Forthcoming in *Philosophy Compass*. Penultimate version.

De Se Attitudes: Ascription and Communication*

Dilip Ninan
Arché, St Andrews

December 1, 2009

Abstract: This paper concerns two points of intersection between *de se* attitudes and the study of natural language: attitude ascription and communication. I first survey some recent work on the semantics of *de se* attitude ascriptions, with particular attention to ascriptions that are true only if the subject of the ascription has the appropriate *de se* attitude. I then examine – and attempt to solve – some problems concerning the role of *de se* attitudes in linguistic communication.

1. Introduction

De se or 'self-locating' attitudes are thoughts one would characteristically express with a sentence containing the first-person pronoun *I* (*me, my*). They are thoughts about oneself when one thinks of oneself in the first-person way. A number of philosophers have argued that *de se* attitudes constitute a distinctive category of thought, insofar as they cannot be reduced to either *de dicto* or *de re* attitudes.[1] This paper concerns two points of intersection between *de se* attitudes and the study of natural language. The first is the semantics of attitude ascriptions: We talk about about first-person beliefs, desires, and other psychological attitudes, and so a semantic theory for our language will have to give an account of this sort of talk. The second point of intersection is communication: When I utter a sentence containing the first-person pronoun *I*, I communicate some information about myself to you. What is that information and how does it relate to the *de se* belief expressed by my utterance? A theory of linguistic communication should provide an answer to this question, an answer which is systematically related to our semantic account of sentences containing *I*.

*For helpful comments, thanks to Derek Ball, Chris Barker, Andy Egan, Alejandro Pérez Carballo, Julia Langkau, Michael Glanzberg, François Recanati, Daniel Rothschild, Paolo Santorio, James Shaw, Brett Sherman, Brian Weatherson, Elia Zardini, and seminar participants at Arché and NYU. Special thanks to Sarah Moss and Seth Yalcin for extended discussions of these topics.

[1]See, for example, Castañeda (1966, 1967), Perry (1977, 1979), and Lewis (1979).

2. *De se* ascription

Consider two situations:

> Situation 1
> John is contemplating his military exploits. Marvelling at his bravery in battle, he thinks to himself, *I am a hero*.

> Situation 2
> A thoroughly inebriated John is watching TV. He watches a man recounting his military exploits. Impressed, John thinks to himself, *That man is a hero*. The man John is watching is John himself, but because he's so intoxicated, John fails to realize this.

In Situation 1, John's belief that he is a hero is *de se*: he is thinking about himself in a first-person way. In Situation 2, his belief that he is a hero is *merely de re*: he is thinking about himself, but not in a first-person way. (I assume familiarity here with the notion of a *de re* attitude, and I also assume that *de se* attitudes are a kind of *de re* attitude.)

Now consider the belief ascription:

1. John believes that he is a hero.

It seems that (1) is true when evaluated with respect to both Situations 1 and 2. Since a *de se* attitude that one is *F* is a *de re* attitude that one is *F*, in both situations John believes *de re* that he is a hero. Kaplan (1989, §XX) concluded from this that sentences like (1) have *coarse* truth conditions: (1) is true iff John believes *de re* that he is a hero. Since this condition is satisfied in both of these scenarios, (1) is true in both.

The issue of *de se* ascription would not be of much interest if the coarse truth conditions theory were true for all natural language attitude ascriptions. For in that case, natural language would simply be insensitive to the difference between *de se* thoughts about oneself and thoughts about oneself that are merely *de re*. For the remainder of this section, I'm going to put aside the question of whether the coarse truth conditions theory is right for sentences like (1). I suspect that it isn't, but the issue is subtle.[2] Instead, I want to focus on two kinds of attitude ascription for which the coarse truth conditions view pretty clearly fails. In both cases, I will present the data against the coarse truth conditions theory, and then go on to sketch some possible semantic accounts of the data. The aim is not to be fully comprehensive, but rather to introduce the topic of *de se* ascription by looking at a few cases in some detail.

[2] Percus and Sauerland (2003) offer an argument – 'the argument from *only*' – which tells against the coarse truth conditions view of sentences like (1). They go on to argue that their data show that sentences like (1) are ambiguous between a *de se* and a mere *de re* reading. While I think their data do tell against the coarse truth conditions view, it's less clear that they establish the ambiguity view, since a contextualist treatment might be possible. See Anand (2006, §1.3) for discussion.

2.1. *PRO*-ascriptions

Consider another hypothetical situation:

> Situation 3
> John is again intoxicated and watching TV. He is watching the speeches of various candidates in the upcoming election. He watches one particularly engaging candidate, and comes to think that this candidate will win. The candidate is none other than John himself; but because he is so intoxicated he doesn't realize that he is the candidate in question. In fact, he is rather pessimistic about his own prospects and thinks to himself, *I'm not going to win the election*.

Now consider the following pair of sentences:

2. (a) John expects that he will win the election.
   (b) John expects to win the election.

The standard judgments about these sentences is that (2a) is true (or has a true reading) in this situation, but that (2b) is unambiguously false in this situation. In order for (2b) to be true, John would need to think to himself, *I am going to win the election*. In other words, (2b) is true iff John has a *de se* expectation to the effect that he will win. If that's right, then sentences like (2b) do not have coarse truth conditions: they are true only if the subject has the appropriate *de se* attitude.

Most of the recent semantic theories designed to predict the fact that (2b) is true iff the subject has the appropriate *de se* attitude rely on a particular account of *de se* content, namely Lewis's (1979) 'centered worlds' account. For this reason, we adopt Lewis's account here.[3] Lewis motivates his account as an improvement over the standard possible worlds accounts of attitude content, but I won't rehearse his arguments here; the interested reader should consult Lewis (1979) and Lewis (1983a). On the possible worlds account, an agent's belief state determines a set of possible worlds, the possible worlds compatible with what she believes; these are the agent's *doxastic alternatives*. A *proposition* is likewise a set of possible worlds. An agent x in world w believes a proposition p iff all the possible worlds compatible with what she believes in w are contained in p.

In order to model irreducibly *de se* attitudes, Lewis encourages us to re-think the notion of a doxastic alternative. Rather than taking doxastic alternatives to be possible worlds, Lewis suggests that we take them to be *centered worlds* instead. Centered worlds are usually taken to be triples consisting of a possible world, a time, and an individual who exists at the world and time in question. The time coordinate is needed to handle *de nunc* attitudes, or thoughts concerning what time it is. Although *de nunc* attitudes raise issues similar to the

[3] For objections to Lewis's approach, see Stalnaker (1981, 2008). Alternatives to Lewis's account can be found in Stalnaker's work, as well as in Perry (1977, 1979) and Kaplan (1989).

ones raised by *de se* attitudes, we will be setting them aside here. So it will be harmless to omit the time coordinate of centered worlds, and simply take them to be pairs consisting of a possible world w and an inhabitant x of w.

A centered world $\langle w', x' \rangle$ is compatible with what an agent x believes in a world w iff x thinks in w that she might be x' in w' (x's beliefs do not exclude the possibility that she is x' in w'). An agent x believes *de se* in w that she is F iff every $\langle w', x' \rangle$ compatible with what she believes in w is such that x' is F in w'. To believe a *centered proposition* (set of centered worlds) p is for every centered world compatible with what one believes to be contained in p.[4]

This account has the ability to distinguish between *de se* and non-*de se* contents. Following Egan (2006, 107), we can say that non-*de se* contents are *boring* centered propositions, where a centered proposition p is boring just in case for any world w, and inhabitants x, y of w, $\langle w, x \rangle \in p$ iff $\langle w, y \rangle \in p$. Since boring centered propositions do not distinguish between worldmates, they are essentially equivalent to possible worlds propositions. A *de se* content is an *interesting*, i.e. non-boring, centered proposition. The account captures the idea that *de se* thoughts are not reducible to non-*de se* ones, since it allows for pure *de se* ignorance, ignorance that cannot be resolved simply by learning more and more boring centered propositions.

Given that we've adopted Lewis's account of the *de se*, we want a semantic theory that has the consequence that (2b) is true iff the content of John's expectation is the set of centered worlds in which the center wins the election. To yield this prediction, we start with a claim about the syntactic structure of (2b). The claim is that the non-finite embedded clause in (2b) has a covert subject, a phonologically null pronoun called *subject control PRO*, which we denote PRO_s. So the real structure of (2b) is something like:

John expects PRO_s to win the election.

The observation that, in attitude contexts, PRO_s gives rise to obligatorily *de se* readings is due to Morgan (1970), and the connection to Lewis's account of the *de se* was first made in Chierchia (1989).

There are a few closely related approaches to the truth conditions of attitude ascriptions which contain PRO_s. All treat *John expects* as it occurs in (2b) as expressing universal quantification over the centered worlds compatible with what John expects, and all treat the complement clause PRO_s *to win the election* as expressing an interesting centered proposition. One particularly elegant way of doing this involves enriching the index ('circumstance of evaluation') of a Kaplan-style two-dimensional semantics to include an individual parameter, and then making the semantic value of PRO_s the individual coordinate of the index. This idea is pursued in Anand and Nevins (2004), von Fintel (2005), and Stephenson (2007, Ch. 4).

To see how this works in more detail, we adopt a Kaplan-style semantic theory according to which semantic values are given relative to a context c and

[4] We will move back and forth between representing centered propositions as sets of centered worlds and characteristic functions thereof.

an index i (see von Fintel and Heim (2004) and von Fintel (2005) for more details on the system being presupposed). Normally we would take both contexts and indices to be world-time-individual triples, but since we are ignoring the temporal aspects of attitudes, we will take contexts and indices to be world-individual pairs. For a context c, w_c is the world of the context, x_c the speaker; and for an index i, w_i is the world of the index, x_i the individual coordinate. The double brackets '$[\![\ ]\!]$' will be used to denote our interpretation function, a three-place function that takes an expression, a context, and an index to a semantic value.

The key to predicting that (2b) is false in Situation 3 is setting the semantic value of PRO_s to the individual coordinate of the index:

$$[\![PRO_s]\!]^{c,i} = x_i$$

(This should be read as saying that the semantic value of PRO_s at a context c and index i is the individual coordinate of i.) With that lexical entry for PRO_s, the intension of a clause like *PRO_s to win the election* will be an interesting centered proposition, i.e. a *de se* content:

$\lambda i.[\![PRO_s \textit{ to win the election}]\!]^{c,i}$
$= \lambda i.\ x_i$ wins the election in w_i.

Read "λi..." as saying "the function which takes an index i to the truth value 1 (truth) iff...". So the above says that the intension of *PRO_s to win the election* is a function which takes an index i to 1 just in case x_i wins the relevant election in w_i. In other words, the semantic value of that clause is the characteristic function of the set of centered worlds in which the center wins the election.

On this semantics, the meaning of an attitude verb is a relation between an individual and a centered proposition. An individual x stands in the belief relation to a centered proposition p just in case p is true at all the centered worlds compatible with what x believes (note that p is true at a centered world $\langle w, x\rangle$ iff $p(w, x) = 1$). So we give the meaning of *expects* as follows:

$[\![\textit{expects}]\!]^{c,i} = \lambda p_{\langle se,t\rangle}.\ \lambda x_e.$ every centered world $\langle w', x'\rangle$ compatible with what x expects in w_i is such that $p(w', x') = 1$.[5]

Putting these pieces together gives us the results that (2b) is true at a context c and index i iff all the centered worlds $\langle w', x'\rangle$ compatible with John expects in w_i are such that x' wins the election in w'. In other words, the sentence is true just in case John has a *de se* expectation that he will win the election. Since John *lacks* such an expectation in Situation 3, we correctly predict that the sentence is false when evaluated at that situation.[6]

Something of a second-person correlate of sentences like (2b) is discussed in Schlenker (1999, 80 - 81). First consider this scenario:

[5] Individuals are of type e, possible worlds of type s, and truth values of type t. So a function of type $\langle se, t\rangle$ is a function from world-individual pairs to truth values.

[6] Other approaches to this data are possible. More 'extensional' approaches are pursued in Chierchia (1989), Schlenker (2003), and von Stechow (2002, 2003). And for a dynamic take on the issues discussed in this section, see Maier (2006, Ch. 3).

Situation 4
John is hosting a party, and he's told by a friend that Mary is behaving boorishly. As he looks for Mary, he runs into a woman he believes is Mary's sister Sue. He says to her, *Mary has to leave—she's offending the other guests*. But, in fact, the woman he is speaking to is Mary, not Sue.

Now consider the following sentences:

3. (a) John told Mary that she had to leave.
   (b) John told Mary to leave.

The contrast between these is intended to parallel our earlier contrast between (2a) and (2b). In the present case, it seems that (3a) has a true reading when evaluated in Situation 4, whereas (3b) is unambiguously false when evaluated in that situation. In order for (3b) to be true, John would have to say to Mary, *You have to leave* or simply, *Please leave*. That is, the telling must be '*de te*' ('of you'), not merely *de re*. Roughly speaking, the content of an utterance will be *de te* if the sentence uttered contains *you* (or is an imperative like *Please leave*).

A common syntactic assumption about the structure of (3b) is that the subject of the embedded clause is *object control PRO* or PRO_o. So the real structure of (3b) is something like:

John told Mary PRO_o to leave

Unlike PRO_s, PRO_o is 'linked' to the object of the attitude verb, rather than to its subject.

A standard way of dealing with this data involves altering slightly our notion of an 'attitudinal alternative'. So far we've been taking attitudinal alternatives to be centered worlds. But to deal with the present data, we can expand our notion of an attitudinal alternative to include a second individual; instead of centered worlds, we now use *pair-centered worlds*, a possible world plus a *pair* of inhabitants of that world (or a triple consisting of a world and two of its inhabitants). The first center of a pair-centered world will continue to represent the speaker or attitude holder; the second center will represent the speaker's addressee. Let's call the first center the *speaker-center* and the second the *addressee-center*.

The idea is that if John says to Mary, *Please leave*, the content of John's command is $\{\langle w, x, y\rangle : y \text{ leaves in } w\}$. But if, as in Situation 4, John says to Mary, *Mary has to leave*, the content of John's command is different, perhaps $\{\langle w, x, y\rangle :$ Mary leaves in $w\}$, perhaps something else.[7] So we want our semantic theory

[7] Although the semantic content of John's utterance of *Mary has to leave* may be the deontically modalized claim that Mary has to leave, his speech act may also count as performing a command whose content is that Mary leaves. Think of a parent saying to her child, *You have to go to bed right now!* The semantic content of that utterance may be a deontically modalized claim, but the parent has also performed a command whose content is that the addressee goes to bed.

to predict that (3b) is true just in case all the pair-centered worlds compatible with what John told Mary to do are such that the addressee-center leaves.

One way of generating this prediction is to assume that *John told Mary* expresses universal quantification over the pair-centered worlds compatible with what John told Mary to do, and that the semantic value of the embedded clause is a pair-centered proposition. To achieve this, we can treat indices as triples consisting of a world and *two* individuals. For uniformity – and because it will be useful later – we also take contexts to be pair-centered worlds, the second center representing the addressee of the utterance context. We will normally denote the speaker-center with the variables 'x', 'x''', etc., and the addressee-center with the variables 'y', 'y''', etc.. The semantic value of PRO_o will be the addressee-center of the index, y_i:

$$[\![PRO_o]\!]^{c,\langle w_i,x_i,y_i\rangle} = y_i$$

So the semantic value of PRO_o *to leave* is:

$\lambda i.[\![PRO_o \textit{ to leave}]\!]^{c,i}$
$= \lambda i.\ y_i$ leaves in w_i.

If *John told Mary* expresses quantification over pair-centered worlds, then (3b) says that all the pair-centered worlds $\langle w, x, y\rangle$ compatible with what John told Mary to do are such that y leaves in w. Since John *didn't* express a *de te* command in Situation 4, the sentence is correctly predicted to be false.[8]

2.2. Monsters

Recent work on *de se* ascription has uncovered a second class of counterexamples to the coarse truth conditions view: ascriptions that appear to contain what Kaplan (1989) called *monsters*. A monster is an operator that shifts the context parameter. Just as the truth of a sentence that contains a modal operator (with widest scope) depends on the truth of the embedded sentence at other possible worlds, the truth of a sentence containing a monster (if such there be) would depend on the truth of the embedded sentence at other *contexts*.

The first arguments for the existence of monsters were due to Schlenker (1999, 2003). One of Schlenker's examples involves the Amharic translation of *says*, which appears to have the option of shifting the context parameter. Consider the following situation:

Situation 5
John says, *I am a hero*, and he doesn't say that anyone else is a hero.

[8]Daniel Rothschild pointed out to me that one can generate this prediction without adding an addressee coordinate to the index if one assumes that the semantic value of PRO_o at c and i is the addressee of the speaker-center of i. That raises the question of whether there is any other data that shows that we *need* to include an addressee-coordinate in the index, as the literature often seems to assume.

In English, in order for someone other than John to report John's assertion without actually quoting him, she would have to use a sentence like (4a) rather than (4b), which would be false in this situation:

4. (a) John said that he is a hero.
   (b) John said that I am a hero.

The data reported by Schlenker is that Amharic-(4b) (i.e. the Amharic translation of (4b)) has two readings. The boring reading is the one on which it means just what English-(4b) means; this reading is false in Situation 5. But the interesting reading is one on which it is *true* in Situation 5: in Amharic, *I* can have a shifted interpretation under *says*, a reading on which it fails to denote the speaker of the actual utterance context.

Since the shifted reading is true in Situation 5, we know that John's asserting *de se* that he is a hero is sufficient for the truth of the shifted reading. A further argument shows that this is necessary as well; a mere third-person *de re* assertion is not enough to ensure the truth of the shifted reading. Consider a context of utterance *c* in which someone other than John is speaking. Amharic-(4b) is ambiguous between two readings, but neither reading is true relative to *c* and Situation 6:

> Situation 6
> Drunken John is again watching TV. He watches a man recounting his military exploits. Impressed, John says, *That man is a hero*. The man John is watching is John himself, but because he's so intoxicated, John fails to realize this. John doesn't say of anyone else that he is a hero.

Amharic-(4b) has the boring, unshifted reading, on which it means that John said that the speaker of *c* is a hero. This is false in Situation 6, because we're assuming John is not the speaker of *c*, and John didn't say of anyone other than himself that he is a hero. Amharic-(4b) also has the interesting, shifted reading, but Amharic speakers judge Amharic-(4b) false in Situation 6. So the shifted reading doesn't have coarse truth conditions; its truth requires John's assertion to be *de se*.

Given these data, we need a semantic theory that will have the consequence that Amharic-(4b) is ambiguous between a reading on which Amharic-*I* picks out the speaker of the actual utterance context, and one on which it picks out the speaker-center of the pair-centered worlds compatible with what John said. *One* way to achieve that result would be to posit a monster that optionally occurs in the structure at the top of the embedded clause; I'll say more about how that option works in a moment.

But before we can conclude that Amharic contains monsters, we need to rule out rival hypotheses. The most obvious one is that Amharic-(4b) involves quotation, and so is not an instance of indirect discourse at all. The main argument against this hypothesis is that while material outside of a quotation does

not typically stand in certain syntactic/semantic relations to material within a quotation (e.g. NPI licensing), similar restrictions are not observed in Amharic attitude ascriptions. We won't go further into this issue here; for relevant discussion see Anand and Nevins (2004), Schlenker (Forthcoming), and especially Anand (2006, §2.3.2).

But even if we set aside the quotational approach, the data at hand does not obviously motivate a monstrous account. Suppose we adopt the enriched two-dimensional (or multi-dimensional) system we discussed earlier, in which both contexts and indices are pair-centered worlds. And suppose that Amharic-*I* is actually ambiguous between two items, a true indexical, I_c, and a shiftable element, I_i. The true indexical denotes the speaker of the utterance context, and the shiftable element gets its value from the speaker coordinate of the index (just like PRO_s):

$$[\![I_c]\!]^{c,i} = x_c$$
$$[\![I_i]\!]^{c,i} = x_i$$

On this account, Amharic-(4b) is ambiguous between the following two readings:

- John said that I_c am a hero.
- John said that I_i am a hero.

The first means just what its English counterpart means: it will be true at a context c just in case John said that the speaker of c is a hero. This gives us the boring, unshifted reading, the reading that is false in Situation 5. But the second will be true just in case the content of John's assertion is the pair-centered proposition in which the speaker-center is a hero, i.e. it will be true if John said, *I am a hero*. This is because the intension of *that* I_i *am a hero* is the pair-centered proposition in which the first center is a hero:

$$\lambda i.[\![\textit{that } I_i \textit{ am a hero}]\!]^{c,\langle w_i, x_i, y_i\rangle}$$
$$= \lambda i.\ x_i \text{ is a hero in } w_i$$

So on this disambiguation, Amharic-(4b) will be true if John asserts *de se* that he is hero (as in Situation 5), but false if his assertion is merely *de re* (as in Situation 6), which is the result we were hoping to predict.

Let's call this view the *pronominal ambiguity view*. On this proposal, no monster is needed to generate the shifty reading. So while Schlenker concluded on the basis of data like Amharic-(4b) that natural language contains monsters, it seems that these data can be handled by a suitably rich context-index theory.[9]

[9] In addition to this question about what sort of *data* would motivate a monstrous analysis, there is also a question about what sort of *theories* count as monstrous. For example, as von Stechow (2002, 2003) points out, the system in Schlenker (2003) does not actually contain monsters in our sense of the term, since no operator in that system ever manipulates the context parameter. See Schlenker (Forthcoming) for further discussion.

This raises a question: could we ever have evidence that a language contained monsters? Or will we always be able to explain away apparently monstrous behavior by appealing to pronominal ambiguity in this way? As it turns out, there is a way to distinguish between these two approaches. And for some constructions in natural language, the monstrous analysis looks more promising than the pronominal ambiguity view. The crucial data are presented in Anand and Nevins (2004), who present a number of arguments for a monstrous analysis of the data they consider; we present just one of their arguments here. Let's look at one of the languages they discuss, Zazaki, in which all indexicals (first-person, second-person, temporal, locative) can optionally shift under Zazaki-*says*. As they show, the pattern of indexical shift is constrained in a way that seems best explained by a monstrous analysis.

Consider what the pronominal ambiguity view would say about Zazaki ascriptions whose complement clause contains more than one occurrence of *I*. That is, consider the Zazaki version of a sentence like (5):

5. John said that I swindled my brother.

On the pronominal ambiguity account, Zazaki-*I* is ambiguous between a true indexical meaning, I_c, and a PRO_s-like meaning, I_i. Since there are two occurrences of *I* in (5) and two ways to disambiguate each one, there should be four possible readings of this sentence:

6. (a) John said that I_c swindled my_c brother.
   (b) John said that I_i swindled my_i brother.
   (c) John said that I_c swindled my_i brother.
   (d) John said that I_i swindled my_c brother.

It shouldn't be too hard to think about what each of these disambiguations would mean. The first is the boring reading on which the sentence just means what its English counterpart means. For example, if Mary is the speaker of context *c*, then (6a) would be true at *c* if John had said, *Mary swindled her brother*. In (6b), both pronouns get the shifted reading, so on that reading, the sentence would be true at *c* if John had said, *I swindled my brother*. (6c) and (6d) are 'mixed' readings: (6c) would be true at *c* if John had said, *Mary swindled my brother*; and (6c) would be true at *c* if John had said, *I swindled Mary's brother*.

Unfortunately for the pronominal ambiguity view, the two mixed readings, (6c) and (6d), are not possible. Zazaki speakers report that Zazaki-(5) would not be true in *c* if John had said, *Mary swindled my brother* nor if he had said, *I swindled Mary's brother*. Only (6a) and (6b) correspond to possible readings of the sentence; either the two occurrences of the indexical shift together or neither shifts. Anand and Nevins show that this empirical generalization – 'Shift Together' – holds for several other languages that contain shifty indexicals. They argue that this datum can be predicted if we assume that Zazaki contains a certain type of monstrous operator.

Their analysis has two components. First, we assume that Zazaki-*I* has a standard Kaplanian entry:

$[\![I]\!]^{c,i} = x_c$

Second, we assume that a Zazaki speech report like Zazaki-(5) is ambiguous between two structures:

7. (a) John said that I swindled my brother.
   (b) John said that OP I swindled my brother.

Given the Kaplanian entry for Zazaki-*I*, (7a) will mean just what its English counterpart means: it will be true at a context c just in case John said that x_c swindled x_c's brother. This is a result we want, since Zazaki-(5) has this boring reading.

(7b) corresponds to the interesting shifted reading, given the following semantics for the operator *OP*:

Where ϕ is a sentence, $[\![OP\ \phi]\!]^{c,i} = [\![\phi]\!]^{i,i}$ [10]

OP is a monster: it shifts the context parameter to the current index. This will generate the shifted reading since the intension of the embedded clause *OP I swindled my brother* will be the pair-centered proposition in which the speaker-center is swindled his brother, which is a *de se* content:

$\lambda i.[\![OP\ I\ swindled\ my\ brother]\!]^{c,i}$
$= \lambda i.[\![I\ swindled\ my\ brother]\!]^{i,i}$
$= \lambda i.\ x_i$ swindled x_i's brother in w_i

So (7b) will be true just in case all the pair-centered worlds compatible with what John said are such that the speaker-center swindled his brother. So the sentence will be true if, for example, John said, *I swindled my brother*. Since on the monstrous analysis, (7a) and (7b) are the only two structures for Zazaki-(5), the theory predicts all and only the right readings for the sentence.

3. *De se* communication

3.1. The problem

So far we've been examining sentences that *ascribe de se* attitudes to individuals; nothing we've said so far suggests that natural language sentences can ever be used to *communicate de se* information.[11] But it is plausible that some sentences can be used to convey such information. If I have a *de se* belief that my pants are on fire, and I say to you, *My pants are on fire*, it would seem that you would be in a

[10] The semantic value of *OP* at c and i is a function from Kaplanian characters to truth values; it maps a character p to truth just in case $p(i)(i) = 1$. So we must also add a new composition rule ('Monstrous Function Application') which allows an operator to combine with the character of its sister.

[11] And nothing we've said so far commits us to the claim that the asserted content of some stand-alone sentence is an interesting centered proposition. So far, interesting centered propositions have only served as the semantic values of *embedded* clauses.

position to know just what I know—namely, that my pants are on fire. Similarly, when I say to you, *Your pants are on fire*, it seems that, if you understood and accepted my utterance, you would gain a piece of *de se* information, the very piece of information I was attempting to convey to you—namely, that your pants are on fire.

On the face of it, then, *de se* information does play a role in communication. But this simple fact creates a problem for the centered worlds approach to the *de se*, a problem first pressed by Stalnaker (1981, 145 - 147). To see what the problem is, it will help to have a simple model of linguistic communication in place, so let me first introduce such a model, one due to Stalnaker (1978).

The basic idea of Stalnaker's account is that conversation proceeds against the background of a shared body of information, and that assertion essentially involves adding to that body of information. The central notion of Stalnaker's model is that of a *speaker presupposition*, a notion he glosses as follows:

> ...the presuppositions of a speaker are the propositions he takes for granted as part of the background of the conversation. A proposition is presupposed if the speaker is disposed to act as if he assumes or believes that the proposition is true, and as if he assumes or believes that his audience assumes or believes that it is true as well. (Stalnaker 1978, 84)

Stalnaker takes propositions to be sets of possible worlds. If we take the set of worlds w such that all of the propositions that a speaker presupposes are true at w, we get her *context set*, the set of possible worlds compatible with what she presupposes. Ideally, all the participants in a conversation will make the same presuppositions, in which case they all have the same context set, the *conversation's* context set. The conversation's context set represents the body of information shared by all the conversational participants. When speakers don't all make the same presuppositions, the conversation is *defective*.

Stalnaker also assumes that the content of an assertion is a possible worlds proposition. If a speaker assertively utters a sentence in conversation and her audience accepts her utterance, then the content of her utterance comes to be presupposed by all participants. In other words, the effect of asserting a proposition p is to remove certain worlds from the context set—namely all the worlds at which p is false. We'll say that a context set *verifies* a proposition p if p is true at every world in the context set.

The framework assumes that the contents of presuppositions, beliefs, and assertions are possible worlds propositions. But suppose I say to you, *My pants are on fire*. I thereby express my *de se* belief that my pants are on fire. Let p be the content of this belief. Then p is presumably what you come to learn when you accept my utterance, and what we both come to presuppose after you accept my utterance. But if Lewis is right, p is not a possible worlds proposition—my sentence expressed my irreducibly *de se* thought, and the content of a *de se* thought cannot, according to Lewis, be represented by a set of possible worlds. But then the content of my thought cannot be represented within Stalnaker's

framework, since the only contents that framework recognises are possible worlds propositions.

A natural idea is to try to reformulate Stalnaker's theory, replacing possible worlds with centered worlds throughout, and permitting the objects of assertion, belief, and presupposition to be interesting centered propositions, *de se* contents. But if we do this, we quickly run into the problem Stalnaker (1981) first raised for the centered worlds approach. To see the problem, suppose that, when, for example, John says to Mary, *My pants are on fire*, we take the content of his utterance to be the centered proposition in which the center's pants are on fire. If Mary accepts his utterance, it seems that we should update the context set by removing from it all the centered worlds in which center's pants are not on fire. That means that Mary now presupposes the centered proposition in which the center's pants are on fire. And for Mary to presuppose that is for her to presuppose that *her* pants are on fire. But this is obviously wrong: if Mary accepts and understands John's utterance, she will come to presuppose that his pants are on fire, not that hers are.

There is also a 'second-person' problem. If John says to Mary, *Your pants are on fire*, and Mary understands and accepts his utterance, then she will come to have a *de se* belief to the effect that her pants are on fire. If you need to convince yourself of this claim, consider the following case. John says to Mary, *Hey! Your pants are on fire!* Now suppose Mary doesn't realize John is speaking to her. What happens is that she sees someone in a mirror, and she thinks that John is talking to that person, and she comes to believe that that person's pants are on fire. As it turns out, that person *is* Mary; she just doesn't realize it. So if she accepts John's utterance, she will come to have a third-person *de re* belief about herself to the effect that her pants are on fire. But even though Mary forms this *de re* belief about herself, there is clearly an important sense in which she hasn't understood what John said, for she hasn't realized what it is that he was trying to get her to believe. John's communicative intention will not be realized unless Mary forms the appropriate *de se* belief.[12]

Why is this a problem? Well, if John says to Mary, *Your pants are on fire*, then if she understands and accepts his utterance, she should come to presuppose *de se* that her pants are on fire. On Lewis's theory, this means she will have to presuppose the centered proposition in which the center's pants are on fire. If the conversation is to remain non-defective, then the context set must now verify that centered proposition, which means that John must also come to presuppose that centered proposition. But for him to presuppose that is for him to presuppose that *his* pants are on fire. But that is absurd, for that is obviously not what he should come to presuppose after telling Mary that her pants are on fire.

[12]This may have something to do with the fact that there's something odd about second-person Moore sentences:

> It's raining but you don't believe it's raining.

I think the oddity here is due to the fact that in order for the addressee to accept an utterance of this sentence she'd have to accept to the corresponding first-person Moore sentence.

Some theorists have responded to these problems by suggesting that utterances containing the first- and second-person singular pronouns communicate *boring* centered propositions, i.e. centered propositions that do not vary in truth value between individuals (Egan 2007, 2009; Moss 2009). Since a boring centered proposition is essentially equivalent to a possible worlds proposition, this view effectively says that we can model the exchange of *de se* information using possible worlds propositions. As far as I can see, this position can only be maintained by abandoning one of the central aspects of Lewis's view—that *de se* information cannot be represented using possible worlds propositions. Perhaps this is an assumption we ought to revise if solving our two problems within Lewis's framework proves intractable. But we shall set this option aside for the moment in order to explore the prospects of solving these problems within Lewis's framework.[13]

3.2. Pair-centered worlds and conversational sequences

So far we've been looking at the problems that arise if we take the objects of assertion to be centered propositions. Do things look any better if we switch from centered propositions to *pair-centered* propositions? At first, it might seem like this move would bring us no closer to a solution to our two problems. For suppose that when John says to Mary, *My pants are on fire*, we take the content of his assertion to be $\{\langle w, x, y\rangle : x\text{'s pants are on fire in } w\}$. And suppose that Mary accepts John's utterance, so that the context set now verifies the pair-centered proposition he asserted (the context set is now a set of pair-centered worlds). This means that Mary now presupposes that pair-centered proposition. But for *Mary* to presuppose that pair-centered proposition is for her to presuppose that *her* pants are on fire. For when Mary presupposes a pair-centered proposition, the first center represents Mary. So we have essentially the same problem that we had earlier with centered propositions.

The problem is that who the first center represents is always a function of who the attitude holder is (the same goes for the second center). When John believes the pair-centered proposition in which the first center's pants are on fire, John's belief is true iff his pants are on fire. But when Mary believes the very same pair-centered proposition, her belief is true iff *her* pants are on fire. So who the first center represents is not stable across John and Mary: relative to John, the first center represents John; relative to Mary, the first center represents Mary. That means that when a pair-centered proposition 'travels' via assertion from John to Mary, the first center 'switches' from representing John to representing Mary. (All of this is true, *mutatis mutandis*, for the second center as well.) This is the source of the problem. We need to stabilize each center, so that when a pair-centered proposition p travels via assertion from John to Mary, who the first center of p represents doesn't change even after Mary 'receives' p.

One way to do this is to relativize the content of any assertion, presupposition, or belief made or held during a conversation to a *conversational sequence*,

[13] Other discussions of this issue can be found in Heim (2004) and Stalnaker (2008, Ch. 3). Yalcin (2007) and Torre (Forthcoming) present accounts similar in spirit to the one presented here.

or an ordered list of the conversational participants. What the context set represents will also be understood in conjunction with the conversational sequence. For a conversation with two participants, the conversational sequence will be a pair. So if John and Mary are the conversational participants, the conversational sequence will either be ⟨John, Mary⟩ or ⟨Mary, John⟩. It doesn't matter which of these we choose—either of them will work, we just have to pick one and stick with it. Let's use ⟨John, Mary⟩.

The content of an assertion (presupposition, etc.) made in a conversation will still be a pair-centered proposition. But instead of having a background stipulation that the first center always represents the speaker, the second her addressee, we now stipulate that the nth center of a pair-centered proposition p always represents the nth member of the conversational sequence. Since our conversational sequence is ⟨John, Mary⟩, the first center of a pair-centered proposition p always represents John, and the second always represents Mary, even when we are using p to characterize one of Mary's attitudes. The conversational sequence stabilizes the centers, so that who the first center represents doesn't switch when a pair-centered proposition travels from John to Mary.

Note that the notion of believing a pair-centered proposition relative to a conversational sequence can be explained in terms of believing a centered proposition. When we say that John believes the pair-centered proposition $\{\langle w, x, y\rangle : P(w, x, y)\}$ relative to the conversational sequence ⟨John, Mary⟩, we could understand this as saying that the following two things are true: (i) Mary is John's addressee, and (ii) John believes the following centered proposition:

$$\{\langle w, x\rangle : \text{there is a } y \text{ such that } y \text{ is } x\text{'s addressee in } w, \text{ and } P(w, x, y)\}.$$

Similarly, when we say that Mary believes the pair-centered proposition $\{\langle w, x, y\rangle : P(w, x, y)\}$ relative to the conversational sequence ⟨John, Mary⟩, we could understand this as saying that the following two things are true: (i) John is Mary's addressee, and (ii) Mary believes the following centered proposition:

$$\{\langle w, y\rangle : \text{there is an } x \text{ such that } x \text{ is } y\text{'s addressee in } w, \text{ and } P(w, x, y)\}.$$

More generally, we can say that an agent a in world w believes a pair-centered proposition $\{\langle w', x', y'\rangle : P(w', x', y')\}$ relative to the conversational sequence $\langle x, y\rangle$ iff either (i) $a = x$ and y is a's addressee in w, and a believes the centered proposition $\{\langle w', x'\rangle : \text{there is a } y' \text{ such that } y' \text{ is } x''\text{s addressee in } w', \text{ and } P(w', x', y')\}$, or (ii) $a = y$ and x is a's addressee in w, and a believes the centered proposition $\{\langle w', y'\rangle : \text{there is an } x' \text{ such that } x' \text{ is } y''\text{s addressee in } w', \text{ and } P(w', x', y')\}$. Note that it follows that an agent can only believe a pair-centered proposition relative to a conversational sequence if she is one of the members of the sequence.

Let's see how this approach helps with our two problems. Suppose that John says to Mary, *My pants are on fire*. The conversational sequence ⟨John, Mary⟩ tells us that the first center of any pair-centered proposition represents John. So, since John said that his own pants were on fire, the content of his utterance is the pair-centered proposition in which the first center's pants are

on fire. If Mary accepts John's utterance, we update the context set with this pair-centered proposition so that the context set contains only pair-centered worlds in which the first center's pants are on fire. That means that Mary now presupposes the pair-centered proposition in which the first center's pants are on fire.

But – and here's the crucial part – for Mary to presuppose this pair-centered proposition is *not* for her to presuppose that *her* pants are on fire, as it was on the old pair-centered theory. For on our new theory, when that pair-centered proposition travels from John to Mary, who the first center represents doesn't change—it continues to represent John even after Mary has received it. That's because who the first center represents is no longer a function of who the attitude holder is; who the first center represents is whoever comes first in the conversational sequence. Since John comes first in the conversational sequence, for Mary to presuppose the pair-centered proposition in which the first center's pants are on fire is for her to presuppose *de te* that John's pants are on fire. John's *de se* assertion results in Mary adopting the corresponding *de te* presupposition.

What about the second-person problem? On our new account, if the conversational sequence is ⟨John, Mary⟩, the content of John's utterance of *Your pants are on fire* is the pair-centered proposition in which the second center's pants are on fire. If Mary accepts his utterance, she will come to presuppose that pair-centered proposition. Given that Mary comes second in the conversational sequence, for her to presuppose that is for her to presuppose *de se* that her pants are on fire. And for John to presuppose it is for him to presuppose *de te* that her pants are on fire. So our new theory has no problem with *you*-utterances: John's *de te* assertion results in Mary's *de se* presupposition (and his *de te* presupposition), as desired.

Note also that the present account has no problem distinguishing between *de se* assertions and non-*de se* ones, and between *de te* assertions and non-*de te* ones. To see this, look at a sentence that is simultaneously *de se* and *de te*, such as *I love you*. Suppose John utters that sentence in the course of his conversation with Mary. Given the conversational sequence ⟨John, Mary⟩, the content of John's utterance is $\{\langle w, x, y\rangle : x \text{ loves } y \text{ in } w\}$. But if John had instead said, *John loves Mary*, the content of his utterance would have been something else—perhaps $\{\langle w, x, y\rangle : \text{John loves Mary in } w\}$, perhaps something else.

So far things look good, but our story is still incomplete. For I've been saying things like:

> Given the conversational sequence, the content of John's utterance of *My pants are on fire* is $\{\langle w, x, y\rangle : x\text{'s pants are on fire in } w\}$

without saying anything about how this kind of claim is connected to the semantic theories we were discussing in §2. What is the connection between the sentence John utters – *My pants are on fire* – and what we want the content of that utterance to be, given the conversational sequence we've chosen, ⟨John, Mary⟩? We want the content of that utterance to be the pair-centered proposition in which the first center's pants are on fire. As it happens, in this particular

case, that pair-centered proposition is the *Kaplan diagonal* of the sentence John uttered, *My pants are on fire*. The Kaplan diagonal of a sentence ϕ as the *set of contexts* at which ϕ is true:

> Kaplan diagonal of a sentence ϕ : $\{c : [\![\phi]\!]^{c,c} = 1\}$

Note that since we're taking contexts to be pair-centered worlds, a set of contexts is a pair-centered proposition, and so the Kaplan diagonal of a sentence is a pair-centered proposition.

Given the conversational sequence ⟨John, Mary⟩, the content we want to associate with John's utterance is just the Kaplan diagonal of the sentence he uttered. The sentence John uttered is *My pants are on fire*. We want the content of his utterance to be the pair-centered proposition in which the first center's pants are on fire. *My pants are on fire* is true at a context $c = \langle w_c, x_c, y_c \rangle$ just in case x_c's pants are on fire in w_c. So the *set* of contexts in which it is true (it's Kaplan diagonal) is the set of all $\langle w_c, x_c, y_c \rangle$ such that x_c's pants are on fire in w_c. But note that this is just another way of describing the pair-centered proposition in which the first center's pants are on fire. So the content of John's utterance is the Kaplan diagonal of the sentence he uttered.

But it's important to see that we get this result only because John happens to come first in the conversational sequence. For think about what happens if *Mary* were to say, *My pants are on fire*. Given that our conversational sequence is ⟨John, Mary⟩, we want the content of Mary's utterance to be the pair-centered proposition in which the *second* center's pants are on fire. But that pair-centered proposition is *not* the diagonal of the sentence she uttered; as before, the diagonal is the pair-centered proposition in which the *first* center's pants are on fire. For what the diagonal of a sentence is is not something that varies from context to context; the diagonal of *My pants are on fire* is the same irrespective of whether John or Mary is the speaker.

Of course, it's easy to retrieve the content we want to assign to Mary's utterance from the diagonal of the sentence she uttered: just swap the centers around. More precisely, the content we want to assign to her utterance is the *inverse* of the diagonal, in the following sense:

> For any pair-centered propositions p and q, q is the *inverse* of p iff for all pair-centered worlds $\langle w, x, y \rangle$: $\langle w, x, y \rangle \in q$ iff $\langle w, y, x \rangle \in p$.

If you have a pair-centered proposition p in which the first center is F and the second center is G, the inverse of p is the pair-centered proposition in which the first center is G and the second is F. Since the diagonal of the sentence Mary uttered is the pair-centered proposition in which the *first* center's pants are on fire, the inverse is the pair-centered proposition in which the *second* center's pants are on fire.

So, given the conversational sequence ⟨John, Mary⟩, when John utters a sentence, we want the content of his utterance to be the diagonal of the sentence he uttered; when Mary utters a sentence, we want the content of her utterance to be the *inverse* of the diagonal of the sentence she uttered. The generalization

is that if you come first in the conversational sequence, the content of your utterance is the diagonal, and if you come second, the content of your utterance is the inverse of the diagonal. This gives us the following definition of utterance content:

> Given a conversational sequence $\langle x, y \rangle$:
>
> - if x utters a sentence ϕ, the content of her utterance is the diagonal of ϕ;
> - if y utters a sentence ϕ, the content of her utterance is the inverse of the diagonal of ϕ.

The content of an utterance is what the participants all come to presuppose if the utterance is accepted; it is what we use to update the context set if the utterance is accepted.

There are at least two ways in which we might seek to extend this general approach. First, we would obviously want to cover conversations that have more than two participants. The generalization of a pair-centered world is a *sequenced world*, i.e. a possible world followed by an n-ary sequence of inhabitants of that world. To a model a conversation that has n participants, we would use a conversational sequence of length n and a set of sequenced worlds as the context set (cf. Stalnaker 2008, 73 - 74). Second, we might wonder whether the general strategy of adding more centers to the objects of thought and talk would be useful in dealing with the problems raised by talk about objects more generally (e.g. Frege's puzzle). I suspect that it would, but making good on this claim is no simple matter, and so I leave it as a topic for future research.[14]

References

Anand, Pranav. 2006. *De De Se*. Ph.D. thesis, MIT.

Anand, Pranav and Nevins, Andrew. 2004. "Shifty Operators in Changing Contexts." In Robert B. Young (ed.), *Proceedings from Semantics and Linguistic Theory XIV*. Cornell University: CLC Publications.

Castañeda, Hector-Neri. 1966. "'He': A Study in the Logic of Self-Consciousness." *Ratio* 8:130 – 157.

—. 1967. "Indicators and Quasi-indicators." *American Philosophical Quarterly* 4:85 – 100.

Chierchia, Gennaro. 1989. "Anaphora and Attitudes *De Se*." In R. Bartsch, J. van Benthem, and van Emde Boas (eds.), *Semantics and Contextual Expression*, 1 – 31. Dordrecht: Foris.

Egan, Andy. 2006. "Secondary Qualities and Self-Location." *Philosophy and Phenomenological Research* 72:97 – 119.

—. 2007. "Epistemic Modals, Relativism, and Assertion." *Philosophical Studies* 133:1 – 22.

—. 2009. "Billboards, Bombs, and Shotgun Weddings." *Synthése* 166:251 – 279.

von Fintel, Kai. 2005. "Lecture notes for Pragmatics in Linguistic Theory." Notes for class taught at the LSA Summer Institute, MIT.

[14] See Ninan (2008, Ch. 3) for a theory of attitude content that employs sequenced worlds propositions.

von Fintel, Kai and Heim, Irene. 2004. "Intensional Semantics Lecture Notes." Notes for class taught at MIT.

Heim, Irene. 2004. "Lectures Notes on Indexicality." Notes for class taught at MIT.

Kaplan, David. 1989. "Demonstratives." In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, 481 – 563. New York: Oxford University Press.

Lewis, David K. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review* 88:513 – 543. Reprinted in Lewis 1983b, 133 - 159. Page references to are to the 1983 reprint.

—. 1983a. "Individuation by Acquaintance and by Stipulation." *Philosophical Review* 373 – 402. Reprinted in Lewis 1999, 373-402. Page references to are to the 1999 reprint.

—. 1983b. *Philosophical Papers, Volume I*. New York: Oxford University Press.

—. 1999. *Papers in Epistemology and Metaphysics*. Cambridge: Cambridge University Press.

Maier, Emar. 2006. *Belief in Context: Towards a Unified Semantics of De Re and De Se Attitude Reports*. Ph.D. thesis, Radboud University Nijmegen.

Morgan, Jerry. 1970. "On the Criterion of NP Deletion." *CLS* 6:380 – 389.

Moss, Sarah. 2009. "Updating as Communication." Unpublished manuscript.

Ninan, Dilip. 2008. *Imagination, Content, and the Self*. Ph.D. thesis, MIT.

Percus, Orin and Sauerland, Uli. 2003. "On the LFs of Attitude Reports." In M. Weisgerber (ed.), *Proceedings of Sinn und Bedeutung 7*. Konstanz: Universität Konstanz.

Perry, John. 1977. "Frege on Demonstratives." *Philosophical Review* 86:474 – 497. Reprinted in Perry 1993, 3-32. Page references are to the 1993 reprint.

—. 1979. "The Problem of the Essential Indexical." *Noûs* 13:3 – 21. Reprinted in Perry 1993, 33-52. Page references are to the 1993 reprint.

—. 1993. *The Problem of the Essential Indexical and Other Essays*. New York: Oxford University Press.

Schlenker, Philippe. 1999. *Propositional Attitudes and Indexicality: A Cross-Categorial Approach*. Ph.D. thesis, MIT.

—. 2003. "A Plea for Monsters." *Linguistics and Philosophy* 26:29–120.

—. Forthcoming. "Indexicality and *de se* reports." In Klaus von Heusinger, Claudia Maienborn, and Paul Portner (eds.), *Handbook of Semantics*. Berlin: Mouton de Gruyter.

Stalnaker, Robert. 1978. "Assertion." In P. Cole (ed.), *Syntax and Semantics 9: Pragmatics*, 315–332. New York: Academic Press. Reprinted in Stalnaker 1999, 78 - 95. Page references to are to the 1999 reprint.

—. 1981. "Indexical Belief." *Synthése* 49:129–149. Reprinted in Stalnaker 1999, 130-149. Page references are to the 1999 reprint.

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

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

von Stechow, Arnim. 2002. "Binding by Verbs: Tense, Person and Mood Under Attitudes." Unpublished manuscript.

—. 2003. "Feature Deletion Under Semantic Binding: Tense, Person, and Mood Under Verbal Quantifiers." In Makoto Kadawaki and Shigeto Kawahara (eds.), *Proceedings of NELS 33*, 379 – 404. Charleston: BookSurge Publishing.

Stephenson, Tamina. 2007. *Towards a Theory of Subjective Meaning*. Ph.D. thesis, MIT.

Torre, Stephan. Forthcoming. "Centered Assertion." *Philosophical Studies* .

Yalcin, Seth. 2007. "Attitudes *De Se* in Context." Unpublished manuscript.