# The Nature of Unsymbolized Thinking

Agustín Vicente[a*] and Fernando Martínez-Manrique[b]

([a]Ikerbasque: Basque Foundation for Science / Linguistics and Basque Studies Department, University of the Basque Country, UPV/EHU, Spain; [b]Dpt. of Philosophy I, University of Granada, Spain)

[*]corresponding author: agustin.vicente@ehu.eus

## Abstract

Using the method of Descriptive Experience Sampling, some subjects report experiences of thinking that do not involve words or any other symbols (Hurlburt and Heavey 2006; Hurlburt and Akhter 2008). Even though the possibility of this unsymbolized thinking has consequences for the debate on the phenomenological status of cognitive states, the phenomenon is still insufficiently examined. This paper analyzes the main properties of unsymbolized thinking and advances an explanation of its origin. According to our analysis, unsymbolized thoughts appear as propositional states, that is, they are experienced as compositional conceptual phenomena, with semantic and syntactic features analogous to those of the contents of utterances. Based on this characterization we hypothesize that unsymbolized thinking is continuous with the activity of inner speech, in particular, it is a form of inner speech where the speech action is aborted even before the intention to talk is implemented by motor commands. We contend that this account provides the best explanation of the distinctive features of the phenomenon, and it helps to understand the sense of agency and ownership associated with it. Finally, we consider a possible objection arising from the experience of unworded inner speech, and we show how our account should inform the debate about cognitive phenomenology.

**Keywords:** cognitive phenomenology; Descriptive Experience Sampling; inner speech; semantic prediction; unsymbolized thinking

## 1. Introduction

One of the key issues in debates concerning the phenomenology of cognition is the possibility of experiencing cognitive states in a way that is not dependent on the experience of another non-cognitive state, e.g., a perceptual or sensory state. One problem of such debates is the shortage of evidence capable of sustaining one position or the other. The discussion is often backed by appealing to intuitions about what this or that scenario –e.g., hearing a sentence without comprehending it– "feels like". Yet phenomenological intuitions do not seem to agree too easily. In this paper we want to examine an experimental finding that seems to support the existence of a kind of cognitive state that can be experienced as such without the aid of any other state: the phenomenon of *unsymbolized thinking* (Hurlburt and Akhert, 2008; Hurlburt, Heavey, and Kelsey, 2013).

Russell Hurlburt and collaborators developed and tested a method to collect data about people's inner experiences at a moment as close as possible to the actual

occurrence of the experiences (Hurlburt and Heavey, 2006). The method is called Descriptive Experience Sampling (DES) and it has two parts. In the first part, subjects carry a beeper during waking hours as they go through their everyday routines. The device beeps randomly several times a day and the task of the subject is to record what was going on in the inner experience she was having at the moment of each beep. The second part is an interview with the researchers shortly after a sampling day. The purpose of this interview is to help the subject to describe in her own terms and in the most detailed way the nature of the reported experiences. The process of sampling and interview is repeated several times so that subjects can gain experience in observing and reporting their mental experiences. This way Hurlburt et al. expect to obtain more reliable and accurate reports.

Analysing their subjects' reports Hurlburt et al. proposed five distinct categories of inner experiences: unsymbolized thinking, inner speech, inner seeing, feelings and sensory awareness. Unsymbolized thinking refers to "thinking a particular, definite thought without the awareness of that thought's being conveyed in words, images, or any other symbols" (Heavey and Hurlburt, 2008: 802). According to collected data, unsymbolized thinking (UT) was reported in 22% of all sampled experiences, and there was considerable interindividual variability, given that the frequency of UT ranged from 0% to 80% –actually, a quarter of the subjects reported no UT at all. One interesting thing about the findings was that the existence of the phenomenon was relatively unexpected for subjects who experienced it. When they reported about states like those they first struggled to try to characterize them as states of a different kind –namely, as instances of inner speech– only to find that their experiences lacked the distinctive symbolic features of inner speech (i.e., words) or any other symbols.

We acknowledge that one may raise doubts on the robustness of those findings. On the one hand, the DES can be criticized on methodological grounds (Scollon, Kim-Prieto, and Diener 2003; Hurlburt and Schwitzgebel, 2007). Yet it is a method that is receiving increasing attention and that can be possibly combined with other methods (Kühn et al., 2014, Hurlburt et al. 2016[1]). On the other hand, some authors reject the existence of UT. Carruthers (2009) suggests that UT may be a result of confabulation,

[11] Based on their DES method, Hurlburt et al. (2013) have distinguished two different experiential phenomena: inner speaking and inner hearing. Kühn et al. (2014) found different neurological bases for each of these phenomena. Hurlburt et al. (2016) also report different neurological correlates for spontaneous and elicited inner speech. That is, fMRI studies seem to provide evidence in favor of Hurlburt et al.'s distinctions drawn on the basis of the DES method, and so in favor of the DES method.

e.g., people report thinking without words or images, but they may be actually using words and/or images. As we will argue, unsymbolized thoughts appear as propositional and effable, and Carruthers's suggestion could help to explain those properties: they arise from the fact that the subjects are actually using words, even though they do not realize they are using them. Hurlburt, Heavey, and Kelsey (2013), in contrast, suggest that confabulation probably goes the other way around: we engage in UT more frequently but, as we tend to identify thinking with innerly speaking, we tend to report using words when in fact we are not using them. We think that this line of response can be expanded once one takes into account that people are able to discriminate between episodes of inner speech and episodes of UT. Given that in the former episodes it is assumed that subjects are not confabulating about the kind of experience they report, it is unclear why they should be confabulating in the latter episodes. Why is it that sometimes people fall under the illusion of not using words, and others they do not fall? Finally, it also has to be noted that subjects are typically surprised to experience episodes of UT (see below), which would be strange according to the confabulation hypothesis.

Even though we prefer to take the phenomenon of unsymbolized thinking at face value, the unconvinced reader can understand our task as conditional: if there is a phenomenon like that, what explanation could it have? In other words, if there is such an experience as that described as UT, we want to ascertain what sort of mental underpinning is more likely to support it. Indeed, if one finds a coherent way to understand how UT can take place one may provide further reasons to support its plausibility. At any rate, characterizing the phenomenon is relevant for, inter alia, the issue of cognitive phenomenology, given that UT appears as a form of cognitive *experience* (see Jorba, 2015).[2] So if the notion is to be put to any use in the debate it is necessary to have analyzed what the phenomenon is. Our aim in this paper is, first, to offer a minimal characterization of the phenomenon. We will argue that the alleged features of UT locate it among compositional conceptual mental states. Second, we want to move beyond Hurlburt et al.'s descriptive program, i.e. to capture and describe pristine experience, and advance a possible explanation of the phenomenon assuming the minimal characterization that we previously offered. Hurlburt and his collaborators'

[2] Even if the existence of UT seems to show that there is indeed cognitive phenomenology, it is not clear whether it supports the stronger claims that are made about such a phenomenology –e.g., that each thought-content has its own what it is like, or that the content of a thought is given by its phenomenological properties. We will briefly revisit this issue in the last section.

research program attempts at characterizing different kinds of conscious states as they are experienced by subjects. That is, the program does not contemplate moving beyond the level, or the world, of experience. Our purpose is different: we want to offer a plausible explanation of one kind of experience, UT, and of the features that it has according to Hurlburt et al. In sum, we want to explore a way to explain the genealogy of UT, and thus, a way to explain what UT is. In particular, we will claim that the best explanation is possibly to regard it as continuous with inner speech. We will back this claim with two arguments: one is that unsymbolized thoughts look semantically and syntactically analogous to the contents of inner utterances; the other is that relating UT and inner speech helps to understand the elements of agency and ownership associated with both. Finally, we will defend our explanation from a possible objection arising from another of the findings reported by Hurlburt et al., namely, the case of unworded inner speech, and will close by considering implications that our account may have for the cognitive phenomenology debate.

## 2. A minimal characterization of UT

As we said, unsymbolized thinking is one of the five categories of the experiences allegedly found by means of the DES method. To be a distinct category means that it is not reducible to features of any of the other categories, even if it can share aspects with some of them. In this section we articulate and discuss the minimal positive characterization of the phenomenon that can be obtained from the reports provided by Hurlburt et al.

In a nutshell, UT can be defined as the feeling of "thinking a particular, definite thought without the awareness of that thought's being conveyed in words, images, or any other symbols" (Heavey and Hurlburt, 2008: 802). Let us summarize the claims by Hurlburt and Akhter (2008: 1366–67) regarding the nature of UT:

1) UT is its own distinct phenomenon.

2) It is a way of *experiencing*.

3) It is experienced to be a *thinking*, not a feeling, not an intention, not an intimation, not a kinesthetic event, not a bodily event.

4) Its content is *explicit*: the "about what" of the thought is plainly apprehended.

5) It is *differentiated*: the "what about it" is not general or vague. Taken together, the explicit and differentiated characteristics imply that the thought's sense is quite clearly articulated.

6) Its content is directly in experience.

7) An unsymbolized thought typically presents itself all at once; there is no rhythm or cadence; no unfolding or sequentiality.

8) It does not include the experience of words, images, or any other symbols.

If this list of properties captures adequately the phenomenon, we argue that they point toward two main features about the nature of an unsymbolized thought[3]. First, it appears as a propositional phenomenon: from 3, 4 and 5 we will argue that unsymbolized thinking is propositional, from 6 and 7 that it is a proposition that is experienced "at one go", and from 8 that is not actually vehicled by words. Second, and a bit more controversially, we will argue that unsymbolized thoughts are effable, that is, even though they do not appear in words they can be given a pretty straightforward and quite definite linguistic rendition. If these two features are on the right track, there are two conclusions that we can draw: first, if unsymbolized thinking is propositional, we contend that it is a conceptual and syntactic phenomenon; second, if it is effable, we contend as well that it is a linguistic phenomenon. Let us devote the next two subsections to developing these conclusions.

### *2.1. The propositional nature of unsymbolized thinking*

Unsymbolized thoughts appear as explicit and differentiated, with a clearly articulated sense. These are the marks, we contend, of propositional thinking, where propositions are typically understood as the semantic contents of sentences. This is suggested by Hurlburt and Akhter when they write: "Just as a complete sentence contains a subject (the about what) and a predicate (the what about it), the typical unsymbolized thought can be said to have those characteristics" (2008: 1367). Now, notice that one of the

[3] This investigation about the nature of UT can be seen as supplementary to Hurlburt's overall project, which is to investigate pristine experience, i.e., how inner life is presented to experiencers. We are not claiming that it is part of the content of subjects' UT-experiences that such experiences are propositional, or syntactically and semantically structured.

properties of UT is that its content is directly in experience. So, inasmuch as UT is propositional, what subjects are experiencing directly is some propositional semantic content. We submit that the best way to understand this is to say that subjects experience the tokening of some mental objects[4] –some mental representations. Given its propositional structure, these representations are arranged in compositional relations mediated by syntactic links of one sort or another. So inasmuch as unsymbolized thoughts are propositional, they have syntactic properties too.

At this point one may object that, according to Hurlburt et al's characterization, "an unsymbolized thought typically presents itself all at once; there is … no unfolding or sequentiality". These features may appear to tell against the presence of syntax, for syntax seems to require sequential structure. As a matter of fact, this is not obviously true, although we think the point is not important. It is not obviously true because it is still not clear whether the only proper syntactic structure is the hierarchical structure of constituents, i.e. phrases in phrase structure grammars. According to the minimalist program, for instance, although the expression of a hierarchical phrase structure has sequential properties and linear ordering, such a linear ordering, though syntactically determined it is not a syntactic property of sentences; rather it belongs to the phonological form (Kayne, 1994). So, on this influential account, syntax per se does not necessarily imply sequentiality.

However, we believe that UTs do have sequential structure, i.e., that, apart from an abstract phrase structure, the component concepts of UTs are arranged in a certain order, such that, for instance, complements may precede heads rather than vice versa (see below). The reason why it is not important if UTs have sequential structure is that such a sequential structure can also be experienced "all at once". That is, the absence of sequentiality in the experience of UT does not require that what is experienced does not have sequential structure or linear ordering. We may grant that the thoughts we experience have minimal sequential properties, i.e. linear ordering (e.g., that they are head-initial or head-final), for linear ordering would give raise to the *experience of sequentiality* (i.e., that parts, or pieces, come one after the other) if, and only if, we try

[4] We are endorsing here a version of Margolis and Laurence's approach to concepts as mental objects, as opposed to abstract objects (Margolis and Laurence 2007). It is beyond this paper to argue for this option, yet we want to observe that endorsing the abstract objects view would give rise to a different set of problems. The 'what aboutness' of unsymbolized thoughts –as well as of any other kinds of thoughts– would have to be accounted for in terms of the relation of an experience with an abstract entity that constitutes its content. The idea that the content is "directly in experience", as claim 6 above states, would be jeopardized, unless a different account of the directness of experience were provided.

to *express* such linear ordering, which would unfold in time. However, if we do not even try to pronounce a certain sentence, we will not experience sequentiality or unfolding.

By the same token, unsymbolized thinking can have sequential structure without displaying ”unfolding”. In experiencing an unsymbolized thought, subjects can be said to experience its structure “all at once”. This is a temporal notion that refers to the idea that there is no temporal deployment of the elements, even if these are syntactically articulated. As the issue of whether linear ordering does or does not pertain to syntax proper is both controversial and not relevant for this paper, in what follows we will use ‘syntax’ liberally to also include certain formal properties of a language such as whether heads precede complements, whether verbs and subjects agree in person, etc.

Stating that UTs have syntax liberally construed does not amount to claiming that they are linguistic in nature. One must allow for the possibility that UTs are experiences of “Mentalese”. In other words, when one wonders about the source of the combinatorial structure exhibited by UT there are two likely answers that come to mind. One is that the structure comes from linguistic production mechanisms, so that the syntax of UT is in fact the syntactic structure of natural language. The other possible answer is that the structure of UTs comes from *thought production* mechanisms, such that thoughts themselves consist of representations linked by a syntax. Of course, the latter idea amounts to the hypothesis of the Language of Thought (Fodor, 1975) [5].

We think that the fact that UT episodes are effable, and that it is relatively easy for subjects to express them, supports the linguistic origin option and disfavours the LOT alternative. Effability is not one of the properties of UTs that Hurlburt et al. highlight most (vs. e.g., explicitness), but their reports make clear that UTs are easy to identify –indeed, they are characterized as having “a univocal meaning”– and verbalize. Verbalization may not be straightforward, as in the case of inner speech proper, and

[5] According to the LOT hypothesis, the vehicle of our thinking is language-like in that it has a set of primitive representations (concepts) and a set of rules (syntax) that defines the set of well-formed compound representations. It was typically assumed that the notion of syntax in LOT was similar to syntax in natural language (NL). However, one may wonder in what respects are the syntax of the LOT and the syntax of a NL alike. According to what we have said above, it is possible to separate syntax proper from other structural properties of an NL such as linear ordering. Does the LOT have just hierarchical phrase structure, or does it have some linear ordering as well? If we go along with minimalists and locate linearization in PF, i.e., linked to speech production, then we could imagine that the LOT has no linear ordering (e.g., it is neither head-final nor head-initial). In such a case, our considerations below would have more force, because what we intend to argue at the end of the day is that UTs have syntactic properties liberally construed, such as linear ordering or agreement.

some subjects may express some uncertainty as to what "rendition" of the meaning is the most accurate (see also Jorba, 2015). However, even in these cases, subjects consider only a narrow set of possible renditions, all of which are very similar[6]. This suggests that the conceptual structure constitutive of the UT is linguistically prepackaged, and in particular, that it has been arranged in accordance with the rules of the subject's particular language. Let us quote Hurlburt and Akhter (2008) in length with respect to one of their subject's reports:

> "Dorothy is tiredly walking down the hall dragging her feet noisily on the carpet. She is thinking, if put into words, something quite like, ''Pick up your feet—it sounds like an old lady". However, there are no words, images, or other symbols experienced in that thinking. Despite the lack of words, the sense of the thought is very explicit: ''pick up your feet" is a more accurate rendition of the experienced thought than would be ''I should pick up my feet"; and ''it sounds like an old lady" is more accurate than ''I sound like an old lady"" (2008: 1364)

If Dorothy were having this thought in Mentalese, it has to be explained how she could recognize it as the thought it is. The thoughts "I sound like an old lady" and "it sounds like an old lady", cannot be distinguished unless either grammatical subjects are made explicit or there is agreement between the subject and the verb (in this case, it is both). Otherwise the thought would be something like "sound like an old lady", which is ambiguous between the two thoughts Hurlburt and Akhter contrast. A speaker of a subject pro-drop language (P-D) with agreement, like Spanish, will differentiate between both thoughts on the basis of subject-verb agreement, while a speaker of a language without agreement (W-A) and no pro-drop will use the overt subject to tell one thought from the other. In principle, it could be that both subjects, (P-D) and (W-A), distinguish the two thoughts using the same strategy, i.e., whatever strategy the LOT might use to express the difference between one thought and the other. But it at least sounds strange that subjects can recognize the way thoughts are made explicit if that way is alien to the way *they* make thoughts explicit. On the other hand, it seems to

[6] Jorba (2015) emphasizes the fact that subjects are less confident about the exact wording of the unsymbolized thought than in cases of inner speech. Still, as she points out, Hurlburt et al. state that the meaning is the same in the different renditions of the UT of their subjects. Indeed, they remark that, unlike experiences of words or images, "unsymbolized thinking has one and only one meaning" (Hurlburt and Heavey, 2006, 241). We take it that the fact that subjects choose among a rather limited range of closely-related wordings suggests that the process of putting UT into words is not a laborious or interpretive one.

be the case that the structure of UTs feels familiar enough so as to believe that their syntax is the syntax of the subject's particular language.

The (liberally conceived) syntax of particular languages differs greatly and in many respects. As we have seen, languages differ in their agreement patterns, and in whether they are or not pro-drop. Word order is not fixed either: some languages, like English, have a standard SVO (Subject-verb-object) order, while others, like Basque, have a SOV order; some construe their adpositional phrases using prepositions, like English again ('from Tokyo') while others, like Japanese, use postpositions ('Tokyo kara': from Tokyo), etc. Moreover, differences can be so extreme that many authors doubt that there can be any kind of Universal Grammar or set of abstract principles that apply to all languages. Evans and Levinson (2009) go as far as to deny that all languages "play" with the same tools, namely, a fixed set of class of words such as names, verbs, adjectives and prepositions, and a set of major phrasal categories such as noun phrase, verb phrase, etc. So the fact that UTs appear as syntactically familiar suggests that they draw on the syntax of one's own natural language.

Now, one could still insist that this is not enough to disregard the possibility that an unsymbolized thought is an instance of LOT. It could simply be that the syntax of LOT is just the syntax of natural language. In other words, an unsymbolized thought would be a thought assembled with the syntax of natural language but not by the language production mechanism. This would prevent UT from being continuous with inner speech, but there is a price to pay for this view. First, this would jeopardize the claims of universality typically associated with the LOT hypothesis, namely, that people share some basic representational repertoire as well as the same compositional processes operating on those representations. Second, even if one takes the universality claim as optional, it is difficult to see what explanatory gain one would obtain from regarding UT as natural-language-like but not language-produced. In our view, the interesting versions of the LOT hypothesis are those that assume that the syntax of the LOT is *independent* of the syntax of language, so the interesting versions of UT-as-LOT should show that the syntax of unsymbolized thoughts is independent from the syntax of language as well. To put it another way, unless one had good independent reasons to hold that thoughts are regimented by natural language categories, to claim that unsymbolized thoughts are "natural-language-like" without being "natural-language-produced" looks like an ad hoc manoeuvre.

### *2.2. The semantics of unsymbolized thinking*

There might be a different version for the natural-language-likeness of thought that relied not on syntax but on semantics. It is controversial to what extent our conceptual structure reflects our linguistic categories. Many authors distinguish between properly conceptual and semantic-conceptual representations (e.g., Slobin, 1996, Levinson, 2003, Malt et al. 2003). It may be that the language we speak has such a pervasive influence in cognition that our conceptual categories are ultimately aligned with the categories provided by our language. If this were so, one could hold that the unsymbolized thought appears as language-like not because it is language-produced but because conceptual boundaries, in general, reflect linguistic boundaries. However, the strong linguistic relativity thesis does not seem to be supported by the current research (Vicente and Martínez-Manrique, 2013). In principle, this means that when thinking we mostly use our conceptual –non-linguistic repertoire. For instance, it seems that results suggest that there are no differences in conceptualization between speakers of Manner-languages like English (i.e. languages that lexicalize the manner of the motion) and of Path-languages like Spanish (i.e. languages that lexicalize the path of the motion: see Malt, Sloman, and Gennari 2003, Papafragou, Hulbert, and Trueswell 2008). Speakers of Manner-languages interpret, memorize and retrieve motion events in basically the same way as speakers of Path-languages do, unless they are told to verbally describe the events after watching them.

Yet, it may be that the UTs of an English speaker differ from the UTs of a Spanish speaker with respect to how they describe a certain motion event. The English speaker may "see" that her UT describes, or is about, how the agent drove back, while the Spanish speaker may "see" that her UT describes, or is about, how the agent *volvió conduciendo* (lit. came back driving). These eventual differences are not easy to spot, and, for all we know, there is as yet no evidence that could support the hypothesis that UTs present different contents to speakers of relevantly different languages. However, a way to understand that UTs have explicit contents is that these contents are explicit with respect to our acquired linguistic categories, i.e. with respect to how we would describe or verbalize them. The idea that UT can be explained as some form of inner speech, and in particular, that it involves the recruitment of semantic representations (see below), could account for this kind of explicitness, as well as for the relative easiness in which

subjects can report the content of their UT. If conscious thoughts used a representational system that cross-categorizes with respect to the semantic system, we should expect more hesitations in the moment of reporting UTs (especially if linguistic categories differ greatly from conceptual categories: Malt et al., 2011). Subjects accept different ways of putting their thoughts in words, ways that we can think they regard as semantically equivalent, but they do not seem to report not being able to completely capture the content of the thought by means of their speech.

In sum, whereas it cannot be discarded that UTs are tokens of LOT sentences made conscious, their effability and familiarity suggests that they are propositional meanings already prepared for being uttered, i.e. semantic representations structured in accordance to the syntax of the experiencer's own language.

## 3. Unsymbolized thinking as continuous with inner speech

We claim that the best explanation of the syntactic and semantic properties of an unsymbolized thought is that it is a linguistic phenomenon, namely, the semantic content of an interrupted inner speech act. This explanation has two further advantages: one is that it provides an account of how the experience of UT comes to life that is based on the mechanisms involved in the experience of inner speech; the other is that it can explain the properties of agency and ownership typically associated with UT by looking at how those properties apply to inner speech. Let us elaborate a little.

### *3.1. Unsymbolized thinking as aborted inner speech*

Even though there is no prevalent model of what inner speech is, there is an influential account (see, e.g., Carruthers, 2011; Swiney and Sousa, 2014). According to this account, which draws on research on motor imagery, an inner speech episode is a prediction issued by the forward models, made on the basis of an efference copy of a motor command for speech production, which is at some point aborted. Inner speech, thus, consists of strings of phonological acoustic representations (this is what the prediction is about) that are broadcast and thus made conscious. The account makes use of an influential monitoring model that traces back to the corollary discharge model of perception proposed by Helmholtz (1860), and extended by von Holst & Mittelstaedt (1950) and Sperry (1950) to deal with motor acts. The model works in the following

way: whenever a motor command is issued, the brain predicts, based on an efference copy (or corollary discharge) and the work of some forward models, what proprioceptive and sensory feedback will ensue. By comparing it to incoming signals, this prediction is used both to correct errors in execution and to identify a change in the world and in the body as self-initiated.

Research on motor imagery (Jeannerod, 2006, Guillot et al., 2012) has it that an episode of motor imagery could be a prediction based on a motor command that is not executed, but inhibited (see below). An episode of inner speech, thus, could be a prediction issued on the basis of an aborted motor command for speech production, a prediction about the incoming sensory signal that the subject would experience if he/she had executed the motor command. In Martínez-Manrique and Vicente (2015) we argued against a narrow view that characterizes inner speech in terms of a specific representational format or product –e.g., as phonological representations– and for the view that regards it as an activity that mobilizes different layers of representations, i.e., semantic, syntactic, articulatory, etc. However, while we reject the identification of inner speech episodes with predictions about acoustic properties, we endorse the idea that inner speech involves generating such predictions (along with higher-level predictions: see below) by the monitoring system just described. We think that this idea can be extended to explain the origin of UT. Let us explain.

The act of innerly speaking begins with a (non-conscious) prior intention to express a certain thought; this prior intention gets more and more specific, until it reaches the level of motor commands. The act, thus, engages the motor system and phonological representations in general (phonemic and articulatory as well as acoustic – which allegedly constitute the prediction made conscious), but it also involves conceptual/semantic representations, which are recruited in the first steps of the generation of the action. In sum, an inner speech act consists in putting a thought in words, which in turn implies making use of conceptual to phonological representations.

The purpose of the efference copies and predictions issued in motor acts is to monitor actions on-line, as well as to confirm authorship. However, a plausible hypothesis is that the monitoring system does not only receive efference copies from motor commands and issue predictions about the incoming sensory signal; it also receives efference copies from higher order intentions and makes predictions on that basis (see Pacherie, 2008, Pickering and Garrod, 2013, Martínez-Manrique and Vicente,

2015). These predictions are needed to monitor how the higher level intentions are realized. That is, when we engage in inner speech we issue predictions at different levels, corresponding to the different levels in the hierarchy of intentions and commands.

Now, the predictions linked to prior intentions can be made conscious in the same way that we can presumably make conscious the predictions linked to motor commands. One possibility suggested by Jeannerod (1995) is that predictions are made conscious just by being predictions of aborted actions, i.e., if an action is aborted after the prediction is issued, the prediction will make it into consciousness. If this were true, then we can say that what is made conscious in inner speech is not just phonological representations, but also their meaning. Since part of what is intended in an act of speaking is to express a certain thought-content (in one's own language), the prediction corresponding to this kind of intention will be the semantic content of the utterance. That is, what we predict is that a certain thought-content, which uses the semantics of our language, is expressed. The content of the prediction is precisely the thought content.

The explanation about the phenomenon of UT that this view about inner speech suggests is the following. In inner speech we form the non-conscious intention to express a certain thought, recruit semantic, syntactic and phonological representations, and issue a motor command to produce overt speech that is subsequently aborted. The result of this is a prediction about the sound of the aborted utterance accompanied by a prediction about its meaning (Pickering and Garrod, 2013). Suppose, however, that we abort the process earlier, in particular, before the message is ready for emission. In that case, only one kind of prediction will be issued, namely, a prediction about the meaning of the message, which will be experienced as such a meaning –and only a meaning. This would be a UT, i.e., a non-symbolized, non-perceptually vehicled, propositional entity that we can easily recognize and verbalize.

It is not clear how many stages can be discerned in speech production (or even if they are really stages), but suppose that a speech act begins with the intention to express a certain thought, and involves, inter alia, mobilizing semantic representations (so that the thought is expressible), and recruiting word representations that are given a certain syntactic structure. Below that, we find the level of phonology, where it is specified how words are expressed and how sounds are articulated. Finally, when everything is

ready for pronunciation, the information is passed to the motor level. Our hypothesis is that we experience UTs when we abort a speech act before we reach the level of phonology. The output is a string of abstract word (or maybe just semantic/conceptual) representations that categorize the world according to the semantics of the language we speak, has the syntax of the language we speak, and conveys the thought we want to express. That is why we easily recognize what the thought is about and why it is also easy for us to verbally report its content. [7]

### *3.2. Agency in unsymbolized thinking*

We take it that this sort of explanation of UT has the further explanatory benefit that it can account for the sense of agency that accompanies most of our thinking by resorting to a promising way to explain the sense of agency in general. The account in question was first defended by Feinberg (1978) and elaborated by Frith (1992), which in turn draws on the corollary discharge model of perception mentioned above. According to this view, we feel as agents of our actions only when our predictions and the incoming signals match. This means that we have to be able to issue predictions, and that these predictions are accurate enough. That is, we require that the action control system works well.

One way in which this insight about the sense of agency in overt acting can be exported to the mental domain is by realizing that lower level predictions can be treated as incoming signals from the point of view of higher level predictions. So, the acoustics of inner speech is a prediction issued on the basis of a motor command, but it is also a to-be compared signal for a higher-level prediction. When we engage in inner speech, most of the times we feel *we* are speaking. A plausible explanation of this feeling is that we are also monitoring predictions and treating them in the way we treat incoming signals.

[7] We might nuance this conclusion. UT, we want to claim, is typically a linguistic phenomenon. However, Hurlburt and Heavey (2006, 239) mention casually that "for some subjects, unsymbolized thinking seems "almost" to be innerly spoken" while for others it "seems "almost" to be seen". It might be that UTs can also be the result of aborting the production of images. We take it that the existence of this kind of "pseudo-imagistic" UT counts in favor of our general proposal, i.e., that UT is not just pure thinking, but some other activity (typically, linguistic) that is interrupted in its earlier stages. Through this paper we have focused on "linguistic UT" because the phenomenon is better documented, and also to keep the discussion manageable. But it does no violence to our proposal to state it in more general terms such as: the experience of UT results from aborting at an earlier point a cascade of intentions that, if aborted later, could produce a perceptual-inner experience.

If we believe that this sort of explanation for the feeling of agency is, as many seem to think, the best one currently in the market (Frith, 2012), then we have to conclude, at least, that there can be no sense of agency without intending to do something. In that respect, the view that when we experience a UT we are simply tokening a thought in consciousness (however that is explained) looks problematic. In contrast, our proposal seems to be able to accommodate the intuition (the fact?) that conscious thinking is a mental action, i.e. something we, persons, do, and feel authorship about. The account of inner speech and UT that has been provided gives us the resources to explain UT as stemming from intentions (intentions to speak) and as involving the control system apparatus. Since we construe UT as a prediction formed by semantic representations structured according to the syntax of our language, UT can also be treated as an incoming signal by a higher layer in the cascade of intentions, for instance, by the prior intention to express a particular thought. Models of speech production, as well as models of action control, are not sufficiently developed at this stage. We can just express our confidence that they will have space for the kind of account we are proposing. At least, we take it that we are on the right track with respect to an explanation of agency in UT concerning two points: first, that our explanation includes intentions; and second, that the explanation also involves the control system of efferent copies, forward models and predictions[8].

## 4. Unsymbolized thinking and unworded speech

We want to close our brief analysis of unsymbolized thinking by examining a possible problem that arises from another type of experience reported by means of the DES

[8] Unbidden thoughts do seem to pose a problem to our account. However, it may also be that unbidden thoughts are a different kind of phenomenon. On the basis of their use of the DES, Hurlburt et al. (2013) have distinguished between inner speaking and inner hearing: sometimes we feel we are the ones speaking to ourselves, but some other times we feel we are just listening to our own inner voice. In an interesting study with fMRI, Hurlburt and colleagues (Kühn et al., 2014) have seen that the brain areas involved in inner speaking are actually different from the brain areas involved in inner hearing. Basically, inner speaking is associated with production areas while inner hearing is associated with comprehension areas. Thus, it may turn out that inner speaking and inner hearing are two different phenomena. This suggests that thinking and having unbidden, or passive, thoughts, may be two different phenomena as well. In such a case, our account of UT might provide an explanation for one kind of phenomenon, i.e. for active thinking. This would still be an advantage over an account that does not explain either of them. And we can hope that we could capitalize on an explanation that accounts for inner hearing (yet to come) in order to account for unbidden thoughts.

method, namely, the experience of *unworded speech*.[9] According to Hurlburt et al., the experience of unworded speech is different from the experience of unsymbolized thinking. Unworded speech refers to episodes in which the subject regards her inner experience as a variety of inner speech in which there are missing parts. This is better understood in the cases of partially unworded speech, such as "you experience yourself as speaking, "That is a very strong _____ – maybe it is a gas leak!" with a temporal space reserved for the word "odor" but the word "odor" itself is not actually in your experience" (Hurlburt and Heavey 2006, 211). The phenomenon is related to the fragmentary form that inner speech can often adopt (Martínez-Manrique and Vicente, 2010). There is, however, a more puzzling category distinguished by Hurlburt et al. in their analysis: the case of totally unworded speech –or unworded speech, simpliciter. This is "the experience of speaking in your own inner voice except that you have no experience of words at all" (2006, 211).

While Hurlburt et al. devote some attention to the contrast between the experience of inner speech and unsymbolized thinking, they say surprisingly little about the difference with unworded speech. They characterize experiences of unworded speech as instances in which "you have the sense of speaking, and are directly aware of the vocal characteristics of that speaking (rate, inflection, timbre, rhythm, etc.), and are directly aware of the meaning of what is being "said," even though no words are present" (2006, 211-12). In contrast, features that typically characterize the experience as one of speaking, such as "sense of control, linear sequence, rhythm, pace, etc" (2006, 219), are absent in the case of unsymbolized thinking. Those are temporal features, i.e., the qualities of a phenomenon that appears as developing in time. Instead, unsymbolized thinking appears somehow like an "instantaneous" thought, without recognizable temporal parts.

Even though we think that further work is needed to refine the characterization of unworded speech, let us grant the assumption that unsymbolized thinking is experienced as a phenomenon that lacks features that are present in unworded speech and concentrate on the issue whether this is enough to regard them as unrelated phenomena. Our claim is that this is not necessarily the case.

Notice, first, that partially and totally unworded speech are described as varieties of the same kind of experienced phenomenon –namely, inner speech–, even though

[9] We will use this a shorthand for 'unworded inner speech', just as Hurlburt and Heavey do.

there are important experiential differences between them –one preserves specific words, the other doesn't. Of course, one could regard this phenomenologically salient difference as a mark of two different kinds of experiences. Indeed, there seems to be a certain arbitrariness in classifying experiences as belonging or not to the same kind, given that kinds can be as fine-grained as one wishes. What matters to us, at least, is to search for kinds that are unified in theory-relevant ways, e.g, because the alleged instances of the kind have a common etiology and similar functions.

Apparently, partially and totally unworded speech are both experienced as instances of speech because they share a temporal profile analogous to the profile of the speech experience. In contrast, unsymbolized thinking does not show such a temporal profile. However, as we said above with respect to syntax, not all is temporal in speech. Temporality appears as a property of giving expression to some meaning, but not of a property of the meaning itself, even if this meaning is arranged in such a way that some concepts "precede" others (e.g., subjects precede objects, or, in general, heads precede their complements). In this respect, unworded speech and unsymbolized thinking can also be regarded as phenomena that preserve a significant component of inner speech, namely, the meaning (syntactically structured). So cases of unsymbolized thinking can be perfectly classified, in continuity with partially and totally unworded speech, as instances of inner speech in which the characteristic elements of *speech* (related to pronunciation), including its temporal features, are lacking, so that only meanings remain. Unworded speech, in contrast, would involve certain characteristics that pertain to emissions. One could perhaps hypothesize, as Alderson-Day and Fernyhough (2014) contend, that unworded speech would not be different from what Fernyhough (2004) calls 'condensed inner speech', which refers to the sort of phenomenon that Vygotskyans call 'thinking in pure meanings' –a phenomenon with a verbal origin.

## 5. Final remarks: UT and the cognitive phenomenology debate

In the previous sections we have characterized UT roughly as the experience of the conceptual-semantic content of an inner speech act that is issued by the linguistic production mechanism but aborted before it reaches the phonological level. We want to end this paper by considering what consequences this picture has, if correct, for the cognitive phenomenology debate.

First, UT shows that while *thinking* that p has a generic, proprietary phenomenology –as opposed, say, to *perceiving* that p or even to *saying* that p–, the phenomenon of UT does not show anywhere as neatly that thinking *that p* has a distinctive phenomenology (Pitt, 2004) which differentiates thinking *that p* from thinking *that q*. The phenomenological difference is reflected in people's first reactions towards UT. The typical reaction of a DES subject when reporting UT for the first time is something like "now I know what it is like to have an unsymbolized thought". For instance, 'Evelyn' is described as "Looking powerless: palms turning slightly up, eyebrows raised, voice uncertain" when she reports that she was "just *thinking* [if NetZero is really much cheaper than Cox Cable]" (Hurlburt and Akhter, 2008, 1365). However, she would not claim: "now I know what it is like to wonder how much cheaper NetZero than Cox Cable is". As they point out, she seems surprised about the fact that she has been thinking a thought without words or images, i.e. that she has had that kind of experience. At that point, it counts as a genuinely new experience for her. However, our bet is that the next time she experiences another UT with a different content, she will feel that she is having that experience again, not that she is having a new experience. That is, it is the kind UT, and not each particular UT, that seems to have a distinctive phenomenology.

Second, even if UT shows the possibility of the experience of thinking without the experience of phonological representations, it requires the initiation of conceptual-linguistic activity that eventually would result in such representations –that is, if the process were not aborted early. So its semantic-related phenomenological properties, such as its explicit and differentiated character, are dependent on the same mechanisms that give rise to the corresponding phenomenological properties of linguistic experience. To put it a different way, even if *thinking* that p –in the sense of having UT– is phenomenologically different from *saying* that p –the latter involves words, the former does not– the experience associated with *thinking* that p typically accompanies the experience of *saying* that p. So the experience of having the UT that *p* is the same as the experience of the semantic content expressed in saying that *p*, detached from the rest of phenomenological properties involved in saying. We submit that it is the fact that this content appears as stand-alone that marks it as a different kind of experience from the subject's point of view. Yet this is compatible with the notion that it is issued by the

same language production mechanisms as the contents of our regular (inner or outer) sayings.

Third, similar considerations apply to the feeling of agency. UT shows that even simple judgments are experienced as authored by the subject. How could unsymbolic conscious judgments be experienced this way, if, to begin with, they are apparently not intended? If UT is, in fact, unsymbolized inner speech, what we experience in UT is a prediction about the meaning of a certain linguistic expression that we never utter. We feel authorship because this prediction is triggered by an intention to speak which is immediately aborted.

Finally, we contend that UT does not provide a sample of "pure thought", if by this one means linguistically uncontaminated thought. Even if our proposal of seeing UT as continuous with speech producing mechanisms were wrong, we claim that our observations about its semantic and syntactic characterization still hold. Given that the structure of UT is pretty much like the structure of semantic contents, were it the case that UT provided instances of "pure thought" this would only mean that thinking is already contaminated by categories from natural language. There is currently no basis to sustain this Whorfian scenario, so claiming that UT is of a piece with the activity related to inner speech stands as the most plausible explanation of the phenomenon.

## Acknowledgments

This is a thoroughly collaborative paper; order of authorship is arbitrary. The authors wish to thank the participants at the II Workshop on the Naturalization of the Mind and Modality, Girona, the Conference *Phenomenology of Cognitive Experiences*, Dublin, and the workshop *Inner Speech: Theories and Models*, Granada, where previous versions of this work were presented. The paper also improved from the comments by two anonymous referees. We are very grateful to the editors of this Special Issue, Marta Jorba and Dermot Moran. Research for this work was funded by Projects IT769-13(Basque Government) and FFI2014-52196-P (Agustín Vicente), and FFI2015-65953-P (Fernando Martínez-Manrique), of the Spanish Ministry of Economy and Competitiveness (MINECO).

## References


Alderson-Day, B., and C. Fernyhough. 2014. “More than One Voice: Investigating the Phenomenological properties of Inner Speech Requires a Variety of Methods.” *Consciousness and Cognition* 24: 113–114.

Carruthers, P. 2009. “How We Know our Own Minds: The Relationship between Mindreading and Metacognition” *Behavioral and Brain Sciences* 32 (02): 121–138.

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

Evans, N., and S. C. Levinson. 2009. “The Myth of Language Universals. Language Diversity and its Importance for Cognitive Science.” *Behavioral and Brain Sciences* 32 (5): 429–492.

Feinberg, I. 1978. “Efference Copy and Corollary Discharge: Implications for Thinking and its Disorders.” *Schizophrenia Bulletin* 4: 636–640.

Fernyhough**,** C. 2004. “Alien Voices and Inner Dialogue: Towards a Developmental Account of Auditory Verbal Hallucinations.” *New Ideas in Psychology* 22: 49–68.

Fodor, J. 1975. *The Language of Thought*. New York: Crowell.

Frith, C. 1992. *The Cognitive Neuropsychology of Schizophrenia*. Hover: Lawrence Erlbaum Associates.

Frith, C. 2012. Explaining Delusions of Control: The Comparator Model 20 Years on. *Consciousness and Cognition* 21 (1): 52–54.

Guillot A., F. Di Rienzo, T. MacIntyre, A. Moran, C. Collet. 2012. “Imagining Is Not Doing but Involves Specific Motor Commands: A Review of Experimental Data Related to Motor Inhibition.” *Frontiers in Human Neuroscience*, 6:247

Heavey, C. and R. T. Hurlburt. 2008. “The Phenomena of Inner Experience.” *Consciousness and Cognition* 17 (3): 798–810.

Helmholtz, H. V. (1860). *Treatise on Physiological Optics*. New York: Dover.

Hurlburt, R. T. and S. A. Akhter. 2008. “Unsymbolized Thinking.” *Consciousness and Cognition* 17 (4): 1364–1374.

Hurlburt, R. T. and C. L. Heavey. 2006. *Exploring Inner Experience*. Amsterdam: John Benjamins.

Hurlburt R. T., C. L. Heavey., and J. M. Kelsey. 2013. “Toward a Phenomenology of Inner Speaking.” *Consciousness and Cognition* 22 (4): 1477–1494.

Hurlburt, R. and E. Schwitzgebel. 2007. *Describing Inner Experience? Proponent Meets Skeptic*. Cambridge, MA: MIT Press.

Hurlburt, R. T., B. Alderson-Day., S. Kühn., and C. Fernyhough. 2016. “Exploring the Ecological Validity of Thinking on Demand: Neural Correlates of Elicited vs. Spontaneously Occurring Inner Speech.” *PLoS ONE*, 11(2): e0147932.

Jorba, M. 2015. “Conscious Thought and the Limits of Restrictivism”. *Crítica. Revista Hispanoamericana de Filosofía* 47 (141): 3–32.

Jeannerod, M. 1995. “Mental Imagery in the Motor Context. *Neuropsychologia* 33: 1419–1432.

Jeannerod, M. 2006. *Motor Cognition: What Actions Tell the Self*. Oxford: Oxford University Press.

Kayne, Richard S. 1994. *The Antisymmetry of Syntax*. MIT Press.

Kühn, S., C. Fernyhough., B. Alderson-Day., and R. T. Hurlburt. 2014. “Inner Experience in the Scanner: Can High Fidelity Apprehensions of Inner Experience Be Integrated with fMRI?” *Frontiers in Psychology* 5: 1–8.

Levinson, S. C. 2003. “Language and Mind: Let's Get the Issues Straight!” In *Language in mind: Advances in the study of language and cognition*, edited by D. Gentner, and S. Goldin-Meadow. Cambridge, MA: MIT Press.

Malt, B., S. Sloman, and S. Gennari. 2003. “Speaking versus Thinking about Objects and Actions.” In *Language in Mind: Advances in the Study of Language and Thought*, edited by D. Gentner and S. Goldin-Meadow. Cambridge, MA: MIT Press.

Malt, B., E. Ameel, S. Gennari, M. Imai, N. Saji, and A. Majid. 2011. “Do Words Reveal Concepts?” In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, edited by L. Carlson, C. Hölscher, and T. Shipley, 519–24. Austin, TX: Cognitive Science Society.

Margolis, E., and S. Laurence. 2007. “The Ontology of Concepts: Abstract Objects or Mental Representations?” *Noûs* 41 (4): 561–593.

Martínez-Manrique, F., and A. Vicente, A. 2010. “What the…!’ The Role of Inner speech in Conscious Thought.” *Journal of Consciousness Studies*, 17 (9-10): 141–67.

Martínez-Manrique, F., and A. Vicente. 2015. The Activity View of Inner Speech. *Frontiers in Psychology*. 6: 232. doi: 10.3389/fpsyg.2015.00232

Pacherie, E. 2008. "The Phenomenology of Action: A Conceptual Framework." *Cognition* 107 (1): 179–217.

Papafragou, A., J. Hulbert, and J. Trueswell. 2008. "Does Language Guide Event Perception? Evidence from Eye Movements." *Cognition* 108: 155–84.

Pickering M., and S. Garrod. 2013. "An Integrated Theory of Language Production and Comprehension." *Behavioral and Brain Sciences* 36 (4): 329–347.

Pitt, D. 2004. "The Phenomenology of Cognition or *What Is It Like to Think that p*?" *Philosophy and Phenomenological Research* 69 (1): 1–36.

Scollon, C. N., C. Kim-Prieto., and E. Diener. 2003. "Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses." *Journal of Happiness Studies* 4 (1): 5–34.

Slobin, D. I. 1996. "From "Thought and Language" to "Thinking for Speaking."" In *Rethinking Linguistic Relativity*, edited by J. Gumperz and S. Levinson. Cambridge: Cambridge University Press.

Sperry, R. W. 1950. "Neural Basis of the Spontaneous Optokinetic Response Produced by Visual Inversion." *Journal of Comparative and Physiological Psychology* 43: 482–489.

Swiney, L., and P. Sousa. 2014. "A New Comparator Account of Auditory Verbal Hallucination: How Motor Prediction Can Plausibly Contribute to the Sense of Agency for Inner Speech." *Frontiers in Human Neuroscience* 8: 675.

Vicente, A., and F. Martínez Manrique. 2013. "The Influence of Language in Conceptualization: Three Views." *Protosociology* 20: 89–106.

von Holst, E. and H. Mittelstaedt. 1950. "Das Reafferenzprinzip." *Die Naturwissenschaften* 20: 464–476.