in Carruthers' sense, as well. There seems no insuperable theoretical barrier to such an idea. The two kinds of content might be 'attached' into a single representation in much the way that multiple predicates may be attached to a subject in an ordinary sentence (Note: in considering the idea that the contents are attached in the manner of multiple predicates to a single subject, I am moving to what I see as the most charitable *extrapolation* of Carruthers' explicit remarks). For instance, in the sentence, 'The *x* is F, G, and Y,' we can say that the contents 'F' 'G' and 'Y' are all attached or 'bound into' single representation. Thus, if the sentence, 'The dog is brown and chases a spherical ball' is in good standing, representationally speaking, then there is no reason in principle that a visual perceptual state could not carry the same sort of complex content. We have, then, a plausible way in which different contents might be 'attached' into a single representation—one which avoids the problems considered above. This may seem to open the door for a similar proposal with respect to inner speech.

However, there is an important disanalogy between the visual and inner speech cases. In the visual case, when the sensory input is 'interpreted' and 'classified', what is determined by the cognitive system is *the kind of entity* that *has* the superficial perceptible properties in question—i.e., *a dog* is classified as being brown, and *a ball* is classified as the thing that is spherical. This is the sort of 'best match' determination for which Kosslyn's work arguably provides some support. But, if we carry forward this idea of a best match to the case where auditory-phonological information is interpreted, the question answered by the cognitive system becomes: what is it that has these sonic features? To what do they correspond? And the answer will be: to thus and such spoken words or sentences of a

natural language. That is: phoneme sets and/or sounds are mapped to *words* and *sentences*. And, more generally, *whatever* properties one goes on to attach, they must be attached *to* a word or sentence—the word or sentence playing the role of subject in the predication.

Now, Carruthers' suggestion is that when the ISE 'dog' is interpreted, the content DOG is attached to it. But the cognitive system would be making a mistake if it did, since the word 'dog' does not have the property of being a dog (by contrast, dogs *do* bear the property of being brown). The problem is that, in order for auditory-phonemic properties to be properly 'attached' to a subject, the subject must be a spoken word; but then this is not the right kind of subject to which to attach the desired semantic properties. The sentence 'Brown dogs run fast' is neither brown, nor fast, nor a dog.

Bearing this in mind, we can consider another possibility. The utterance 'Jones deserves justice' has both a sound *and* a certain meaning. These are both properties *of the utterance*, let us suppose. There is no general bar against a mental state simultaneously representing both. The natural language equivalent of such a representation would be: 'The utterance 'Jones deserves justice' has thus and such a sound and means that Jones deserves justice.' Two predicates—one relating to sound, the other to a meaning—are attached to the subject, which is a particular utterance.

However, its cumbersomeness aside, this still does not get the attachment theorist what he wants. For *representing* a meaning and *having* a meaning are two very different things. The utterance 'Jones deserves justice' means that Jones deserves justice, not that a certain utterance has a certain meaning. Thus 'Jones deserves justice' is 'about' Jones deserving justice, and not about the meaning of the phrase 'Jones deserves justice.' If (by hypothesis) the ISE 'Jones deserves justice' represents the utterance 'Jones deserves justice' as having a certain meaning, then it still does not represent—still is not *about*—Jones deserving justice. So, to the extent that we had metacognitive awareness of ISEs, it would only be metacognitive awareness of thoughts about the meanings of various sentences—not thoughts about people, places, and (non-linguistic) things. ISEs would not, in fact, be capable of carrying the contents of the vast majority of our propositional thoughts. Moreover, it was never anyone's pre-theoretical view that, when we engage in inner speech, we are having thoughts that, in effect, make either true or false claims about

the meanings of sentences. The pre-theoretical view is that the ISE '*x* is F' means that *x* is F.

A referee proposes that Carruthers could respond by granting that ISEs do not in fact have the same content as our propositional thoughts, yet maintain that ISEs have contents of roughly the kind just considered, which would still allow one to infer the content of one's own propositional attitudes. Here the idea might be that the ISE 'Jones deserves justice' has the content: '*X* has thus and such a sound and is a saying that Jones deserves justice' (the referee suggests substituting 'is a saying that' for 'means that'). As with the approach just considered, the ISE would not itself mean that Jones deserves justice. Rather, it would mean that something is a saying that Jones deserves justice. However, one could potentially take the information that there was such a 'saying' into account when interpreting one's own propositional attitudes (assuming one could recognize the 'saying' as one's own).

However, this proposal has still has serious problems. First, it conflicts with what we would say about the complementary acts of auditory and visual perception. When we hear someone else say that *p,* we represent the sound of their utterance and, after decoding, its semantic content, *p*. Do we *also* represent that the auditory event was a case of someone's *saying that p*? Well we *could*, but there is no reason to think that this is inevitably a part of hearing and understanding speech. Keep in mind, it is one thing to represent, in a general way, that someone is speaking; it is quite another to represent, of each heard utterance that *p*, that it is a saying that *p*. It is the latter that this proposal needs to be the case, for it is only in that case that the contents of each ISE could be globally broadcast. Merely representing that some inner speech is occurring would not by itself allow one to know the subject matter of the ISEs. Even less plausible is the idea that this content concerning 'sayings' would be *attached* to each auditory representation of a heard sentence. For this would be akin (in the visual case) to holding that, whenever one reads a written sentence—e.g., 'he forgot the bananas'—one's visual representation of the letters has attached to it a representation that those letters are a *writing that he forgot the bananas*. This is implausible on its face, nor am I aware of any independent motivation for such a view. So there is no analogy from the auditory or visual perception of language to support the claim that inner speech has contents of the kind proposed; rather, all signs point in the opposite direction.

Finally, the idea that ISEs represent sentences as being sayings that thus and such is at odds with the phenomenology of inner speech. Recall that, on Carruthers' account, ISEs are globally broadcast and are therefore conscious and introspectable. If ISEs really have contents such as 'X has thus a such sound and is a saying that thus and such' then, when we engage in inner speech, we should be aware of thinking that each ISE is a saying that thus and such. But this is not at all what it is like to have inner speech. In having the ISE 'He forgot the bananas,' one seems to consciously entertain the proposition 'he forgot the bananas' (if any proposition at all), and not the proposition: 'X is a saying that he forgot the bananas.'

3.2. The Co-occurrence View and Carruthers

At this point it may seem that the problems of the attachment view can easily be avoided if one simply conceives of inner speech as involving the co-occurrence of two distinct states—one state representing word sounds, the other representing what those represented sounds typically represent (i.e., a propositional content). So, if the first state represents the sound of the sentence 'Bob always rides horses,' the second (distinct-but-co-occurring) state represents (i.e., means) that Bob always rides horses.

While this approach is internally consistent, it conflicts with the rationale that many have offered for why inner speech is especially well suited to aiding metacognition. Looking first at Carruthers, it is a key part of his overall theory of the mind (and of metacognition and introspection) that only mental states with sensory character are 'globally broadcast' and made available to multiple consumer systems, such as the mindreading module (2011, p. 47-55). While conceptual contents may sometimes be globally broadcast, this can only occur when such contents are 'bound into the content of any given sensory state and broadcast along with it' (2011, p. 48). In explaining how the mindreading system—which, on his theory, only has access to globally broadcast states—generates judgments about one's own propositional attitudes, Carruthers proposes that it often takes as input (from the global workspace) semantic conceptual contents that are attached, or bound into, the auditory phonological representations characteristic of inner speech (2011, p. 73-74).

For the main thesis of his 2011 book on self-knowledge is that people lack any

direct, non-interpretive form of access to their own propositional attitudes. Instead, we come to know our own beliefs, desires, and other propositional attitudes through swift acts of self-interpretation, just as we form judgments about the beliefs and desires of others through acts of interpretation. In one's own case, however, one has an increased data set for such interpretations, since one *does* have non-interpretive (or 'direct', or 'introspective') access to one's own inner speech. By attending to the contents of thoughts that get expressed into inner speech, we can come to have good *evidence for* the judgments or decisions we have recently made—even when our outward behavior provides no relevant cues (he calls this the 'Interpretive Sensory-Account' (ISA) of self-knowledge of propositional attitudes).[9]

Note again that such semantic or 'conceptual' contents are never globally broadcast *on their own*—they must always be clothed in a sensory representation of some kind, in order to gain admittance to the global workspace. For if representations without sensory character could be globally broadcast, there would be no reason why our judgments, decisions, and propositional attitudes generally could not be introspectively accessed (i.e., self-ascribed without self-interpretation). For they, too, could then be globally broadcast and made directly available to the mindreading 'consumer system.' This explains his insistence that the semantic contents of inner speech are literally attached or bound into the auditory/phonetic contents, and are not simply features of co-present but distinct mental states. Adopting a 'co-occurrence' view would leave him without a principled reason for holding that (amodal) propositional attitudes cannot be globally broadcast in the manner of perceptual states.

Thus his approach faces a dilemma: he must either find a workable version of the attachment view, or provide a theoretically motivated reason for why propositional contents can only be globally broadcast when they occur at *the same time as*

[9] While Carruthers maintains that ISEs could not constitute *any* of our propositional attitudes themselves (2011, p. 102-106), he does allow that 'one can coin a looser sense of the term 'thinking' in which episodes in inner speech *can* count as forms of thinking.' In this looser sense of the term, "for thinking to be taking place is just for one to be tokening some event…with propositional content, which plays *some* role in issuing in judgments, decisions, or other changes in attitude or action" (p. 107). It remains essential to his view that these inner speech "thoughts" themselves have sensory character—that the propositional contents are "attached" to representations of word sounds.

representations of the sounds of sentences one would use to express them.[10] Going the latter route, Carruthers would also need a reason for thinking that the globally broadcast states with propositional content could never themselves simply *be* one's propositional attitudes. Perhaps these challenges can be met; however, his work to this point offers no clear means for addressing them.

3.3 The co-occurrence view and Frankish

The co-occurrence view will be equally unsatisfying for the numerous theorists who hold that the metacognitive value—and, for some, the general cognitive value—of inner speech is importantly tied to its occurring in a natural language. Here we can take Frankish (2004) as a leading example (though see also Clark (1998), Jackendoff (1996), and Bermúdez (2003)). According to Frankish 'our conscious thoughts and judgments are often framed in natural language,' where a natural language forms 'the medium of representation' (2004, p. 23). 'Some reasoning processes,' he adds, 'constitutively involve the manipulation of natural-language sentences—written, vocalized, or, most often, articulated in inner speech' (p. 32). As he puts it, 'the linguistic activities *implement* the reasoning process' (p. 33). He is very explicit that we think *with* natural language sentences when we engage in inner speech.

Metacognition enters the picture in his account of how some of these inner speech episodes are transformed into occurrent beliefs (or 'superbeliefs,' in his terminology). 'A sentence (of inner speech),' he writes, 'acquires the causal role of a belief or desire when we endorse it and commit ourselves to taking its content as a premise or goal in our conscious reasoning' (p. 152). It is a metacognitive awareness of some episode of inner speech (and its content), combined with a decision to adopt a certain perspective or 'policy' with respect to the introspectively discerned content, that result in the inner speech episode's taking on the causal role of a belief.

[10] Carruthers could respond that *visual* imagery (or other kinds of auditory imagery) might still provide relevant clues for self-interpretation, and attempt to do without inner speech altogether. However, there is a specificity to the kinds of contents presumed to be carried by inner speech—a presumption that it can carry the same kinds of contents as natural language sentences (involving, e.g. conditionals, negation, imperceptible entities, and so on)—that it would be hard for his theory to do without. For instance, it is not clear how one could judge that that one had just formed a belief in a counterfactual conditional by taking as 'data' a series of visual images.

But note that, if the co-occurrence view is right, we no longer have any reason to hold that the reflection on inner speech episodes that Frankish and others appeal to is reflection upon representations that occur *in* a natural language. For consider again the ISE, 'Jones deserves justice.' Once we have distinguished between two mental states—one that represents the sound of the English sentence 'Jones deserves justice' (but does not itself occur in English), and another that means that Jones deserves justice, there is no reason to think that the latter occurs *in English*. The phenomenologically-grounded intuition that it does is fully explained by the fact that the state carrying that propositional content is *accompanied by* another mental state that is *about* the sound of an English sentence.[11]

In response, Frankish could conceivably just accept that the propositional contents associated with inner speech are not themselves had by internal symbols or sentences of a natural language. This would require revision of his 'cognitive conception' of language, according to which much of one's conscious propositional thought takes place in a natural language. It would no longer be true that inner speech 'constitutively involve[s] the manipulation of natural-language sentences,' where such activities 'implement the [computational] reasoning process' (p. 32-33). Further, it would not leave us with a clear picture of why inner speech is especially closely related to the conscious propositional thoughts that Frankish holds can, under the right circumstances, constitute occurrent beliefs. For it is now unclear why conscious propositional thoughts, themselves not occurring in a natural language, should ever be accompanied by representations of the sounds of sentences one would use to express them. Why should the two kinds of representation bear any special relationship? The proposed link now appears completely arbitrary.

These difficulties likely account for Frankish's endorsement of an 'attachment' view, which promises some more intimate connection between the two kinds of representation. Like Carruthers, Frankish holds that inner speech involves 'images of hearing the sentences uttered or of the vocal movements necessary to utter them,' where these 'come with

[11] Some theorists, such as Christopher Gauker (2011), offer independent arguments for thinking that human propositional thought always occurs *in* a natural language. However, Frankish is not among them (nor are the other theorists under discussion).

semantic interpretations attached' (p. 57). He does not provide further details on how we are to think of the nature of the attachment. As argued above, the most promising attachment view is the subject-predicate view. Yet this view returns the wrong result for Frankish as well. Recall that, on Frankish's view, our metacognitive acceptance or rejection of the contents of ISEs is what determines whether or not they constitute occurrent beliefs (or desires). His idea is that if one has the ISE, 'Capital punishment is impermissible,' one is then able to take a higher order perspective on that content, either accepting or rejecting the impermissibility of capital punishment. But on the subject-predicate version of the attachment view, the content of that ISE must actually be something along the lines of: 'X has thus and such a sound and is a saying (or means) that capital punishment is impermissible.' Accepting or rejecting *this* content is quite a different matter from accepting or rejecting the claim that capital punishment is impermissible. It is, instead, accepting or rejecting the idea that some putative utterance had a certain sound and was a saying (or meant) that capital punishment is impermissible. One's views on capital punishment will have little to do with one's acceptance or rejection of this content. Further, the two other arguments against this style of attachment view given at the end of Section 3.1.1 apply here as well (these were the arguments by analogy to language perception, and from phenomenology).

3.4 On the Putative Linguistic Structure of Inner Speech Episodes

The points above equally call into question some less fully developed views of the relation between inner speech and metacognition. For instance, Jackendoff (1996) holds that 'thought per se is *never* conscious.' Rather we are only ever consciously aware of the (inner) linguistic *expression* of our thoughts, when they are expressed with inner speech utterances—the 'talking voice in the head' (1996, p. 10). 'We become aware of thought taking place,' he says, 'only when it manifests itself in linguistic form' (*ibid.*). However, on the co-occurrence view (which, I am now assuming, is the only workable view on which ISEs have propositional contents), our propositional thoughts do not occur *in* a natural language. At least, the theorists in question offer no tenable reasons for thinking that they do. As discussed above, phenomenology does not support the assumption. At best, we can say that conscious propositional thought is often *accompanied by* auditory images of

sentences one might use to express those thoughts. But this conflicts with Jackendoff's claim that we are only aware of our thoughts when they occur 'in linguistic form.'

Similarly, Andy Clark endorses Jackendoff's 'key claim that linguistic formulation makes complex thoughts available to processes of mental attention' (1998, p. 178). According to Clark, in generating inner speech we create a 'special kind of mental object' that is 'apt for scrutiny from multiple different cognitive angles.' This allows for 'second-order cognition'—that is, thought about one's own thoughts (*ibid.*). Clark explains:

> By 'freezing' our own thoughts in the memorable, context-resistant and modality transcending format of a sentence we thus create a special kind of mental object – an object which is apt for scrutiny from multiple different cognitive angles…ideally suited to figure in the evaluative, critical, and tightly focused operations distinctive of second-order cognition (1998, p. 178).

Here Clark also seems to assume that inner speech involves thoughts that occur *in* a natural language—and that their doing so explains why ISEs are especially well suited to facilitate second-order (i.e., meta) cognition. Yet, having sorted the auditory-phonological component of an ISE and the propositional component into separate mental states (in line with the co-occurrence view), there is (again) no reason offered for thinking that either occurs *in* a natural language. So, whatever features natural language representations may have, they cannot serve to explain why inner speech would have any special connection to metacognition.

Finally, Martínez-Manrique & Vicente defend what they call the 'simple view' of inner speech's role in metacognition, according to which:

> We recruit the means we have to focus someone else's attention on our thoughts in order to focus our own attention on our own thoughts….Our silent speech activates the linguistic module and the pragmatics module, which probably work in tandem to extract the thought that has been communicated, i.e. the thought we try to tell ourselves (2010, p. 162).

Perhaps we do tell ourselves our own thoughts in order to become aware of them. But it is really not clear why we would. There is an obvious reason why we generate linguistic utterances to focus *others'* attention on our thoughts: our thoughts are not already in their heads. We need a way of transmitting them. But that cannot be the reason we tell

ourselves our own thoughts. There is no obvious need to transmit oneself a thought one already has. And, at any rate, if one does need a means to transmit oneself one's own thoughts, why does one not *also* need a means to transmit oneself one's own inner speech? Simply saying that that latter is conscious does not explain why the former cannot be conscious as well, especially given that the semantic and auditory/phonological contents of inner speech are not 'attached' in any way. So the analogy Martínez-Manrique & Vicente offer does not explain *why* inner speech plays a role in metacognition, if indeed it does.

4. The Content and Structure of Inner Speech: Two Possibilities

If we wish to tie inner speech in some special way to metacognition (now a considerable *if*), it seems the most we can say is that, somehow, by thinking *about* the sounds of English sentences, we are able to trigger—and perhaps gain second-order awareness of—ontologically distinct thoughts (themselves not in English[12]) that share the contents of the English sentences whose sounds are being thought about. On such a view, there is no barrier to amodal representations with propositional contents being conscious. So the question then becomes why propositional thoughts would only become conscious, or become metacognitively accessible, when accompanied by an auditory phonological representation of a sentence one might use to express it. That question may have an answer. Yet no one, to my knowledge, has offered one. Nor has the question previously been framed in this way.

When we turn to the question of how we should positively characterize the contents of inner speech (given the above), it seems we have two options. The first is the impoverished view, according to which ISEs *only* represent word sounds (even if they are often *accompanied by* propositional, amodal thoughts with related contents). The difference between this and the second, semantically rich view is in whether the accompanying state bearing propositional content is counted as a part of inner speech itself (and hence whether 'inner speech' refers to the co-occurrence of two separate mental

[12] Nothing said here *rules out* the possibility that these ontologically distinct propositional thoughts occur in English (or some other natural language). The point is that the theorists in question offer no sustainable reason for thinking that they do.

states). Faced with these two options, the question of how we ought to use the term 'inner speech'—to refer to half of the complex, or the two together—might seem insubstantial. However, terminological decisions have a way of driving inquiry, so it is wise to make such choices in a principled way when one can.

It seems one can offer good reasons for using the term 'inner speech' in each of the two ways under consideration. One reason to hold that inner speech involves both states, is that there is an important psychological difference between a person who can *speak* a language and someone who can merely aurally represent the words of that language (perhaps having memorized the sound of a phrase in some foreign language). The person who speaks the language has an understanding of that language—she has thoughts that mean what her utterances mean. If understanding what one is saying is essential to *saying* anything at all, then there is some reason to include the state in virtue of which one counts as understanding one's own utterance as a part of something called inner *speech.* This is a distinction psychology—even folk-psychology—should plausibly mark. By counting the propositional thought that *p* as a part of the ISE '*p*', one thus distinguishes the ISE 'It is raining'—which can only be had by someone who understands English—from the episode of merely representing the sound of the sentence 'It is raining,' which can be tokened by someone who speaks no English.

On the other hand, one might reject the idea that inner speech is importantly speech-like, and so feel no pull towards respecting its status as such. After all, if we have already abandoned the idea that the propositional contents associated with inner speech are carried by (internal) natural language sentences, one may as well also let go of the idea that inner speech is in fact a kind of speech (assuming that a necessary feature of speech is that it occurs *in* a public language). If all sides agree that inner speech is *not* a kind of speech occurring *in* a natural language, then there is little reason to use the term 'inner speech' to refer to something that can only be done by people who can comprehendingly speak a particular natural language. On this view, calling ISEs 'inner speech' is a misnomer; the idea that IESs are a kind of speech is grounded in a use/mention confusion that we would do well to avoid. With ISEs we think *about* natural language sentences, not *with* them; ISEs mention natural language sentences, they do not use them. Inner speech—the subjective phenomenon initially picked out as 'the little voice in the head,' or the 'silent

soliloquy'—turns out only to represent word sounds and sentences. Sometimes this auditory verbal imagery is accompanied by related propositional thoughts, sometimes not. From this perspective, there is no important reason to reserve the name 'inner speech' only for episodes where the auditory verbal imagery is so accompanied.

It is not obvious which of these two positions is ultimately preferable. It is enough for now to keep clearly in mind the considerations in favor of each, leaving both on the table as viable options. Neither, however, leaves us with a clear picture of why inner speech would be especially involved in metacognition. And each is counterintuitive in its own way. The impoverished view is unexpected in its entailment that inner speech is only ever about words and word sounds. The semantically rich view is counterintuitive in holding that what we normally might think of as a single mental occurrence—the ISE 'I should grade those papers'—is really the occurrence of two quite distinct states, neither of which obviously occurs in a natural language, and either of which could (so far as we know) occur without the other. Their frequent pairing seems to require explanation.

5. Conclusion

The ubiquity of inner speech in the mental lives of most humans suggests that it plays some important cognitive role. Many have thought that the special role it plays, at least much of the time, is metacognitive in nature. However, this paper has argued that the pieces do not fit together when it comes to explaining why inner speech, as such, would be especially involved in metacognition. We face a kind of dilemma. On the one hand, if we think of ISEs as simply having auditory phonological contents, then reflection on such states will only ever be reflection upon representations of sounds. On the other hand, one might insist that ISEs have rich semantic (and propositional) contents in addition to their auditory-phonological contents. However, it seems that the only way of making good on that claim is to allow that ISEs are a kind of compound state, consisting of two distinct but contemporaneous states—one with a propositional/semantic content, and the other with a content relating to word sounds. For reasons discussed above, this conflicts with the reasons theorists have given for why inner speech would be especially useful to metacognition.

What role, then, does inner speech play in metacognition? Whatever answers we wish to develop, they should not *simply* appeal to its (putative) linguistic structure, its (putative) sensory character, or its (putative) tendency to occur consciously. Future proposals must cohere with one of the two broad possibilities outlined above for understanding the content and structure of inner speech episodes. Only then can we make real progress on understanding the cognitive—and possibly metacognitive—role of inner speech.

Department of Philosophy
University of Cincinnati

References

Baddeley, A. D. 1966. Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quaerterly Journal of Experimental Psychology, 18*, 362-365.

Baddeley, A. D. 2007. *Working memory, thought and action*. Oxford University Press.

Bermudez, J. L. 2003. *Thinking without Words*. Oxford: Oxford University Press.

Bermudez, J. L. 2005. *Philosophy of Psychology: a contemporary introduction*. New York: Routledge.

Carruthers, P. 2010. Introspection: Divided and Partly Eliminated. *Philosophy and Phenomenological Research, 80*(1), 76-111.

Carruthers, P. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Clark, A. (1998). Magic words: how language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and Thought: Interdisciplinary Themes* (pp. 162-183). Cambridge: Cambridge University Press.

Conrad, R., & Hull, A. J. 1964. Information, acoustic confusion and memory span. *British Journal of Psychology, 55*, 429-432.

Feinberg, T. E., Gonzalez Rothi, L. J., & Heliman, K. M. 1986. 'Inner Speech' in Conduction Aphasia. *Archives of Neurology, 43*, 591-593.

Frankish, K. 2004. *Mind and Supermind*. Cambridge: Cambridge University Press.

Gauker, C. 2011. *Words and Images: An Essay on the Origin of Ideas*. Oxford: Oxford University Press.

Geva, S., Bennett, S., Warburton, E. A., & Patterson, K. 2011. Discrepancy between inner and overt speech: Implications for post-stroke aphasia and normal language processing. *Aphasiology, 25*(3), 323-343. doi: 10.1080/02687038.2010.511236

Heavey, C. L., & Hurlburt, R. T. 2008. The phenomena of inner experience. *Consciousness and Cognition, 17*, 798-810.

Hurlburt, R. T. 1993. *Sampling Inner Experience in Disturbed Affect*. New York: Plenum Press.

Jackendoff, R. 1996. How language helps us think. *Pragmatics and Cognition, 4*(1), 1-34.

Klinger, E., & Cox, W. M. 1987-1988. Dimensions of thought flow in everyday life. *Imagination, Cognition and Personality, 7*, 105-128.

Kosslyn, S. 1994. *Image and Brain: The Resolution of the Imagery Debate*. Cambrdige, MA: MIT Press.

Levine, D. N., Calvano, R., & Popovics, A. 1982. Language in the absence of inner speech. *Neuropsychologia, 20*(4), 391-409.

Machery, E. 2005. You Don't Know How You Think: Introspection and Language of Thought. *British Journal for the Philosophy of Science, 56*, 469-485.

Martinez-Manrique, F., & Vicente, A. 2010. 'What the...!' The role of inner speech in conscious thought. *Journal of Consciousness Studies, 17*(9-10), 141-167.

Morin, A. 2005. Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms, and empirical evidence. *Journal of Consciousness Studies, 12*(4-5), 115-134.

Oppenheim, G. M., & Dell, G. S. 2008. Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition, 106*, 528-537.

Oppenheim, G. M., & Dell, G. S. 2010. Motor movement matters: the flexible abstractness of inner speech. *Mem Cognit, 38*(8), 1147-1160. doi: 10.3758/MC.38.8.1147

Ozdemir, R., Roelofs, A., & Levelt, W. J. 2007. Perceptual uniqueness point effects in monitoring internal speech. *Cognition, 105*(2), 457-465. doi: 10.1016/j.cognition.2006.10.006

Postma, A., & Noordanus, C. 1996. Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech, 39*, 375-392.