Subject obviation and self-locating knowledge*

Julie Goncharov

forthcoming in *Lingvisticae Investigationes*
(preprint)

Abstract

Subject obviation is a restriction on having coreferential subjects in sentences like *#Je veux que je parte* 'I want that I leave'. Since the early 1980s, linguists have tried to explain this restriction, attributing it to different domains of grammar (syntax, semantics, semantics-pragmatics interface). In this paper, I defend the view that subject obviation is due to a property of "general intelligence" rather than linguistic competence (Ruwet 1984). In my analysis, the role of "general intelligence" is played by the principle of non-triviality that disallows ascription of propositions whose content does not exclude any possibility from a belief state (Stalnaker 1978, 1988). Embedded propositions in obviative sentences, I suggest, involve self-locating information, i.e., information about who and where we are in the world (Perry 1977, Lewis 1979). Because these propositions involve self-locating information, their use in some attitude ascriptions violates non-triviality. That is to say, sentences with subject obviation are deviant for the same reason as *#I believe I am sane*, when the speaker has no doubts about her sanity. I formulate my analysis within Stalnaker's framework (Stalnaker 2008, 2014). I conclude by showing how my approach sheds new light on phenomena that challenge grammar-based accounts of subject obviation.

Keywords: subject obviation, subjunctive, reference, non-triviality, pragmatics

1 Introduction

Subject obviation is a restriction on having coreferential subjects in some attitude ascriptions with finite complement clauses.[1] For example, *vouloir* 'want' in French can be used with a subjunctive complement, but not when the subject of the main clause and the embedded subject refer to the same individual; compare (1)a with (1)b. This restriction is not absolute; coreferential constructions can be used when the interpretation of the subject in the main clause does not fully align (in the sense to be discussed) with the interpretation of the embedded subject; see (1)c.

*Acknowledgement to be added.

[1]The initial name for this restriction was 'subjunctive obviation'. 'Subject obviation' is a recent label that is used for a broader range of constructions (e.g., Kaufmann 2019; Stegovec 2019). Although this paper focuses on what was previously called 'subjunctive obviation', I will use the name 'subject obviation' for two reasons. First, I believe a unified account with other similar phenomena is justified. Second, as we will see, 'subjunctive obviation' is a misnomer because the restriction is not limited to subjunctives.

(1) a. #Je veux que je parte. (French)
'I want that I leave-SBJV'
b. Je veux que tu partes.
'I want that you leave-SBJV'
c. Je veux absolument que j'amuse ces enfants.
'I want absolutely that I amuse-SBJV these children' (Ruwet 1991, 2, 30)

Subject obviation is found in different languages and with a variety of attitude predicates.[2] The examples in (2) and (3) show subject obviation in Italian and Hungarian. The sentences in (2)a and (3)a have the same structure as the French example in (1)a, and both are judged unacceptable. I will call such sentences *obviative*. I will call sentences with non-coreferential subjects, like in (1)b and (2)b, *non-obviative*, and sentences that have coreferential subjects but are not judged unacceptable, like in (1)c, (2)c, and (3)b, *ameliorated*.

(2) a. #Penso che io parta domani. (Italian)
'I think that I leave-SBJV tomorrow'
b. Penso che parta domani.
'I think that he/she leave-SBJV tomorrow'
c. Penso che io abbia fatto molti errori.
'I think that I have-SBJV made many mistakes' (Costantini 2023, 29–30)

(3) a. #Remélem, hogy (nem) szédülök. (Hungarian)
'I hope that I (don't) have-IND vertigo'
b. Remélem, hogy nem untatlak.
'I hope that I not bore-IND you' (Szabolcsi 2021, 10)

Generative linguists have studied subject obviation since the early 1980s, and there are many things we now know about this restriction. We know the restriction is not due to a syntactic condition, because ameliorated sentences come with varying degrees of acceptability, whereas syntactic well-formedness judgements are assumed to be categorical (Ruwet 1984). Research also shows that subject obviation is not due to competition between infinitival and finite complements, as it occurs in languages lacking infinitival alternatives (Szabolcsi 2021; Costantini 2023). Nor can subject obviation be explained by competition between subjunctive and indicative moods, since subject obviation is attested with both (Szabolcsi 2021). Finally, subject obviation is not directly tied to decisive modality (desires, hopes, commands), since it also appears with epistemic attitudes (Costantini 2016, 2023).

Most analyses of subject obviation have attributed the restriction to a domain of grammar (syntax, semantics, semantics-pragmatics interface). In this paper, I defend the view that subject obviation is due to a property of "general intelligence" rather than linguistic competence, which was initially articulated in Ruwet (1984). In my proposal, the role of "general intelligence" will be played by the principle of non-triviality that disallows ascription of propositions whose content does not exclude any possibility from a belief state (Stalnaker 1978, 1988).

The structure of the paper is as follows: The first three sections build towards my explanation of subject obviation (sections 2-4). I begin by discussing the explanation given by Ruwet (1984) (section 2). This discussion is important because my own explanation of subject obviation shares

[2](E.g., Bouchard 1983; Picallo 1985; Ruwet 1984; Kempchinsky 1986, 2009; Farkas 1992; Schlenker 2005; Costantini 2006, 2016, 2023; Szabolcsi 2010, 2021; Stegovec 2019.)

two key features with Ruwet's explanation: First, like Ruwet, I attribute subject obviation to a property of "general intelligence" rather than linguistic competence. This feature is discussed in section 3. Second, also like Ruwet (and several scholars after him), I argue that obviative sentences are unacceptable because there is an illegitimately tight "self-to-self" relation between the two coreferential subjects. This second feature is discussed in section 4. In section 5, I show how the main facts about subject obviation are derived using the apparatus developed in sections 3 and 4. In the last two sections (sections 6 and 7), I compare my analysis with current semantic-pragmatic accounts of subject obviation and discuss its advantages. Section 8 concludes the paper.

2 Ruwet's explanation

I begin with the explanation of subject obviation in Ruwet (1984).[3] Although Ruwet's study is one of the first on the topic, and since then we have learned more about subject obviation and grammar in general, his work is a good starting point for two reasons: First, it allows us to see the main elements in the analysis of subject obviation. Second, Ruwet's explanation has an important feature that recent accounts lack but that is central for my analysis.

Ruwet's explanation can be divided into four points. With the first and the last points, I will agree without reservation. The second point was shown to be wrong by the studies that came after Ruwet's, so I will reject it. The third point is what distinguishes Ruwet's explanation from other accounts of subject obviation. It holds that subject obviation is due to a property of "general intelligence" rather than linguistic competence. I will agree with this claim broadly but disagree about the specific mechanism of "general intelligence" that is the source of subject obviation.

Ruwet's explanation of subject obviation goes as follows: First, he maintains that obviative sentences, as in (4)a repeated from above, are syntactically well-formed. The main empirical contribution of Ruwet's work is that he identifies a range of strategies that can improve acceptability of obviative sentences. These strategies include (among others) the use of passives or modals and changing the Aktionsart or the viewpoint aspect; see (4)b-e. If (4)a were syntactically ill-formed, Ruwet argues, it would have been difficult to explain the different degrees of acceptability of ameliorated sentences in (4)b-e because syntactic well-formedness is, presumably, a categorical judgement.

(4) a. #Je veux que je parte. (French)
'I want that I leave-SBJV'
b. ?Je veux que je sois enterré dans mon village natal.
'I want that I be-SBJV buried in the village of my birth'
c. ?Je veux que je puisse attaquer à l'aube.
'I want that I can-SBJV attack at dawn'
d. ?Je veux que je réussisse.
'I want that I succeed-SBJV'
e. Je veux (absolument) que je sois parti dans dix minutes.
'I want (absolutely) that I be-SBJV gone in ten minutes' (Ruwet 1991, 20, 21, 23, 26)

Second, Ruwet argues that the unnaturalness of (4)a comes from the fact that it competes with

[3]The paper was first published in French as Ruwet (1984) and later translated into English and published as Ruwet (1991). I refer to this work by its first date of publication but use the English version for citations.

the infinitival construction *Je veux partir* 'I want to leave' that expresses the same thought. As mentioned in the introduction, the competition conjecture cannot be maintained. Since Ruwet's work, it has been shown that, in some languages, subject obviation obtains with attitude reports that do not have a suitable infinitival alternative for an obviative sentence. We will see evidence against the competition conjecture at the end of the section.

Third, the competition between a subjunctive and an infinitive construction, Ruwet continues, is due to a property of "general intelligence or central cognitive processes".[4] The property that he has in mind is "an **iconic** link between the (superficial) form of the sentence – simple or complex [...] – and the content, experienced as relatively simple or relatively complex".[5] In other words, Ruwet hypothesises that formally simple infinitival constructions like *Je veux partir* tend to be iconically associated with relatively simple construals, whereas more complex bi-clausal constructions like *Je veux que je parte* tend to be used to convey relatively complex relations. Thus, if we want to express a relatively simple thought, we tend to choose a simple infinitival construction, which makes sentences with subjunctive clauses less natural for expressing that same thought.

Fourth, Ruwet's classification of a construal as "relatively simple" or "relatively complex" has to do with the interpretation of a coreferential relation between the subject of the main clause and the embedded subject. A construal is "relatively simple" when the two subjects "are viewed from fundamentally the same point of view" or "the internal distance between the two instances of the self [brought in by the two subjects – Author] tends to vanish".[6] A construal is "relatively complex" when "the relation of self-to-self [...] involves an internal differentiation and highlights two distinct facets of the self, and/or introduces a certain distance between self and self".[7] This intuition that a "self-to-self" relation can be either "viewed from fundamentally the same perspective" or involve a "shift in perspective"[8] and "distancing between self and self"[9] is also central to most recent accounts of subject obviation. It comes under different guises: as a distinction between a (simple) *de se* and an event *de se* interpretation (Schlenker 2005), as the difference between objective information and "direct experience" (Szabolcsi 2021), or the difference in how knowledge is obtained – through reasoning or introspection (Costantini 2016, 2023). I think the intuition is correct. In this paper, I will formalise it using self-locating knowledge.

The four points above capture the main elements of the analysis of subject obviation that are present in many accounts of this restriction. They include: (i) the explanation for ameliorated sentences, (ii) the competition conjecture, (iii) the source of subject obviation (iconicity, for Ruwet), and (iv) the reason for subject obviation (the "self-to-self" relation, for Ruwet).

As I said earlier, my explanation of subject obviation will share key features with Ruwet's explanation. I agree with Ruwet's reasoning about ameliorated sentences: their varying degrees of acceptability show that the source of subject obviation is not a syntactic restriction. Nor is it, I believe, a semantic or a semantic-pragmatic restriction (for the reasons that will be discussed in sections 6 and 7). The source of subject obviation (I partially agree with Ruwet) is a property of "general intelligence", which is shared between linguistic and non-linguistic representations. In my explanation, the role of "general intelligence" will be played by the principle of non-triviality. When

[4](Ruwet 1991, 19)
[5](ibid., 8)
[6](ibid., 16)
[7](ibid.)
[8](ibid., 13)
[9](ibid., 15)

applied to linguistic representations, this principle disallows ascription of propositions whose content does not exclude any possibility from a belief state. This principle will be formulated and discussed in the next section (section 3). I will propose that embedded propositions in obviative sentences are unacceptable because they violate the principle of non-triviality. What causes embedded propositions in obviative sentences to violate the principle of non-triviality (and here again, I agree with Ruwet's initial insight and similar ideas in recent accounts) is the presence of a tight "self-to-self" relation between the interpretation of the subject of the main clause and that of the embedded clause. This relation, which I will formulate in terms of self-locating knowledge, will be discussed in section 4.

However, in my explanation of subject obviation, there will be no room for iconicity or "an iconic link", which is central for Ruwet's explanation. I set aside the objection that the concept of iconicity is hard to articulate; rather, I dismiss the usefulness of iconicity on the ground that it requires the competition conjecture, which has been shown to be wrong. As mentioned above, the competition conjecture says that attitude reports with subjunctive complements and coreferential subjects are unnatural because they compete with similar and simpler infinitival constructions. But, as Szabolcsi (2021) demonstrates, subject obviation obtains with attitude predicates that do not have infinitival alternatives. In Hungarian, subject obviation is found with attitude predicates that take either a subjunctive complement, such as *akarni* 'want' in (5)a, or an indicative complement, such as *ramélni* 'hope' and *sajnálni* 'regret/be sorry' in (5)b,c.

(5) a. #Azt akarom, hogy meg-látogassam Marit. (Hungarian)
'I want that I visit-SBJV Mary'
b. #Remélem, hogy fél lábon állok.
'I hope that I stand-IND on one leg'
c. #Sajnálom, hogy ugrándozok.
'I regret that I jump-IND around' (Szabolcsi 2021, 17)

Szabolcsi (2021) argues that since only *akarni* 'want' has a corresponding infinitival alternative, the competition conjecture cannot be right. I would add a qualification to this claim: the competition conjecture cannot be the main element in the analysis of subject obviation if we want to develop a unified explanation for subject obviation. But it remains to be seen whether the availability of an infinitival or other alternative construction in a language may contribute to the unacceptability of obviative sentences. The data from Hungarian show that the competition conjecture cannot be maintained, and the competition-based analyses, including Ruwet's analysis in terms of iconicity, cannot explain all cases of subject obviation.

In the next two sections, I continue building towards my analysis of subject obviation, elaborating on the source of subject obviation, which I formalise as the principle of non-triviality, and the reason for subject obviation, which I propose to capture as self-locating knowledge. After discussing the principle of non-triviality and self-locating knowledge, I will show how this apparatus can help us account for obviative and ameliorated sentences.

3 Non-triviality, belief ascription, and autonomy of pragmatics

This section and the next one are based on the ideas developed by Stalnaker, the way I understand them. I will use more formalism than Stalnaker usually does; this is to make his ideas easier

to use for a formal linguist. In some places, I will have to simplify the picture for reasons of space, but I hope to stay faithful to the spirit of the ideas. Because Stalnaker's ideas are rather on the philosophical side of the linguistics-philosophy divide and may not be all that familiar to the linguistic audience, I will structure my presentation in this section as follows: I will start with the idea that is most familiar to linguists, namely, that the role of an assertion is to exclude from the context in which the assertion is made those possibilities that are incompatible with the asserted proposition. I will call this *a principle of non-triviality*. After that, I will describe Stalnaker's view on belief ascription and how the principle of non-triviality is extended to belief. At this point, I will introduce a (partial) knowledge model that we will use later on to talk about subject obviation and I will define the principle of non-triviality in that model. Finally, I will discuss in what sense we can view the principle of non-triviality as a property of "general intelligence". As we will see, it has to do with the thesis of *an autonomy of pragmatics* that Stalnaker defends in his book *Context* (Stalnaker 2014).

In formal semantics, it is standard to assume that propositions are functions from possible worlds to truth values. Stalnaker also makes this assumption. But his conception of a possible world is different from the conception usually assumed in formal semantics, which comes from Lewis's work. As we will see, this difference affects a number of theoretical choices, so it is worthwhile to keep it in mind throughout the paper. Lewis's view on possible worlds is often called "modal realism", while Stalnaker's view is often referred to as "actualism". In Lewis's view, possible worlds are things of the same sort as our actual world; they differ from the actual world "not in kind but only in what goes on at them".[10] For Lewis, to believe in possible worlds is to believe in "the existence of entities that might be called 'ways things could have been' ".[11] In Stalnaker's view, possible worlds are (mere) possibilities, properties of the universe, relative to which truth is determined. For Stalnaker, to believe in possible worlds is to believe simply that they have a certain structure that "people distinguish [. . .] in their rational activities".[12]

Stalnaker's conception of possible worlds as possibilities distinguished by a rational agent (rather than real entities just like the actual world but not actualised) is fundamental for his contextualism. It allows him to assimilate the context in which an utterance takes place to the domain of possible worlds, which determines the meaning of the uttered sentence. Here, propositions are not functions from the domain of *all* possibilities but only from the domain of those possibilities that are distinguished in the context of utterance.[13] This provides the basis for the position that Stalnaker is most known for in linguistics, namely, that the role of an assertion is to exclude from the context in which the assertion is made those possibilities that are incompatible with the proposition expressed by the asserted sentence (Stalnaker 1978). The corollary of this statement is that a sentence cannot be asserted in a context that entails the proposition expressed by that sentence. I will call this corollary *the principle of non-triviality*.

[10](Lewis 1973, 85)

[11](ibid., 84)

[12](Stalnaker 1984, 57)

[13]In contextualism, the relation between semantics and pragmatics is more complex than in the relativist view, usually assumed in formal semantics where a sentence is evaluated with respect to a model and a context of utterance. This is a complicated issue that I will not take up in this paper. What is important to keep in mind for the purpose of this paper is that for Stalnaker, "context is not just information that mediates between utterance and proposition" either pre- or post-semantically when the meaning of indexicals is determined or implicatures are computed, rather "it is the material out of which propositions are constructed" (Stalnaker 1999, 156).

In Stalnaker's framework, the principle of non-triviality can be also formulated at the level of belief ascription (Stalnaker 1988, 2014). This is the level that I will have in mind when talking about the principle of non-triviality in this paper. For Stalnaker, the sentence '*Alice believes that* p' expresses a proposition that is a function of the proposition expressed by the embedded sentence '*that* p'. That is to say, the whole proposition expressed by '*Alice believes that* p' fulfils its assertoric function (i.e., excludes non-compatible possibilities from the context of utterance) in virtue of the embedded proposition '*that* p' excluding non-compatible possibilities from Alice's belief state. Following Stalnaker, I will refer to the context relative to which the whole proposition is evaluated as *a basic context* and to the context for the embedded proposition as *a derived context*. A basic context is a set of possible worlds compatible with what is known or assumed by conversational participants. A derived context for '*Alice believes that* p' is determined as follows: For each possible world in the basic context, Alice is in a particular belief state (which itself is a set of possible worlds compatible with Alice's beliefs in that world). The union of these belief states is a set of possible worlds that the speaker takes to be compatible with Alice's beliefs in the basic context. This union is the derived context in which the ascribed proposition expressed by '*that* p' is interpreted.[14] Having in mind the principle of non-triviality, I will say that if the derived context entails the proposition expressed by '*that* p', '*that* p' is not ascribable in that context. As I will argue in the next section, subject obviation obtains when there is an attempt to create an attitude ascription with a non-ascribable proposition. Therefore, I propose that the source of subject obviation is a violation of the principle of non-triviality.

To make the principle of non-triviality more concrete, let me introduce a (partial) knowledge model. Our (partial) model here is a tuple $\langle W, S, R\rangle$, where W is a (non-empty) set of possible worlds under Stalnaker's conception of possible worlds. As a reminder of this conceptual change, I will use $x, y, z, ...$ as variables for possible worlds (instead of customary $w, w', w'', ...$). I will use Greek letters $\alpha, \beta, \gamma, ...$ to name specific worlds when I need an illustration (α will always represent the actual world). S is a (non-empty) set of individuals or subjects. I will use capital letters $A, B, C, ...$ for members of this set. R is a regular epistemic relation on W (transitive, Euclidean, serial). In this model, let us say that R_A is a Hintikka-style indexed epistemic relation for Alice. Then, for any world x, $\{y : xR_Ay\}$ is a set of possible worlds compatible with what Alice believes to be the case in x. Let us say that the sentence '*Alice believes that* p' is uttered in the basic context C (where C is a set of possible worlds in each of which the sentence was uttered). Then, the derived context for interpreting '*that* p' will be the union of Alice's belief states in C, i.e., $\bigcup_{x\in C}\{y : xR_Ay\}$. Given this, the principle of non-triviality with respect to '*that* p' can be formulated as follows:

(6) *The principle of non-triviality*
a proposition p is ascribable to an individual A in a context set C iff
$\bigcup_{x\in C}\{y : xR_Ay\} \cap p \neq \emptyset$ and $\bigcup_{x\in C}\{y : xR_Ay\} \cap W - p \neq \emptyset$

To say that a proposition is "ascribable to some individual in a context" is to say that the speaker of the utterance that takes place in that context can use a sentence that expresses this proposition as a complement of an attitude predicate.

Let me illustrate the principle of non-triviality, using a simple example. Suppose Alice at some point truthfully says *I believe Mabel is having vertigo*. Suppose we have three possible worlds α, β, and γ, such that in α and β, Mabel indeed is feeling dizzy and light-headed but not in γ. Suppose

[14](Stalnaker 1999, 157)

further that the context set C consists of two possible worlds α and γ (that is, $C = \{\alpha, \gamma\}$) and that Alice's accessibility relation R_A relates every world to itself and α and β to each other (that is, $R_A = \{\langle\alpha, \alpha\rangle, \langle\alpha, \beta\rangle, \langle\beta, \beta\rangle, \langle\beta, \alpha\rangle, \langle\gamma, \gamma\rangle\}$). To distinguish α from β, let us say that a fair coin was tossed and the results came out as Heads in α and Tails in β and γ, but we are not interested in Alice's belief about the results of the coin toss. This setup is schematised in Figure 1 (where Vm = Mabel is having vertigo, H = Heads, T = Tails).

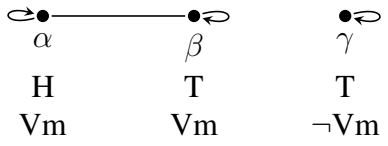

Figure 1: Illustration of the principle of non-triviality, applied to Alice's utterance 'I believe Mabel is having vertigo'. The utterance is not trivial since it excludes γ from the context set $\{\alpha, \gamma\}$. (Vm = Mabel is having vertigo, H = Heads, T = Tails)

In this setup, the principle of non-triviality is satisfied since the derived context (the union of possible worlds accessible from the context set C) is $\{\alpha, \beta, \gamma\}$ and it is compatible with both Mabel having vertigo and Mabel not having vertigo. Also, the belief ascription is true in the actual world α since $\{\alpha, \beta\} \subseteq \{x : \text{Mabel is having vertigo in } x\}$. We can say that the role of the belief ascription is to eliminate the possible world γ, in which Alice believes that Mabel is not having vertigo, from the context set C.

The principle of non-triviality as it is formulated in (6) bears a close resemblance to a more familiar uncertainty presupposition that is proposed for some attitude predicates. To see this, suppose that the context set C consists only of one world – the actual world α. Then the principle of non-triviality reduces to:

(7) a proposition p is ascribable to an individual A in a context set C iff
$\{x : \alpha R_A x\} \cap p \neq \emptyset$ and $\{x : \alpha R_A x\} \cap W - p \neq \emptyset$

I will discuss the connection between the principle of non-triviality and the uncertainty presupposition in section 6. There, I will argue that using the principle of non-triviality puts us in a better position for explaining subject obviation than using the uncertainty presupposition.

The last point to discuss in this section is in what sense the principle of non-triviality is a property of "general intelligence". Recall that, following Ruwet, I want to argue that the source of subject obviation is a property of "general intelligence" rather than linguistic competence. The property of "general intelligence" that Ruwet uses is iconicity, which is "general" in the sense that it applies to *all* representations and is not limited to linguistic representations or, as I put it earlier, not part of linguistic competence. The principle of non-triviality is "general" in exactly the same sense: it applies to all representations. In other words, "general intelligence", as I define it here, is a set of features and mechanisms that allow us to understand (or theorise about) behaviour and reasoning of rational agents independently of the means they are using to express their behaviour or reasoning (including the independence from language). That is to say, these are features and mechanisms that we can use to understand (or theorise about) the behaviour and reasoning of non-linguistic creatures with cognitive capacities comparable to ours, as well as the behaviour and reasoning expressed through language.[15]

[15]The relation between "general intelligence" and linguistic competence is a complicated matter. It relates to most

In this section, I formulated the principle of non-triviality in Stalnaker's framework that is explicitly designed to represent speech acts and contents of propositional attitudes in a way that is independent of language (or independent of the "vehicle" to use Stalnaker's own metaphor) that rational agents use to express their thoughts. Stalnaker maintains that "it is possible and fruitful to theorise about the structure and function of discourse independently of specific theory about the mechanisms that languages use to serve those functions".[16] He calls this position a thesis of *an autonomy of pragmatics* and defends it in his 2014 book. One of the outcomes of the autonomy of pragmatics thesis is that there might be linguistically manifested phenomena that are better explained as consequences of general reasoning than of linguistic competence. In this paper, I suggest that subject obviation is one such phenomenon.

4 Phenomenal information and self-locating knowledge

In the previous section, we discussed the principle of non-triviality which, I want to suggest, is the source of subject obviation. This principle disallows ascription of propositions whose content cannot be used to exclude any possibility from a belief state. In this section, I will discuss what makes embedded propositions in obviative sentences non-ascribable. As mentioned earlier, Ruwet (1984) and several scholars after him had an intuition (correct in my view) that obviative sentences are unnatural because there is an illegitimately tight "self-to-self" relation between the subject of the main clause and the subject of the embedded clause. This intuition will be the starting point of our discussion in this section. After presenting various proposals for capturing this intuition in the literature, I will suggest that these proposals can be subsumed under the idea that embedded propositions in obviative sentences convey phenomenal information. The goal of the remainder of this section, then, will be to introduce phenomenal information and its analysis in terms of self-locating knowledge.

Describing the illegitimately tight "self-to-self" relation between the two coreferential subjects in obviative sentences, Ruwet first of all points out that the distancing that the two subjects lack does not amount to different interpretations of indexical expressions, as in the famous McCawley example *I dreamed that I was Brigitte Bardot and that I was kissing myself*. Nor is it the same as the distinction sometimes made between "a person and his portrait [. . .], an actor and the character he plays" or "an author and his literary work".[17] What he has in mind is a finer distinction between

fundamental debates in the philosophy of language, including debates about the language of thought, the problem of intentionality, and the theory of meaning. I do not want to raise these issues here because they will lead us far beyond what I set to do in this paper. However, I want to clarify that my proposal to account for subject obviation in terms of "general intelligence" is more profound than the iteration of the accepted view in the recent literature that subject obviation is due to a pragmatic restriction. As far as I can tell, pragmatic restrictions in these accounts are formulated as integral parts of linguistic competence; they either govern speech acts or are encoded into the meaning of a lexical item. By contrast, the principle of non-triviality operates on non-linguistic representations as well. Consider, for example, wondering whether there is someone in the next room (without putting this question into words). Upon entering the empty room, you automatically do the reasoning and exclude the possibility of someone being there (presumably, without using language). I thank an anonymous reviewer for asking how my account relates to the existing pragmatic accounts of subject obviation, which led to this clarification. I agree with the reviewer that the answer depends on which theory of the semantics-pragmatics interface is assumed. As noted in fn. 13, Stalnaker's framework used in this paper assumes the contextualist view.

[16](Stalnaker 2014, 1)

[17](Ruwet 1991, 12)

"the soul, broadly construed, and the body and their conflation in the person considered globally" or the distinction drawn by Plato between the rational part of the soul and its desires.[18] This kind of distinctions, he says, can be present even in simple sentences like *I am torn between my love for my family and my love for my country* or *Hey, I am talking to you*. Later on, Szabolcsi will call this discontinuity of the self "mind-boggling".[19]

Schlenker (2005) uses the concept of an event *de se* ('of self') interpretation to describe the tight "self-to-self" relation in obviative sentences, which he borrows from Higginbotham (2003). The examples in (8) illustrate the contrast between a simple *de se* and an event *de se* interpretation. Whereas the sentence in (8)a requires only that the speaker remembers the fact of his going to school in the 5th grade and thus can be truthfully uttered by any adult, the sentence in (8)b can be asserted only by an adult with an exceptionally good memory who remembers a particular event of his going to school at this young age.

(8) a. I remember that I walked to school in the 5th grade. (simple *de se*)
b. I remember walking to school in the 5th grade. (event *de se*)

Schlenker's explanation of subject obviation is based on the competition conjecture, which, as we saw earlier, is not viable. However, his intuition that subjunctive clauses in obviative sentences, in addition to having a simple *de se* interpretation, include a perspectival component associated with an event *de se* interpretation – so that the event is seen from within – is (I think) correct and points to the same idea that Ruwet tries to explain in his paper.

Szabolcsi (2021) makes another step towards clarifying the nature of the "self-to-self" relation in obviative sentences. She adds to the picture Hungarian data that directly point to the experiential nature of the information that is conveyed by embedded propositions in obviative sentences. Here are some of her examples:

(9) a. Remélem, hogy benne vagyok a csapatban. (Hungarian)
'I hope that I'm on the team'
b. Remélem, hogy nem untatlak.
'I hope that I'm not boring you'
c. Remélem, hogy biztonságban vagyok.
'I hope that I'm safe'
d. #Remélem, hogy fél lábon állok.
'I hope that I'm standing on one leg'
e. #Remélem, hogy (nem) szédülök.
'I hope that I (don't) have vertigo'
f. #Remélem, hogy (nem) fázom.
'I hope that I'm (not) cold'
g. #Remélem, hogy ugrándozok.
'I hope that I'm jumping around'
h. #Remélem, hogy simogatom a macskát.
'I hope that I'm stroking the cat' (Szabolcsi 2021, 10)

All the unacceptable examples in (9) have what Szabolcsi calls "classical predicates of direct

[18](ibid.)
[19](Szabolcsi 2021, 12)

experience", which convey the agent's own perceptual or cognitive state.[20] This is what, according to Szabolcsi, makes them obviative.

Finally, discussing subject obviation with epistemic attitudes in Italian, Costantini (2016, 2023) postulates that a sentence is obviative when the embedded proposition is arrived at through introspection. Introspection here is described as a process pertaining to self-knowledge that gives the believer "a direct access to mental states and is highly epistemically secure".[21]

I want to suggest that the various ways of describing the illegitimately tight "self-to-self" relation in obviative sentences that we saw above can be subsumed under the idea that embedded clauses in obviative sentences contain *phenomenal information*. "Phenomenal" is a term that philosophers use to describe a type of knowledge of what it is like to be in a particular cognitive state.[22] For example, what it is like to be in a cognitive state corresponding to seeing the colour red, or feeling cold, or experiencing vertigo. There are several approaches to explaining phenomenal knowledge in philosophy of mind and epistemology; some of them attempt to reduce phenomenal knowledge to more familiar self-locating knowledge (e.g., Perry 1999; Stalnaker 2008, 2014). In this paper, I will adopt the reductionist approach developed by Stalnaker. An advantage of his reductionist approach (for us) is that it is framed in a model suitable for explaining linguistic phenomena.

Self-locating information is information about who and where we are in the world. Self-location can be with respect to place, time, or a person's own identity. This kind of information is usually associated with the interpretation of indexical expressions (*I, here, today*) and is contrasted with ordinary (impersonal) belief. This contrast is usually illustrated using intricate scenarios involving learned amnesiacs, mountain-dwelling gods, spilled goods, or garments on fire.[23] However, Stalnaker (among others) correctly points out that self-locating knowledge is more ubiquitous than these exotic scenarios make us believe. Self-locating knowledge is usually taken for granted, and its presence becomes noticeable only in its absence. Here's Stalnaker's own mundane example of self-locating knowledge that shows the ubiquity of this kind of knowledge:

> "It is Monday afternoon. After shopping in the mall, I take the elevator down to level B of the parking garage. I had gone up a different elevator, one in the centre of the garage. The one I came down is either at the east or the west end, I am not sure which – there is an elevator at each end. I know my car is about in the middle along the northern edge, but is that to the right or to the left? I have a clear mental map of level B, but it has no 'you are here' marker, so I don't know how to orient myself on it. The garage is pretty symmetrical, so it is hard to tell by looking around just where I am. I do know that there is a pale green Prius with Massachusetts license plate [...] to my right as I come out of the elevator, but knowing that does not help, since of course my mental map of level B does not tell me what cars are parked in what places." (Stalnaker 2014, 121)

In this situation, the agent lacks self-locating knowledge because no objective information available from his position can allow the agent to choose between two epistemic possibilities: an actual world in which, say, the agent descends at the east end and has to go right to find his car at the northern

[20](ibid., 10)

[21](Costantini 2023, 33)

[22]The famous illustration of phenomenal knowledge and its (alleged) difference from ordinary belief is a thought experiment in which Mary, a brilliant scientist who knows everything there is to know about the physics and physiology of colour-seeing, but who was raised in a black-and-white world, sees the colour red for the first time (Jackson 1982).

[23](E.g., Castañeda 1966; Perry 1977, 1979; Lewis 1979, among others.)

edge, and a counterfactual possibility in which the agent descends at the west end and has to turn left to find his car.

To analyse self-locating knowledge, Stalnaker (2008, 2014) uses the concept of centred possible worlds proposed by Lewis (Lewis 1979). A centred possible world is a pair $\langle c, x\rangle$ where x is a possible world that represents an objective possible situation and c is a person with whom the knower or believer identifies themselves in x. In Lewis's system, centred possible worlds are more fine-grained objects than uncentred possible worlds in the sense that a single uncentred possible world can correspond to a number of centred possible worlds. This is what happens when the agent is confused about his or her self-location, as in the parking garage story above. The distinction between centred and uncentred possible worlds allows Lewis to capture the intuitive difference between ordinary and self-locating knowledge. The content of self-locating knowledge is a proposition centred on the individual whose knowledge is being represented. By contrast, the content of ordinary knowledge is a proposition where centres are irrelevant.[24]

Although Lewis's system (commonly assumed in formal semantics) is successful in capturing the intuitive difference between ordinary and self-locating knowledge, the system, as it stands, has serious limitations. As Stalnaker (2008) points out, because in Lewis's system the content of ordinary and self-locating knowledge is represented by different kinds of propositions (i.e., a set of centred worlds where the centre is irrelevant versus a set of centred worlds with the believer at the centre), the system is solipsistic. It cannot be used to explain how self-locating information is communicated, or how ordinary and self-locating knowledge of different individuals is integrated, or even how self-locating beliefs of one individual are revised. As an illustration of self-locating knowledge being communicated, suppose that in the parking garage story above, the unfortunate agent confused about whether he must turn left or right to find his car sees a young couple and asks them whether he is at the east or west end of the garage. They reply, 'You are at the west end'. There is nothing self-locating about the couple's reply, but upon hearing and accepting their reply, the agent acquires self-locating information and concludes that he must turn left to find his car.[25]

To overcome these limitations, Stalnaker proposes a number of modifications of Lewis's system. Here, I discuss two of his modifications that are immediately relevant to the topic of this paper. The first modification has to do with how an accessibility relation between possible worlds specifies whose belief is being represented. Lewis uses the classical Hintikka-style accessibility relation that specifies whose beliefs are being represented by an index on the relation, e.g., xR_Ay says that y is compatible with Alice's beliefs in x. However, the Hintikka-style accessibility relation is insufficient for capturing intentionality of belief, in general, and is especially problematic for self-locating belief. For example, if Alice is talking to Bennett, her beliefs about the beliefs of the person she is talking to may be different from her beliefs about what Bennett believes. This may be the case when Alice does not realise that she is talking to Bennett. In this situation, we might assent to '*Alice believes that the person she is talking to believes that* p' without assenting to '*Alice believes that Bennett believes that* p'. The classical indexed accessibility relation does not allow us

[24]More precisely, Lewis uses centred properties broadly construed (Lewis 1979). But, since there is a one-to-one correspondence between his properties and possible worlds, it is proper to talk about centred possible worlds. It is also worth noting that in Lewis's system objective information is a special case of self-locating information (a case where the centre is irrelevant).

[25]In principle, Lewis's story can be adjusted to account for the possibility to communicate and update self-locating belief. But because Lewis uses modal realism, the adjustments will blur the line between ordinary and self-locating belief and we will lose a natural way of addressing the problem of intentionality (see **sta11?**).

to express this difference.[26]

Stalnaker proposes that there is only one accessibility relation established between centred possible worlds. That is, $\langle A, x\rangle R\langle B, y\rangle$ says that A in x locates herself as B in y. Unlike Hintikka and Lewis, Stalnaker specifies the believer in the relata rather than on the accessibility relation; see (10).[27]

(10) a. Hintikka/Lewis: y is compatible with what A knows in x iff xR_Ay
b. Stalnaker: y is compatible with what A knows in x iff for some centre c, $\langle A, x\rangle R\langle c, y\rangle$

With this modification, we can represent the difference between Alice's beliefs about the beliefs of the person she is talking to and Alice's beliefs about Bennett's beliefs (even if she is talking to Bennett). Suppose f is an individual concept (i.e., a function from a possible world to an individual) that, for any possible world, picks out the person with whom Alice is talking in that world. Then, the beliefs of the person Alice is talking to will be centred on the individual (whoever it might be) who is the value of f; see (11)a. At the same time, Bennett's beliefs (according to Alice) will be centred on Bennett; see (11)b. These two sets of beliefs do not have to be the same.

(11) a. '*Alice believes that the person she is talking to believes that* p' is true in α iff
for all worlds x and y and all subjects C,
if $\langle A, \alpha\rangle R\langle A, x\rangle$ and $\langle f(x), x\rangle R\langle C, y\rangle$, then $y \in p$
b. '*Alice believes that Bennett believes that* p' is true in α iff
for all worlds x and y,
if $\langle A, \alpha\rangle R\langle A, x\rangle$ and $\langle B, x\rangle R\langle B, y\rangle$, then $y \in p$

The second modification proposed by Stalnaker changes the relation between centred and uncentred possible worlds. As mentioned above, for Lewis, centred possible worlds are fine-grained objects, in the sense that multiple centred possible worlds may correspond to one uncentred possible world. For Stalnaker, however, there is one-to-one correspondence between centred and uncentred possible worlds; see (12).[28] In other words, in Stalnaker's system, if an individual locates herself differently in a possible world (resulting in two distinct centred possible worlds), this corresponds to two distinct uncentred possible worlds.[29]

(12) *One-to-one correspondence principle*
for all $c, c', c'' \in S$ and all $x, y \in W$, if $\langle c, x\rangle R\langle c', y\rangle$ and $\langle c, x\rangle R\langle c'', y\rangle$, then $c' = c''$

To sharpen the difference between Lewis's and Stalnaker's views, let me illustrate it with a vivid analogy using Stalnaker's own words:

> "A misleading picture sometimes accompanies the Lewis account of self-locating belief: belief about what possible world you are in is like belief about what country you are in, while belief about where in the world you are is like a more specific belief about where in the country you are (what village, street corner, or mountain top). But ordinary belief about where you are in the world is always also belief about what possible world you are in (what possible state of the world is actual). If I am not sure as I drive along

[26](Stalnaker 2014, 120)
[27](ibid., 119)
[28](ibid., 118)
[29]This modification reflects the disagreement between Stalnaker and Lewis about the metaphysics of possible worlds discussed at the beginning of section 3.

> the highway toward New York, whether I am still in Massachusetts, then I am not sure whether I am in a possible world in which this stretch of highway is located in Massachusetts." (Stalnaker 2008, 51)

The "misleading picture", Stalnaker argues, is due to the fact that we erroneously accept the *principle of phenomenal indistinguishability*. He formulates the principle as follows: "If a possibility is an epistemic alternative for a knower at a time (that is, it is compatible with his or her knowledge), then it is phenomenally indistinguishable from the actual world to the knower at that time".[30] This principle comes from the empiricist's view that we have direct epistemic access to phenomenal experiences or to our internal world. Stalnaker argues that our knowledge of the internal world is just as indirect as our knowledge of the external world; therefore we should be skeptical about the principle of phenomenal indistinguishability.

Here is one of the arguments that Stalnaker makes against the principle of phenomenal indistinguishability (adapted for our purposes). Suppose Alice is a brilliant medical student who knows everything there is to know about physiological processes of various illnesses. But she was born with a sensory deprivation disorder so severe that she never had any experiences, such as pain, sadness, dizziness, tranquility, cheerfulness, happiness, or euphoria. Alice agrees to participate in an experiment where she will be given a drug inducing one of two conditions: vertigo (in which case she will experience dizziness and lightheadedness) or mania (in which case she will experience exhilaration and euphoria). Alice does not know which drug she will be given, but she knows that the choice will depend on a toss of a fair coin: if Heads, she will be given a vertigo-inducing drug; if Tails a mania-inducing drug. The argument goes as follows: because Alice has never experienced any sensations, both before and after the experiment she is in a state of ignorance, not able to distinguish between the epistemic possibility where the coin landed Heads and she was given a vertigo-inducing drug, and the one where the coin landed Tails and she was given a mania-inducing drug. These two possibilities are depicted in Figure 2 (where Va = Alice was given a vertigo-inducing drug, Ma = Alice was given a mania-inducing drug, H = Heads, T = Tails). The worlds α and β in Figure 2 differ both physically and phenomenally.

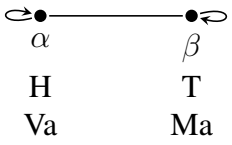

Figure 2: Illustration of Alice's state of ignorance, both before and after the experiment. Alice is not able to distinguish between α and β, because she is not in a position to recognise either lightheadedness associated with vertigo or euphoria associated with mania. (Va = Alice was given a vertigo-inducing drug, Ma = Alice was given a mania-inducing drug, H = Heads, T = Tails)

Stalnaker argues that to accept the principle of phenomenal indistinguishability is to suppose that there is a possible world β* where the coin landed Tails and Alice was given a mania-inducing drug, but she had vertigo-associated sensations. In other words, β* is physically like β but phenomenally like α. Those who accept β* as an epistemic possibility, Stalnaker says, accept the reality of phenomenal experiences. I find Stalnaker's argument that β* is not a viable epistemic possibility in the situation above convincing (assuming that the drug worked as expected and Alice did not

[30](Stalnaker 2008, 88)

have an abnormal reaction to it). Thus, I agree that we should abandon the principle of phenomenal indistinguishability and endorse, instead, the position that if an agent is ignorant or confused with respect to phenomenal (or self-locating) information, the agent is ignorant or confused about which possible world she is in. This position is reflected in the one-to-one correspondence principle.

Let me show how the two modifications discussed above work, using a simple example with self-ascription of phenomenal information. Suppose Alice, at a particular moment, says or thinks to herself:

(13) I believe I am having vertigo.

Our modified (partial) knowledge model is a tuple $\langle W, S, E, R\rangle$, where W and S are, as above, a (non-empty) set of possible worlds and a (non-empty) set of individuals or subjects respectively. E is a set of centred possible worlds (i.e., a set of pairs $\langle c, x\rangle$ where $c \in S$ and $x \in W$) that meets the following condition: the individual represented by the centre c exists in the possible world x of that centre. R is an epistemic accessibility relation (transitive, Euclidean, serial) that differs from the classical Hintikka-style accessibility relation in two respects. First, it is a relation on E (i.e., it relates centred possible worlds rather than uncentred possible worlds), and the individual whose belief is represented is specified in the relata rather than as an index (the first modification). Second, R relates centred possible worlds in a way that satisfies the one-to-one correspondence principle; that is, if a possible world has a different centre, this corresponds to a distinct uncentred possible world (the second modification).

In our example in (13), Alice plays three roles: the speaker, the attitude holder, and the person experiencing vertigo. As a speaker, Alice assumes the basic context for the utterance. Let us say the basic context C that Alice assumes consists of two worlds α and γ. In this basic context, Alice, the speaker, represents beliefs of the attitude holder (who happens to be Alice herself). This representation of Alice's belief states in C determines the derived context for the embedded proposition *that I am having vertigo*. Let us say that the accessibility relation R in our present case is the same as in Figure 1. That is, the derived context is the set $\{\alpha, \beta, \gamma\}$; see Figure 3. To satisfy the principle of non-triviality, the derived context must include an epistemic alternative in which Alice, the experiencer, is not having vertigo. Since having vertigo is phenomenal (or self-locating) information, the epistemic alternative in which Alice is not having vertigo cannot be a possible world in which Alice locates herself as Alice with the same "experiential profile"[31] as the one she has in the actual world. This is because we abandoned the principle of phenomenal indistinguishability (see above). The epistemic alternative in which Alice is not having vertigo must be a possible world in which Alice locates herself as a different individual, possibly with the same name, shared memories, etc., but with a different "experiential profile" at the relevant time. For simplicity, I will represent Alice with a different "experiential profile" than the one in the actual world as Mabel.

To express Alice's ignorance (or confusion) about her own identity, we can use what Stalnaker calls an I-concept. An I-concept with respect to a possible world x is an individual concept f (a function from a possible world to an individual) whose value for any possible world y accessible from x is the centre of y. That is, f is an I-concept with respect to x iff for any world y and any individual B, if $\langle f(x), x\rangle R\langle B, y\rangle$, then $B = f(y)$. An I-concept may be rigid; this occurs when the believer is not ignorant or confused about their self-location or phenomenal experience. In this

[31] (E.g., Godfrey-Smith 2020.)

case, the value of f is the same individual in all worlds accessible from x. An I-concept may be non-rigid; this is the case when the believer is ignorant or confused about their self-location or phenomenal experience. In the latter case, the value of f differs for different worlds accessible from x. In Alice's case, which we are discussing here, the I-concept that represents Alice with respect to the actual world α is non-rigid: it picks Alice in α and β but Mabel in γ (that is, $f(\alpha)$ = A, $f(\beta)$ = A, $f(\gamma)$ = M).[32] The full setup is summarised in Figure 3 (where Va = Alice is having vertigo, Vm = Mabel is having vertigo, H = Heads, T = Tails).

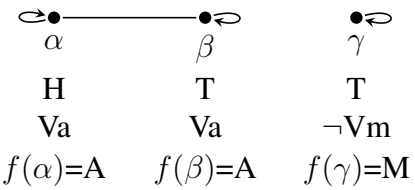

Figure 3: Illustration of the principle of non-triviality, applied to Alice's utterance 'I believe I am having vertigo'. The utterance is non-trivial because the derived context $\{\alpha, \beta, \gamma\}$ contains the world γ, in which Alice, who locates herself as Mabel in that world, is not having vertigo. (Va = Alice is having vertigo, Vm = Mabel is having vertigo, H = Heads, T = Tails)

In this setup, the principle of non-triviality is satisfied since the derived context (the union of possible worlds accessible from the basic context set C) is $\{\alpha, \beta, \gamma\}$, and it is compatible with both propositions *that I am having vertigo* (true in α and β where Alice locates herself as Alice) and *that I am not having vertigo* (true in γ where Alice locates herself as Mabel). That is, $\bigcup_{x \in C}\{y : xR_f y\} \cap \{z : f(z)$ is having vertigo in $z\} \neq \emptyset$ and $\bigcup_{x \in C}\{y : xR_f y\} \cap W - \{z : f(z)$ is having vertigo in $z\} \neq \emptyset$ (where $xR_f y$ iff $\langle f(x), x\rangle R\langle f(y), y\rangle$ for any x and y). Also, the belief ascription as a whole is true in the actual world α since $\{\alpha, \beta\} \subseteq \{x : f(x)$ is having vertigo in $x\}$. Finally, we can say that the role of the belief ascription is to eliminate from the context set C the possible world γ, in which Alice believes of herself that she is not having vertigo.

Before concluding this section, let me briefly mention that to capture the familiar cases of self-locating knowledge in this framework (the cases with learned amnesiacs, mountain-dwelling gods, etc., including the case where one is lost in the underground parking garage), we must state that the alternative with the different centre is accessible from the world of evaluation. For example, in Figure 3, this amounts to saying that γ is accessible from α (i.e., $\langle\langle A, \alpha\rangle, \langle M, \gamma\rangle\rangle \in R$). In this case, Alice cannot sincerely say, 'I believe that I am having vertigo'.

Let me summarise our discussion so far. I started in section 2 with the explanation of subject obviation proposed by Ruwet (1984). The exposition of his view was important because, as I said, my own explanation of subject obviation resembles his in key features. Like Ruwet, I attribute

[32] This is a simplified picture. Because γ is not accessible from α, we need two I-concepts in C: one, say f_1, for Alice's representation of herself with respect to α, and the other one, say f_2, for Alice's representation of herself with respect to γ. Such a setup is more expressive and allows us to capture cases in which Alice "chooses" between being in a knowledgeable or ignorant cognitive state. However, this refinement complicates the representation of indexical expressions, which I represent simply as the value of an I-concept f; for example, representing *I am having vertigo* as $\{x : f(x)$ is having vertigo in $x\}$. We independently need a more elaborate representation for indexical expressions, and there is extensive linguistic work on this topic. However, I leave this discussion outside of the scope of this paper, as it is an independent issue that, once suitably settled, can be easily incorporated into the explanation of subject obviation proposed here.

subject obviation to a property of "general intelligence". And also like Ruwet (and several scholars after him), I say that obviative sentences are unnatural because there is an illegitimately tight "self-to-self" relation between the two coreferential subjects. Then, I proceeded to elaborate on each of these two key features. In section 3, I discussed the principle of non-triviality, which, I suggested, is the property of "general intelligence" to which subject obviation can be attributed. The principle of non-triviality disallows ascription of a proposition whose content does not exclude any possibility from a belief state. I showed how this principle can be formulated in the framework developed by Stalnaker (Stalnaker 1978, 1988, 2014). After that, in section 4, I discussed phenomenal (or self-locating) information to argue that this kind of information makes embedded propositions in obviative sentences non-ascribable. Self-locating knowledge is also captured in Stalnaker's system (Stalnaker 2008, 2014). In the next section, I show how the apparatus presented in the preceding sections enables us to explain subject obviation.

5 Obviative, non-obviative, and ameliorated sentences

The time has come to look at obviative, non-obviative, and ameliorated sentences. Above, I discussed self-ascriptions like *I believe I am having vertigo* as uttered by Alice. I said that in order to satisfy the principle of non-triviality, Alice's cognitive state in the context should include an epistemic possibility in which Alice is not having vertigo. Since having vertigo is phenomenal (or self-locating) information, as discussed in the previous section, the epistemic possibility in which Alice is not having vertigo is also a possible world in which Alice locates herself as an individual with a different "experiential profile".

But suppose Alice utters the same sentence in a context that forbids her from doubting or being confused about her self-location. In this case, the self-ascription becomes unnatural. Consider the sentences in (14) uttered in a non-pathological, non-dynamic situation that rules out Alice's slightest doubt about how she feels in the world.

(14) [Context: a non-pathological, non-dynamic situation, in which Alice is fully aware of her perceptual or cognitive state and has no doubts about it.]
   a. #I believe I am having vertigo.
   b. #I believe I am sane.
   c. #I believe I am standing on one leg.
   d. #I believe I am jumping around.

The examples in (14) sound unnatural because there is tension between the principle of non-triviality, which asks us to suppose that there is a possibility in which Alice is unaware of her cognitive state, and the context (or assumed common ground) that suppresses this supposition.

I want to suggest that this kind of tension is what makes sentences with coreferential subjects and self-locating information in the embedded clause obviative. Consider again subject obviation in Hungarian, repeated from (9). The embedded propositions in these examples convey phenomenal information, as we discussed earlier.

(15) a. #Remélem, hogy (nem) szédülök. (Hungarian)
   'I hope that I (don't) have vertigo'
   b. #Remélem, hogy fél lábon állok.
   'I hope that I'm standing on one leg'

c. #Remélem, hogy (nem) fázom.
'I hope that I'm (not) cold'
d. #Remélem, hogy ugrándozok.
'I hope that I'm jumping around'
e. #Remélem, hogy simogatom a macskát.
'I hope that I'm stroking the cat' (Szabolcsi 2021, 10)

Now, suppose that 'hope' in Hungarian is a predicate that induces a "particularly intimate" relation between the attitude holder and the subject in the embedded clause.[33] This "particularly intimate" relation disallows the attitude holder from viewing herself as distinct individuals in the context in which an attitude ascription with 'hope' is made. Let me call such "particularly intimate" attitude predicates *rigid* attitudes and, using the apparatus from the previous section, define them as predicates which admit only a rigid I-concept in the context in which they are used:[34]

(16) An attitude predicate is *rigid* iff
for all $x, y \in C,\ f(x) = f(y)$
where f is an I-concept

My proposal is that obviative sentences are sentences in which there is an attempt to ascribe phenomenal (or self-locating) information using a rigid attitude. This cannot be successful because any ascription should satisfy the principle of non-triviality. To satisfy the principle of non-triviality when self-locating information is ascribed, there should be an epistemic alternative in which the agent locates herself differently than in the actual world; however, a rigid attitude forbids varying self-location in the context. This results in an unresolvable tension, causing us to perceive the sentence as unnatural.

Let me illustrate this using the English gloss of the Hungarian example in (15)a. Suppose Alice says '#I hope I am having vertigo'. We can describe Alice's epistemic situation using the setup that we used for (13); see Figure 4 (where Va = Alice is having vertigo, Vm = Mabel is having vertigo, H = Heads, T = Tails). The difference between our present (obviative) sentence and the sentence in (13) is the type of the attitude. 'Hope' is a rigid attitude which disallows a non-rigid I-concept. The sentence with 'hope' is irreparably unnatural because if it is evaluated in the context set containing only α and β, the principle of non-triviality is violated; at the same time, 'hope' cannot be used in the context set containing α and γ because 'hope' is rigid. Note that an epistemic alternative where Alice, representing herself as Alice, is not having vertigo is unavailable, as it would require accepting the principle of phenomenal indistinguishability, which, as we saw above, is unfounded.

If the proposal that subject obviation is an interplay between the type of an attitude and the type of ascribed information (within the frame set by "general intelligence") can be maintained, there are two questions that need to be answered: (a) what makes a particular attitude (in a particular language) rigid, and (b) how do we draw a line between phenomenal and objective information? These are complicated questions, and in this paper, I will not be able to give a satisfactory response to either of them. But in sections 6 and 7, I will try to show that my way of formulating these questions places us in a better position to answer them compared to recent semantic-pragmatic accounts of subject obviation.

[33] This is an expression used by Ruwet in connection with French *vouloir* 'want' (Ruwet 1991, 18).

[34] In a more elaborate version of my proposal (see fn. 32), the rigidity condition should be reformulated as a condition on having identical I-concepts in the same context.

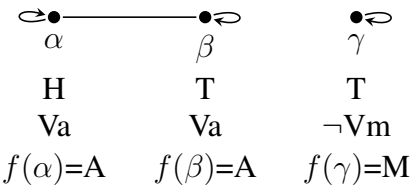

Figure 4: Illustration of unresolvable tension in case of obviative sentences such as Alice's utterance '#I hope I am having vertigo'. In the context $\{\alpha, \beta\}$, the principle of non-triviality is violated. In the context $\{\alpha, \gamma\}$, a rigid attitude cannot be used. (Va = Alice is having vertigo, Vm = Mabel is having vertigo, H = Heads, T = Tails)

Let us now look at ameliorated sentences, repeated in (17) from above. In these sentences, the embedded proposition does not involve phenomenal (or self-locating) information. So, the principle of non-triviality can be satisfied without introducing a non-rigid I-concept in the context. The absence of the tension between the need to satisfy the principle of non-triviality and the requirement on a rigid attitude makes these sentences more acceptable.

(17) a. Remélem, hogy benne vagyok a csapatban. (Hungarian)
'I hope that I'm on the team'
b. Remélem, hogy nem untatlak.
'I hope that I'm not boring you'
c. Remélem, hogy biztonságban vagyok.
'I hope that I'm safe' (Szabolcsi 2021, 10)

Again, as a quick illustration, let us consider the English gloss of the example in (17)a. Suppose Alice says 'I hope I am on the team'. In this case, we can represent Alice's epistemic situation as in Figure 5 (where Ma = Alice is on the team, H = Heads, T = Tails). Since being on the team is not self-locating information, adding an epistemic possibility in which Alice is not on the team does not require Alice to locate herself differently from the actual world. Therefore, there is no tension between satisfying the principle of non-triviality and adhering to the requirement for a rigid attitude.

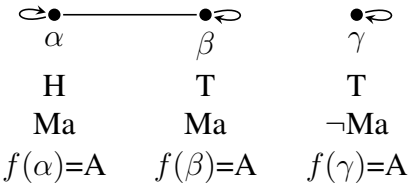

Figure 5: Illustration of an ameliorated sentence such as Alice's utterance 'I hope I am on the team'. There is no tension between the principle of non-triviality and rigidity of 'hope' because the content of ascribed information is not self-locating. (Ma = Alice is on the team, H = Heads, T = Tails)

At this point, let me briefly comment on Ruwet's empirical observation that ameliorated sentences have varying degrees of acceptability and that different factors (and their combinations) may contribute to the amelioration effect. This observation is in line with my proposal in this paper. Amelioration occurs when the embedded clause is not interpreted as conveying phenomenal (or self-locating) information. In the framework assumed in this paper, the divide between self-locating and objective content is not the divide in a *type* of content. In Stalnaker's (and Lewis's) view, self-locating and ordinary belief are of the same type (both are sets of centred possible worlds). The

difference is that in ordinary belief the centre can be ignored. What sets self-locating belief aside is the subject's relation to the content of this belief and not the content itself. Of course, there are clear cases of phenomenal and clear cases of objective information, on which we will all agree. For example, having vertigo versus being on the team. But there are also borderline cases like being safe or being allowed to do something, whose classification as phenomenal versus objective depends on their interpretation. I think this dependence on interpretation is what accounts for varying degrees of acceptability of ameliorated sentences.

To complete the picture, let us consider a non-obviative counterpart of (15)a where the subject of the main predicate and the embedded subject refer to different individuals. Suppose Alice says in Hungarian 'I hope Mabel is having vertigo'. Then, Alice's epistemic situation can be represented as in Figure 6 (where Vm = Mabel is having vertigo, H = Heads, T = Tails). Of course, Mabel's having vertigo is not self-locating information for Alice, so the tension that we have in obviative sentences does not arise.

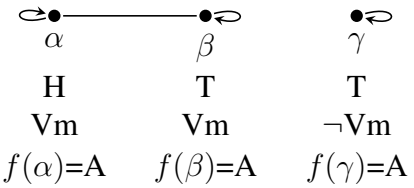

Figure 6: Illustration of a non-obviative sentence such as Alice's utterance 'I hope Mabel is having vertigo'. The sentence is unproblematic because 'that Mabel is having vertigo' is not self-locating for Alice. (Vm = Mabel is having vertigo, H = Heads, T = Tails)

Before concluding this section, let me briefly return to Ruwet's initial examples, repeated in (18), and show how my analysis explains these examples.

(18) a. #Je veux que je parte. (French)
'I want that I leave-SBJV'
b. Je veux que tu partes.
'I want that you leave-SBJV'
c. Je veux absolument que j'amuse ces enfants.
'I want absolutely that I amuse-SBJV these children' (Ruwet 1991, 2, 30)

To extend my proposal to these examples, we need to add two assumptions: First, following Heim (1992) and von Fintel (1999), we need to say that desire predicates like *vouloir* 'want' involve doxastic alternatives. Second, we need to show that intentional actions like *partir* 'leave' involve self-locating information. I consider the first assumption to be relatively uncontroversial, although there are analyses of desire predicates that do not directly invoke doxastic alternatives.[35] However, more needs to be said about the second assumption.

I argue elsewhere that the distinction between intentional actions like leaving and non-intentional actions like amusing children can be reduced to the distinction between self-locating and ordinary belief.[36] Amusing children is a non-intentional action in the sense that the outcome of the action (i.e., the children being amused) is not fully controlled by the agent. Intentional actions locate the

[35](E.g., Villalta 2008; Anand and Hacquard 2013; Condoravdi and Lauer 2016.)
[36](Author, Year)

agent with respect to his or her *reason for acting*.[37] That locating oneself with respect to reason for acting is similar to self-location with respect to identity, space, or time can be shown by the fact that the agent can be confused about his or her reason for acting. (Recall the confusion in the underground parking that I used to illustrate self-location with respect to space.) A typical case of confusion with respect to reason for acting is "fridge amnesia" – the experience of opening the fridge and momentarily forgetting why you did so. Like in the case of other self-locating knowledge, in the case of "fridge amnesia" the agent knows all relevant objective facts (e.g., that he is standing in his kitchen, holding the fridge door open and peering inside). Yet, the agent is unable to choose his next action because he is lacking self-locating information.

Due to space constraints, I will not give full details about how the examples in (18) are captured. But I hope that with the two assumptions above, it is clear how my analysis illustrated in Figures 4-6 using 'hope' and phenomenal information can be extended to Ruwet's initial cases involving desire predicates and intentional actions.[38]

A further question arising in relation to Ruwet's initial examples is what explains the acceptability of attitude reports with infinitival complements as in (19).[39] For Ruwet, sentences with infinitival and finite complements compete with each other for expressing a tight "self-to-self" relation between the matrix and the embedded subject. Sentences with finite complements are unnatural because there is a simpler way to express the same thought using an infinitival complement. As discussed in section 2, the competition conjecture cannot be maintained. So we need to find another answer to the question why (19) is acceptable.

(19) Je veux PRO partir. (French)
'I want to leave-INF'

The alleged problem with (19) stems from the common assumption that PRO has an essentially *de se* interpretation. That is, its interpretation is similar to that of *I*. Thus, without the competition conjecture, we predict that (19) should be as bad as (18)a. However, as we are about to see, the assumption about PRO is not uncontroversial. If PRO is not essentially *de se*, the problem disappears because we no longer have good reason to predict that (19) should be unacceptable.

The assumption that PRO is essentially *de se* goes back (at least) to Morgan (1970) and Chierchia (1989). It is based on the observation that infinitival complements cannot be used in misidentification scenarios. For example, it is reported that in the scenario in (20) where John does not recognise himself as the engaging candidate, only (20)a can be used to describe John's expectations. (20)b in this scenario is false.[40]

(20) [Context: John is watching the speeches of various candidates in the upcoming election. He finds one candidate particularly engaging and thinks that this candidate will win. But because

[37]The notion "reason for acting" is borrowed from Anscombe (1957).

[38]An anonymous reviewer correctly points out that Ruwet's initial examples involve future-oriented attitude reports and asks how my analysis can be extended to future actions. At this point, the representation of future intentions remains unclear to me. However, I think the first step towards understanding future intentions is to recognise that expressions of the future are ambiguous. As Anscombe (1957) points out, sentences like *I am going to fail this exam* can be interpreted either as a prediction (e.g., I am so bad at this subject that I predict that I will fail the exam) or as an expression of intention (e.g., I might be good at the subject, but to annoy my parents, I intend to fail the exam). Obviative sentences seem to lack the prediction interpretation, indicating that future actions in these sentences are closely tied to present intentions.

[39]I thank an anonymous reviewer for raising this question.

[40]The example in (20) is adapted from Ninan (2010).

John is intoxicated, he does not realise that he is the candidate in question. In fact, he is pessimistic about his own prospects and thinks to himself, 'I'm not going to win the election.']

a. John expects that he will win the election.
b. #John expects to win the election.

However, Cappelen and Dever (2013) argue that the assumption above is incorrect: PRO is not essentially *de se*; it can be used in scenarios where the subject misrecognises himself. Here is their example that shows that PRO can be used in case of misrecognition. If we accept that the sentence *He expects to win the election* in their scenario is true, we should conclude that PRO is not essentially *de se*.

> "John is running for mayor of the local town. Earlier in the day, he gave a campaign speech, and he is now relaxing in the local pub. In an attempt to put the stresses of the campaign behind him, he has been drinking steadily all evening. The local news is showing on a television behind the bar, and is playing excerpts from John's speech. The speech is a complete disaster – full of objectionable policies and errors on matters of facts. It's obvious to us, sitting in the bar, that the speech has destroyed any chance John had of being elected. John, however, in his current inebriated state, finds the speech delightful. Too drunk to recognize himself on screen, he gestures broadly toward the television and declares, 'That guy's going to be our next mayor!'. I turn to you and say, 'Can you believe it? John's so drunk he actually expects to win the election.'" (Cappelen and Dever 2013, 163)

Of course, denying that PRO is essentially *de se* does not give us a positive analysis of PRO. We still want to understand what PRO is and how to explain the contrast in (20). My main point here is that infinitival constructions, such as in (19), are not part of the analysis of subject obviation because the connection between these two constructions is based on two questionable suppositions: the competition conjecture and the assumption that PRO is essentially *de se*.

6 Non-triviality, uncertainty, and introspection

As we saw in the previous section, a part of my explanation of subject obviation is an interplay between the meaning of the matrix predicate and the kind of information in the embedded clause. This idea is also present in the explanation in Ruwet (1984) and some recent semantic-pragmatic accounts of subject obviation. I already discussed the differences and similarities between my explanation of subject obviation and Ruwet's (section 2). In this section, I compare my explanation with two recent semantic-pragmatic accounts.

The first account I look at is a Kaufmann-Szabolcsi account (Kaufmann 2019; Szabolcsi 2021).[41] According to this account, the source of subject obviation is the same as the source of *directive obviation*. Directive obviation is an inability of first person exclusive forms to be subjects of regular root imperatives or subjunctives used as imperatives. For example, in Greek, *na*-subjunctives, as in (21)a, can be used as directive speech acts but not in the first person singular form, as in (21)b. Directive obviation can be described as a ban on having the *director* and the *instigator* refer to the same individual, the speaker in the case of first person singular.

[41] (See also Author, Year.)

(21) a. Na aniksis to parathiro. (Greek)
SBJV open-2SG the window
'Open the window!'
b. #Avrio na stilo ena e-mail stin Ana.
tomorrow SBJV send-1SG an e-mail at.the Anna
'Tomorrow, I should send an email to Anna.' (Oikonomou 2016, 73, 168)

Kaufmann (2019) builds her account of directive obviation on her earlier work, where directive speech acts are carried out with modalised sentences (i.e., *Open the window!* ≈ *You must/should open the window*) that come with a set of pragmatic presuppositions restricting the context. These restrictions derive the non-assertive character of imperatives and non-canonical directives. In particular, there are two general conditions on the use of imperatives: (a) the director of an imperative presents herself as uncertain about the course of events (Epistemic Uncertainty Condition) and (b) the director believes that the commanded proposition becomes true if the instigator takes it to be necessary (a combination of Decisive Modality and Director's Anticipation). If the director and the instigator are one and the same individual, the two conditions come into conflict. As a result, there can be no context that satisfies both of these conditions, and directives like (21)b are judged unacceptable.

Kaufmann (2019) makes a cursory remark that her account of directive obviation can extend to subject obviation in sentences with desire verbs like *vouloir* 'want'. Szabolcsi (2021) accepts this conjecture without giving a detailed analysis of how the account might work. I presume the idea is that the uncertainty presupposition of *want* (e.g., Heim 1992; von Fintel 1999) comes in conflict with the certainty attributed to the subject of the embedded clause. This certainty, I take it, is due to the type of information conveyed in the embedded clause.

The virtue of this account is that it proposes a general explanation of subject obviation in terms of semantic-pragmatic restrictions that can unify different obviative phenomena. The shortcomings of this account will, of course, depend on a concrete implementation of the general schema. However, already at the general level, it is possible to name a number of problematic points. First, if decisive modality remains an important ingredient of the analysis, it is unclear how the account can be extended to cases of subject obviation with epistemic attitudes. We saw subject obviation with epistemic attitudes in Italian earlier in the paper (see also (22) below). Second, since the uncertainty condition is attributed to the presuppositional meaning of an attitude predicate, factive predicates that show subject obviation cannot be explained by this account. I discuss subject obviation with factives in the next section. Finally, to account for cross-linguistic variation in subject obviation, the analysis would need to employ lexical (presuppositional) differences between those predicates that show subject obviation and those that do not. If two predicates are lexically similar, it may turn out to be difficult to locate such a difference. This point will be discussed in the next section.

Let me now look at another pragmatic-semantic account of subject obviation. Costantini (2016, 2023) studies subject obviation with epistemic attitudes in Italian; see (22) repeated from above. His explanation attributes subject obviation to a conflict between the interpretation of the matrix verb and the information conveyed by the embedded clause. In his view, (22)a is obviative because the verb is "a predicate implying an indirect access to a proposition" and the embedded clause conveys "a proposition accessible through introspection".[42]

[42](Costantini 2023, 37)

(22) a. #Penso che io parta domani. (Italian)
'I think that I leave-SBJV tomorrow'
b. Penso che io abbia fatto molti errori.
'I think thatI have-SBJV made many mistakes' (Costantini 2023, 29-30)

The merit of this account is that it broadens the empirical landscape of the phenomenon. Previous studies of subject obviation focused on desire predicates which led to erroneous generalisations. However, there are also serious shortcomings of this account. First, the account is rather descriptive. Terms like "indirect access" and "introspection" are not formally defined, and no formal details of the hypothesised interaction between these two features are provided. For this reason, it is hard to see how the account in Costantini (2016, 2023) can be extended beyond epistemic predicates in Italian. For example, "an indirect access to a proposition" might be a plausible restriction on an epistemic attitude, but it is unclear what it would mean in case of desire predicates. Second, cross-linguistic variation is also problematic for Costantini's account. The situation here is more pressing than for the Kaufmann-Szabolcsi account because subject obviation with desire predicates is frequent (at least in Indo-European languages), while subject obviation with epistemic predicates is rare. For example, in Romance languages subject obviation with 'believe' obtains in Italian but not in Spanish or French. We will see these data in the next section.

Let us sum up the problematic points for the two recent semantic-pragmatic accounts of subject obviation discussed above. Both accounts are restricted in empirical scope; the Kaufmann-Szabolcsi account is hard to extend beyond desiderative modality, while Costantini's account is restricted to epistemic attitudes. Both accounts have to address cross-linguistic variation in a superficial way, explaining for each predicate that shows subject obviation and each predicate that does not, why they behave the way they do. This is because the source of subject obviation in these accounts is linked to the meaning of the attitude. Moreover, the Kaufmann-Szabolcsi account needs a separate mechanism to explain subject obviation with factives, while Costantini's account needs formalisation. The core problem (I believe) is that these accounts attribute the source of subject obviation to some feature of linguistic competence; this leads us on a wild goose chase. I argue that we should instead echo Mercutio: "Nay, if thy wits run the wild goose chase, I have done"[43] and explore the possibility that subject obviation is due to a property of "general intelligence". In the next section, I show how my analysis in terms of "general intelligence" addresses the issues with factives and cross-linguistic variation.

7 Avoiding a wild goose chase

In this section, I will first discuss, in more detail, the problem that factives pose for the semantic-pragmatic accounts of subject obviation. Then, I will show how my explanation of subject obviation avoids this problem. Finally, I will make some remarks about cross-linguistic variation in subject obviation.

[43] W. Shakespeare *Romeo and Juliet*, Act 2 scene 4.

Factives

Subject obviation obtains with factive attitudes like 'regret' and 'know'. Consider the Italian examples in (23) and the Hungarian examples in (24). Notice that in Italian, *rammaricarsi* 'regret' selects the subjunctive mood, whereas in Hungarian, *sajnál* 'regret' selects the indicative. This, once again, illustrates the previously discussed point that subject obviation is not limited to subjunctive clauses. The Hungarian example in (24)b also shows the familiar amelioration effect found in other subject obviation cases.

(23) a. #Mi rammarico che io parta domani. (Italian)
'I regret that I leave-SBJV tomorrow'
b. #Non so se io parta domani.
'I don't know if I leave-SBJV tomorrow' (Costantini 2016, 127)

(24) a. #Sajnálom, hogy ugrándozok. (Hungarian)
'I regret that I jump-IND around'
b. Sajnálom, hogy untatlak.
'I regret that I bore-IND you' (Szabolcsi 2021, 11–12)

Discussing subject obviation with *sajnál* 'regret' in Hungarian, Szabolcsi (2021) notes that it cannot be easily explained in terms of a conflict between a certainty and uncertainty conditions on common ground.[44] This is because the uncertainty condition is not plausible in the case of factives since common ground (and hence, the attitude holder's beliefs, assuming common ground is defined as the shared beliefs of conversational participants) is assumed to entail the prejacent of a factive predicate. As an alternative, Szabolcsi outlines an explanation in terms of counterfactual reasoning, which (as she claims) approximates the certainty-uncertainty conflict. Hungarian 'regret' (she continues) can be paraphrased using 'wish', as in (25)a. This makes sentences with 'regret' similar to English sentences such as *I find it regrettable* or *I wish it weren't the case*. The paraphrase in (25)a entails the propositions in (25)b, which (according to Szabolcsi) together have "an appropriate whiff of contradiction to them".[45]

(25) a. I wish I weren't jumping around (paraphrase of (24)a)
b. If it were up to me whether I am jumping around, I would not be jumping around
AND I am jumping around (per factivity of 'regret'/'I wish I weren't')
AND It is up to me whether I am jumping around (per RESP[46])
(Szabolcsi 2021, 16)

The propositions in (25)b, indeed, seem to form an inconsistent set. But without specifying the interpretation of the counterfactual and the status of the factive and RESP-induced implications, it is hard to say what exactly is wrong when (25)b is used to communicate (24)a. Szabolcsi does not provide these details. In this section, I will discuss one way of formalising the reasoning in (25)b. This will clarify why (25)b appears inconsistent and provide grounds for evaluating the counterfactual strategy proposed in Szabolcsi (2021). As we will see, counterfactual reasoning is neither necessary nor sufficient for explaining subject obviation with factives. It may be true that

[44] Costantini (2016, 2023) makes only cursory remarks about factives.

[45] (Szabolcsi 2021, 16)

[46] RESP is a responsibility operator proposed by Farkas to account for control constructions and subjunctive obviation (Farkas 1988, 1992). It is often used in the literature on subject obviation to describe obviation with intentional actions.

some factives, like *sajnál* 'regret' in Hungarian, have a counterfactual interpretation, but this is not the reason they show subject obviation. Factives like *sajnál* 'regret', as I will argue below, show subject obviation for the same reason as non-factive attitudes.

To see that counterfactual reasoning is not necessary, we should note that it cannot explain subject obviation with all factives. In (23)b, we saw that *sapere* 'know' in Italian shows subject obviation, but counterfactual reasoning is implausible for epistemic factives. One might object that in (23)b the factive implication is not guaranteed because of the negation. In the so-called projective environments (under negation, in questions, with possibility adverbs, etc.), the presence of a factive implication depends on the context or question under discussion.[47] So, one could say that in (23)b no factive implication is projected and subject obviation is due to the same certainty-uncertainty conflict as in non-factive examples. However, subject obviation with *sapere* 'know' obtains in positive sentences as well; see (26). Negation in (23)b is used just to ensure that *sapere* 'know' selects the subjunctive mood, resulting in robust unacceptability. In positive sentences like (26), *sapere* 'know' takes the indicative and subject obviation holds only with a neutral intonation; a high pitch on *sapere* 'know' makes (26) acceptable (see Costantini 2023, fn. 5 for discussion).

(26) ?So che sto male. (Italian)
'I know I feel sick.' (Costantini 2023, fn. 5)

In addition to *sapere* 'know', other factives in Italian, such as *essere sorpresi* 'be surprised' and *avere saputo* 'come to know', show subject obviation; see (27). These factives are epistemic in nature and, as mentioned above, it is unclear how the counterfactual strategy in (25)b could be used to explain subject obviation with epistemic factives. It is true that the constructions in (27) have a dynamic flavour, in the sense that they convey that the attitude holder's beliefs or expectations have been revised to align with the actual situation described by the prejacent. One might attempt to use this dynamic component of 'be surprised' and 'come to know' to develop an explanation akin to the counterfactual strategy. But these explanations will not be fully parallel; moreover, non-dynamic epistemic factives like 'know' in (23)b and (26) will remain unaddressed.

(27) a. #Sono sorpreso che io parta domani.[48] (Italian)
'I'm surprised that I leave-SBJV tomorrow'
b. #Ho saputo che sto male.
'I've come to know that I am-IND feeling sick.'
c. #Ho saputo che sto leggendo il giornale.
'I've come to know that I am-IND reading the newspaper.' (Costantini 2023, 43)

Cross-linguistic variation in subject obviation is a complex issue which we will discuss later in this section. But some challenges are already evident in our discussion of factives. We saw that for some factives, like 'regret' in Hungarian, counterfactual reasoning may be used (according to Szabolcsi, at least), but for other factives, like epistemic factives in Italian, an explanation in terms of counterfactual reasoning is implausible and a different strategy must be found. I think it is clear that if the source of subject obviation is the same across languages, counterfactual reasoning cannot be essential to explaining it for factives.

To show that counterfactual reasoning may not be sufficient for explaining subject obviation even with 'regret' in Hungarian, I will first discuss a way to formalise the counterfactual strategy.

[47](E.g., Tonhauser et al. 2013; Simons et al. 2016; Roberts and Simons 2024.)
[48]Thanks to Francesco Costantini (p.c.) for this example.

This will allow us to see exactly where the conflict lies. Then, I will introduce an example that avoids this type of conflict but still shows subject obviation. This will demonstrate that counterfactual reasoning is not sufficient for explaining subject obviation.

Let us look again at the counterfactual strategy repeated in (28) below. Recall that Szabolcsi (2021) argues that 'regret' in Hungarian can be paraphrased as 'wish', which produces an obviative sentence like '#I wish I weren't jumping around'. This paraphrase entails the propositions in (28), forming a logically inconsistent set. This explains, arguably, the unnaturalness of obviative sentences.

(28) a. If it were up to me whether I am jumping around, I would not be jumping around
b. I am jumping around (factive implication)
c. It is up to me whether I am jumping around (RESP-induced implication)

We can capture the conditional proposition in (28)a using the account in Stalnaker (1968). According to this account, the sentence 'if ϕ, then ψ' is true in a world x iff ψ is true in the nearest world to x in which ϕ is true; see (29). The nearest world is determined by a selection function s that maps a proposition and a possible world into a possible world. The selection function (among other restrictions) has to choose the actual world α if the antecedent ϕ is true in α (that is, if $\alpha \in \phi$, then $s(\phi, \alpha) = \alpha$).

(29) a. 'if it were up to me whether I am jumping around, I would not be jumping around' is true in x iff
s(it is up to me whether I am jumping around, x) $\in \{\, y : \neg$ I am jumping around in $y \,\}$

For convenience, let us assign the RESP-induced implication 'it is up to me whether to be jumping around' to ϕ, and the factive implication 'I am jumping around' to ψ. We can gloss over the status of factivity and RESP-induced implications and say that for a 'regret'-sentence in Hungarian to be assertable, the context set C (which includes the actual world α) must entail ϕ and ψ. That is, $C \subseteq \phi$ and $C \subseteq \psi$. Since the actual world α is a member of C, both ϕ and ψ are true in α. Since ϕ is also the antecedent in the conditional proposition in (28)a, the selection function s is required to select the actual world as the nearest world where the consequent is true. Since the consequent is equivalent to $\neg\psi$, we arrive at a contradiction that both ψ (the factive implication) and $\neg\psi$ (the consequent of the conditional) are true in α. This reasoning is schematically shown in (30).

(30) a. $\alpha \in C$ (by the definition of a context set)
b. $C \subseteq \phi$ (by RESP-induced implication)
c. $C \subseteq \psi$ (by factivity)
d. $\alpha \in \phi$ (from a, b)
e. $s(\phi, \alpha) = \alpha$ (from d and the requirement on s)
f. $\alpha \in W - \psi$ (from e and the meaning of the conditional)
g. $\alpha \in \psi$ (from a, c)

The counterfactual strategy appears to successfully derive a contradiction potentially causing subject obviation with factives. However, the contradiction that emerges from the reasoning in (30) is a contradiction for the wrong reason. What the strategy aims at (I think) is a conflict between what the agent intended or wished to do (if it were up to them) and what the agent actually did. For this, we need a link between the RESP-induced implication and the factive implication. However, no such link is required to derive the contradiction in (30). According to (30), sentences like 'If it

were up to me whether I am eating cheese, I wouldn't be jumping around' are contradictory for the same reason as our target example in (28)a. Consider (31) in a context where it is up to the speaker whether to eat cheese, but they decide against it, opting to jump around instead. To see how the counter-intuitive contradiction in (31) can be derived, all we need is to substitute ϕ in (30) with 'it is up to me whether I am eating cheese'.

(31) a. If it were up to me whether I am eating cheese, I wouldn't be jumping around
b. It is up to me whether I am eating cheese (RESP-induced implication)
c. I am jumping around (factive implication)

Of course, in the actual 'regret'-sentence in (24)a, the RESP-induced implication and the factive implication are generated by the same lexical material; thus, they are guaranteed to be about the same action. My point here is not that counterfactual reasoning *plus* the link between the RESP-induced implication and the factive implication cannot explain subject obviation with factives. As we saw above, it can. Rather, my point is that counterfactual reasoning alone does not produce the intended explanation. The fact that the reasoning in (30) derives a contradiction also when there is no link between RESP-induced and factive implications, as in (31), shows that counterfactual reasoning is not enough for our explanation. We must also establish a link between RESP-induced and factive implications to obtain a contradiction for the right reason.

In my explanation of subject obviation, there is no connection between the presuppositional meaning of an attitude and obviation. The tension that is the source of subject obviation is created by the need to satisfy the principle of non-triviality and the requirement on a rigid attitude. Consider the English gloss of the factive example parallel to the one in section 5, where Alice says '#I regret I am having vertigo'. Alice's epistemic situation in this example can be represented as in Figure 7 (where Va = Alice is having vertigo, Vm = Mabel is having vertigo, H = Heads, T = Tails).

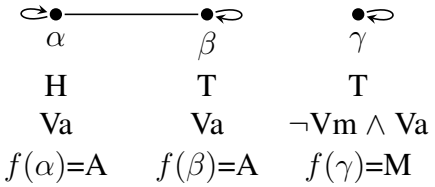

Figure 7: Illustration of an obviative sentence with a factive predicate such as Alice's utterance '#I regret I am having vertigo'. Obviation is due to the same tension as with non-factive attitudes: in the context $\{\alpha, \beta\}$, the principle of non-triviality is violated, while in the context $\{\alpha, \gamma\}$, a rigid attitude cannot be used. Factivity can be added without problem. (Va = Alice is having vertigo, Vm = Mabel is having vertigo, H = Heads, T = Tails)

Nothing in this setup prevents the speaker from presupposing that Alice is having vertigo in every world in the context set. Assuming that factivity can be treated as speaker presupposition (e.g., Stalnaker 2014), the setup depicted in Figure 7 can satisfy factivity and still give rise to subject obviation. Moreover, there is no problem in extending my explanation to epistemic factives like 'know' and 'be surprised', which, as we saw above, cannot be explained by the counterfactual strategy.

Cross-linguistic variation

Now, let me briefly comment on another problematic point for the semantic-pragmatic accounts discussed in the previous section: cross-linguistic variation in subject obviation. Consider the verb

'believe'. Subject obviation obtains with 'believe' in Italian but not in Spanish or French; see (32). Note, again, that the mood choice is irrelevant here.

(32) a. #Credo che io la convinca. (Italian)
'I believe I convince-SBJV her'
b. Creo que me marcho mañana. (Spanish)
'I believe that me leave-SBJV tomorrow'
c. Je crois que je suis malade. (French)
'I believe that I be-IND ill'

To account for this variation, the semantic-pragmatic accounts must argue that 'believe' in Italian is lexically different from 'believe' in Spanish or French. For example, that 'believe' in Spanish and French might not convey indirect access to knowledge or may lack an uncertainty presupposition. It remains uncertain whether such conjectures can be supported by evidence, but I think that finding a single property responsible for all cases of subject obviation is unlikely. If my analysis is on the right track, the question about which property of an attitude predicate causes subject obviation can be reformulated. Rather than asking which common property causes subject obviation, we will instead ask which property (potentially different in each case) disallows an interpretation with a non-rigid I-concept. I believe that the latter question is less demanding and more likely to yield an answer.

8 Conclusion

In this paper, I defended the view that subject obviation is a property of "general intelligence" rather than linguistic competence. I started with the explanation of subject obviation proposed by Ruwet (1984), who first articulated this view. In my proposal, the role of "general intelligence" was played by the principle of non-triviality that disallows ascription of propositions whose content does not exclude any possibility from a belief state (Stalnaker 1978, 1988). This principle was combined with Ruwet's intuition (also found in several recent accounts) that obviative sentences involve an illegitimately tight "self-to-self" relation between the two coreferential subjects. I proposed that this tight "self-to-self" relation can be captured as self-locating information and it is this kind of information that makes embedded propositions in obviative sentences non-ascribable. I used Stalnaker's framework to capture self-locating knowledge (Stalnaker 2008, 2014). I concluded by discussing how my proposal allows us to see in a fresh light some phenomena that are problematic for recent semantic-pragmatic accounts of subject obviation.

References

Anand, Pranav, and Valentine Hacquard. 2013. "Epistemics and Attitudes." *Semantics and Pragmatics* 6 (October). https://doi.org/10.3765/sp.6.8.

Anscombe, G. E. M. 1957. *Intention*. Cambridge, London: Harvard University Press.

Bouchard, Denis. 1983. "The Avoid Pronoun Principle and the Elsewhere Principle." In *Proceedings of NELS 13*, 29–36. Amherst: University of Massachusetts, GLSA.

Cappelen, Herman, and Josh Dever. 2013. *The Inessential Indexical: On the Philosophical*

Insignificance of Perspective and the First Person. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199686742.001.0001.

Castañeda, Hector-Neri. 1966. "'He': A Study in the Logic of Self-Consciousness." *Ratio* 8: 130–57.

Chierchia, Gennaro. 1989. "Anaphor and Attitudes de Se." In *Semantics and Contextual Expressions*, edited by van Benthem Bartsch and van Emde Boas, 1–31. Kluwer/Reidel.

Condoravdi, Cleo, and Sven Lauer. 2016. "Anankastic Conditionals Are Just Conditionals." *Semantics and Pragmatics*, November. https://doi.org/10.3765/sp.9.8.

Costantini, Francesco. 2006. "Subjunctive Obviation: An Interface Perspective." PhD thesis, Ca' Foscari University of Venice.

———. 2016. "Subject Obviation as a Semantic Failure: A Preliminary Account." *Annali Di Ca' Foscari. Serie Occidentale* 50: 109–31. https://doi.org/10.14277/2499-1562/AnnOc-50-16-5.

———. 2023. "On Some Epistemic Access Effects." In *Agency and Intentions in Language*, edited by Julie Goncharov and Hedde Zeijlstra. Vol. 6. Brill Research Perspectives in Linguistics. Leiden: Brill. https://brill.com/display/title/68330?contents=editorial-content.

Farkas, Donka F. 1988. "On Obligatory Control." *Linguistics and Philosophy* 11 (1): 27–58. https://doi.org/10.1007/bf00635756.

———. 1992. "On Obviation." In *Lexical Matters*, edited by Ivan A. Sag and Anna Szabolcsi, 85–110. Stanford: Center for the Study of Language and Information.

Fintel, Kai von. 1999. "NPI Licensing, Strawson Entailment, and Context Dependency." *Journal of Semantics* 16 (2): 97–148.

Godfrey-Smith, Peter. 2020. *Metazoa: Animal Minds and the Birth of Consciousness*. Harper Collins.

Heim, Irene. 1992. "Presupposition Projection and the Semantics of Attitude Verbs." *Journal of Semantics* 9 (3): 183–221. https://doi.org/10.1093/jos/9.3.183.

Higginbotham, James. 2003. "Remembering, Imagining, and the First Person." In *Epistemology of Language*, edited by A. Barber, 496–533. Oxford: Oxford University Press.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127–36.

Kaufmann, Magdalena. 2019. "Who Controls Who (or What)." *Semantics and Linguistic Theory* 29 (December): 636–64. https://doi.org/10.3765/salt.v29i0.4643.

Kempchinsky, Paula. 1986. "Romance Subjunctive Clauses and Logical Form." PhD thesis, UCLA.

———. 2009. "What Can the Subjunctive Disjoint Reference Effect Tell Us about the Subjunctive?" *Lingua* 119 (12): 1788–1810. https://doi.org/10.1016/j.lingua.2008.11.009.

Lewis, David. 1973. "Causation." *Journal of Philosophy* 70 (17): 556–67. https://doi.org/10.2307/2025310.

———. 1979. "Attitudes de Dicto and de Se." *The Philosophical Review* 88 (4): 513–43. https://doi.org/10.2307/2184843.

Morgan, Jerry. 1970. "On the Criterion of Identity for Noun Phrase Deletion." *Chicago Linguistic Society* 6: 380–89.

Ninan, Dilip. 2010. "De Se Attitudes: Ascription and Communication." *Philosophy Compass* 5 (7): 551–67. https://doi.org/10.1111/j.1747-9991.2010.00290.x.

Oikonomou, Despina. 2016. "Covert Modals in Root Contexts." PhD thesis, MIT.

Perry, John. 1977. "Frege on Demonstratives." *The Philosophical Review* 86 (4): 474–97. https://doi.org/10.2307/2184564.

———. 1979. "The Problem of the Essential Indexical." *Noûs* 13 (1): 3–21. https://doi.org/10.230

7/2214792.

———. 1999. *Knowledge, Possibility and Consciousness*. Cambridge, Mass.: MIT Press.

Picallo, Carmen. 1985. "Opaque Domains." PhD thesis, CUNY.

Roberts, Craige, and Mandy Simons. 2024. "Preconditions and Projection: Explaining Non-Anaphoric Presupposition." *Linguistics and Philosophy* 47 (4): 703–48. https://doi.org/10.1007/s10988-024-09413-9.

Ruwet, Nicolas. 1984. "Je Veux Partir/* Je Veux Que Je Parte. À Propos de La Distribution Des Complétives à Temps Fini Et Des Compléments à l'infinitif En Français." *Cahiers de Grammaire* 7: 74–138.

———. 1991. "*Je Veux Partir/Je Veux Que Je Parte*: On the Distribution of Finite Complements and Infinitival Complements in French." In *Syntax and Human Experience*, edited by John Goldsmith, 1–55. Chicago: University of Chicago Press.

Schlenker, Philippe. 2005. "The Lazy Frenchman's Approach to the Subjunctive." In *Romance Languages and Linguistic Theory*, 269–309.

Simons, Mandy, David Beaver, Craige Roberts, and Judith Tonhauser. 2016. "The Best Question: Explaining the Projection Behavior of Factives." *Discourse Processes* 54 (3): 187--206. https://doi.org/10.1080/0163853x.2016.1150660.

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*, edited by Nicholas Rescher, 98–112. Blackwell.

———. 1978. "Assertion." *Syntax and Semantics* 9: 315–32.

———. 1984. *Inquiry*. Cambridge: MIT Press.

———. 1988. "Belief Attribution and Context." In *Contents of Thought*, edited by Robert Grimm and Daniel Merrill, 140–56. Tucson: University of Arizona Press.

———. 1999. *Context and Content*. Oxford: Oxford University Press.

———. 2008. *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

———. 2014. *Context*. Oxford: Oxford University Press.

Stegovec, Adrian. 2019. "Perspectival Control and Obviation in Directive Clauses." *Natural Language Semantics* 27 (1): 47–94. https://doi.org/10.1007/s11050-019-09150-x.

Szabolcsi, Anna. 2010. "Infinitives Vs. Subjunctives: What Do We Learn from Obviation and from Exemptions from Obviation?"

———. 2021. "Obviation in Hungarian: What Is Its Scope, and Is It Due to Competition?" *Glossa: A Journal of General Linguistics* 6 (1): 1–28. https://doi.org/10.5334/gjgl.1421.

Tonhauser, Judith, David Beaver, Craige Roberts, and Mandy Simons. 2013. "Toward a Taxonomy of Projective Content." *Language* 89 (1): 66–109. https://doi.org/10.1353/lan.2013.0001.

Villalta, Elisabeth. 2008. "Mood and Gradability: An Investigation of the Subjunctive Mood in Spanish." *Linguistics and Philosophy* 31: 467–522.