In: Emar Maier, Corien Bary & Janneke Huitink, eds. (2005) "Proceedings of SuB9"
www.ru.nl/ncs/sub9

DE SE REDUCTIONISM TAKES ON MONSTERS*

Emar Maier
Radboud University Nijmegen
e.maier@phil.ru.nl

Abstract

Chierchia (1989) and others have used the contrast between *George hopes that he will win* and *Georges hopes to win* in mistaken-self-identity scenarios, to argue for dedicated *de se* LFs. The argument, further strengthened by evidence of shiftable indexicals, appears applicable against any reductionist account that sees *de se* as merely a particular subtype of *de re*. My *Acquaintance Resolution* framework is an attempt at such a reduction, and this paper seeks to extend that theory with a logical principle of introspection for belief, to account for the data within a unified treatment of *de re* and *de se*.

1 Introduction

An arsonist has set fire to Tujiko's and Noriko's pants. Just before she feels her own pants burning, Tujiko catches a glimpse of her fiery pants in a mirror and, not recognizing herself, she points and yells: "Oh my god! That girl's pants are on fire!". Alarmed, Noriko looks down and notices she's on fire, screaming:"Help! My pants are on fire!"

This is your typical mistaken-self-identity scenario, modeled after an original from Kaplan (1989). It is meant to show the difference between *de se* and *de re* beliefs about oneself. On the one hand, the two utterances express a similar belief that is in both cases about the utterer herself, viz. that her pants are on fire. On the other hand the cognitive difference of the two expressed beliefs differs vastly and we may expect two different reactions: Noriko will panic and/or roll on the floor, whereas Tujiko will attempt to help the 'other' girl. On most accounts it follows that both girls' beliefs are *de re* about themselves, since they are both referring more or less directly to themselves with their utterances of *that girl* or *my*. The obvious difference is in the *way* they manage to refer to themselves. Noriko's belief is 'from a first person perspective', or *de se*, whereas Tujiko's use of the third person description shows a third person perspective on the same belief, which is still *de re* but not *de se* (i.e. mere *de re*).

The above tentatively suggests a treatment of *de se* attitudes as a subclass of *de re*. In 2 we will make this precise by giving a semantics of *de re* and *de se* belief. In 3 we will switch to belief reports, and see what the aforementioned semantics predicts as a semantics of reports. Section 4 presents a problem for the reductionist account of *de se* reports sketched above. In 5 I present my own reductionist attempt, which will be extended in 6 to cover the problematic constructions of 4.

*This paper and its companion (Maier to appear) provide different extensions to (Maier 2004), each addressing a different set of counterarguments to the *de se* reductionist proposal. I wish to thank Philippe Schlenker for raising and explaining the problem addressed in this paper.

2 The relational account of *de re* belief

To carry out the reduction of *de se* to *de re* we must be precise about what *de re* belief is. For this purpose we use a combination of Kaplanian (1969) acquaintance relations and Lewisian (1979) self-ascription. To start with the Kaplanian ingredient, the motivating example was Quine's (1956) Ortcutt scenario:

> There is a certain man in a brown hat whom Ralph has glimpsed several times under questionable circumstances on which we need not enter here; suffice it to say that Ralph suspects he is a spy. Also there is a gray-haired man, vaguely known to Ralph as rather a pillar of the community, whom Ralph is not aware of having seen except once at the beach. Now Ralph does not know it, but the men are one and the same [viz. Ortcutt]. (Quine 1956, 179)

From the first half we conclude that Ralph in fact believes *de re* of Ortcutt that he is a spy, but from the pillar-of-the-community bit it follows that Ralph believes *de re* of Ortcutt that he is not a spy. How to account for these two facts, without dismissing Ralph as logically insane?

Kaplan (1969) comes with a simple answer: *de re* belief is just *de dicto* belief with descriptive content provided by the way the believer is (perceptually) acquainted with the *res*. Applied to the Ortcutt example, the *res* is Ortcutt, and the relevant acquaintance relations have Ralph seeing someone in a brown hat, and seeing some guy with gray hair at the beach. In Kaplan's terminology: there are two *vivid names* of Ortcutt for Ralph: *the man in the brown hat* and *the gray-haired man at the beach*. The logical forms of the two seemingly contradictory beliefs then come out as *Ralph believes that the man in the brown hat is a spy* and *Ralph believes that the gray-haired man at the beach is not a spy*, wherein the belief relation may be explicated simply as relating an individual to a proposition, i.e. a set of worlds.[1]

To capture *de se* beliefs, however, we need more structure than just propositions as the objects of believe. This was the conclusion of Lewis's (1979) argumentation, based on examples where people are mistaken about who they are or are referring to. Take Kaplan's (1989) pants-on-fire scenario discussed above. What does Tujiko learn upon realizing that the girl she sees is herself? The difference is that now she can say 'My pants are on fire!', i.e. the *de re* belief has become *de se*, but has she learned a new *proposition*? No, says Lewis, proposition-wise nothing has changed; whether she refers to herself with *that girl* (pointing at the mirror), or with *I*, the expressed proposition constituting her belief is that Tujiko's pants are on fire. Lewis' solution is that belief is self-ascription of properties: first, Tujiko self-ascribes the property of seeing someone whose pants are on fire, then she realizes her mistake and comes to self-ascribe the property of having one's pants on fire. In possible worlds semantics, these properties are set-theoretically represented as $\{\langle a, w\rangle | a$ sees someone with fiery pants in $w\}$ and $\{\langle a, w\rangle | a$'s pants are on fire in $w\}$, respectively, and self-ascription is a new primitive notion replacing the modal attitude operator.

We now combine the above two theories into a unified analysis of *de re* and *de se*. First, make the Kaplanian definition of *de re* belief sensitive to properties:

[1] Although Kaplan himself advocates a sententialist view on propositional attitudes in the cited paper, we simply translate his theory to the more standard analysis of propositions as sets of possible worlds.

(1) x believes *de re* of y that it has P iff there is a two-place relation R s.t.
   (i) R is a sufficiently vivid acquaintance relation
   (ii) R holds between x and y (in the actual world)
   (iii) x self-ascribes the property of bearing R to something P

Applied to Tujiko and Noriko we get that both girls believe *de re* about themselves that they are on fire, because for each girl x there is an R that holds between x and x and satisfies the other two criteria, for Tujiko we can take R to be *seeing someone in the mirror*, for Noriko we can even just take the relation of equality, since unlike Tujiko, she believes to 'bear equality to someone whose pants are on fire' (iii).

Next, define *de se* as *de re* under the acquaintance relation of equality:

(2) x believes *de se* to be P iff x believes *de re* of x that he is P, with equality as the 2-place acquaintance relation R

Now, Noriko's belief is *de se*, but Tujiko's is merely *de re*. This reduction of *de se* to *de re* can be traced back to Lewis (1979, 156)[2] but Cresswell and von Stechow (1982) were the first to clearly separate belief and belief reports, and extend the above analysis of *de re* belief to a semantics of *belief reports*, with which the rest of the paper is concerned.

3 Belief reports

Belief reports are sentences typically used to convey that someone has some belief or other. As I said, the remainder of this paper provides a semantics for (a certain subclass of) belief reports, that is, a systematic[3] way of deriving logical forms (representations of truth-conditions in a logical language) from surface structures of the form *NP believes that NP VP*. The obvious starting point being that a sentence of that form is true iff the referent of the first NP believes *de re* of the referent of the second NP that that last has the property denoted by the VP, for example:

(3) $[\![\text{Ralph believes that Ortcutt is a spy}]\!]_w = 1$
   iff $[\![\text{Ralph}]\!]_w$ believes (in w) *de re* of $[\![\text{Ortcutt}]\!]_w$ that it is a spy
   iff there is an R s.t.
      (i) R is a sufficiently vivid acquaintance relation
      (ii) $R([\![\text{Ralph}]\!]_w, [\![\text{Ortcutt}]\!]_w)$
      (iii) $[\![\text{Ralph}]\!]_w$ self-ascribes the property of bearing R to a spy

We already saw that in Quine's example this is verified by taking R to be the relation of seeing someone in a brown hat. An analogous derivation, with $R(x, y)$ is *x sees y at the beach*, shows the truth of the report *Ralph believes that Ortcutt is not a spy*.

From now on we restrict attention to reports of beliefs about oneself. As Kaplan (1989) has pointed out, in the mistaken identity scenario where Tujiko doesn't recognize her own mirror image, the following reports, as uttered by an informed spectator, are both true:

[2] "[*de se* belief] is ascription of properties to oneself under the relation of identity. Certainly identity is a relation of acquaintance par excellence. So belief *de se* falls under belief *de re*." (Lewis 1979, p.156)

[3] Not necessarily *compositional* in the oldskool Amsterdam sense of the word…

(4) a. Noriko believes she's on fire
b. Tujiko believes she's on fire

Kaplan concluded that 'purely indexical distinctions', such as the difference between Tujiko's *de re* and Noriko's *de se* attitudes, can not be conveyed by reports in natural language: there are no *de se* reports, only *de se* attitudes. In the reductionist framework discussed above, this boils down to saying that for a report to be true there has to be *some* acquaintance relation, and natural language has no way of specifying on the surface, *which* acquaintance relation. This is exactly what is captured by the straightforward semantics exemplified in (3), which would indeed predict truth for both sentences in (4), in accordance with Kaplan's (1989) conjecture.

Such reductions of *de se* to *de re*, denying the existence of dedicated *de se* LFs for natural language reports have been proposed and defended by Boër and Lycan (1980), Cresswell and von Stechow (1982), von Stechow (1982), Reinhart (1990), and, reformulated in terms of Kaplan's two-dimensional character theory, by Kaplan (1989) and Zimmermann (1991). Lately, however, there has been a surge in counterarguments, one of which is the topic for the remainder of the paper.

4 Anti-reductionism

There are two groups of counterarguments against the general reductionist setup, one appears in work on monsters and *de se* reports (Chierchia 1989, Schlenker 2003, von Stechow 2002), and the other involves quantified belief reports, most notably embedding under 'only' (Percus and Sauerland 2003). This last I discuss elsewhere, for now we'll focus on the monsters.

First of all, Chierchia (1989) argues for separate *de se* and *de re* LFs on the basis of infinitive constructions like, in (Pseudo-)English:[4]

(5) a. Noriko believes to be on fire
b. #Tujiko believes to be on fire

Unlike with the corresponding 3rd person reports in (4), where the one about Tujiko was perhaps a bit forced (or even misleading) but still true, there is a real semantic contrast here: (5a) is fine, (5b) is plain false.

Given our earlier result that there can be no *de se* reports, the question arises how to account for these data? Chierchia's own account postulates an ambiguity: there are *de se* and *de re* LFs, and sentences as in (4) are ambiguous, whereas the infinitives in (5) correspond solely with a *de se* LF. On his account, a *de re* belief complement is of a sentential type, with a free variable bound by a *res* from the outside. Such complements become

[4]In English, this kind *believes to be* without object is rather rare, if not unacceptable, though Google comes up with e.g. "The author believes to be aware of related intellectual property rights [...]" (`www.ietf.org/ietf/IPR/infineon_ietf_ipr.pdf`). The similar *hopes to be* used by Schlenker (2003) and others is fine in English, and the contrast is the same, but the semantics of that attitude verb introduces some independent difficulties. Chierchia's original examples were in Italian where *crede di essere* is the standard way of ascribing *de se* beliefs, and I can confirm that in Dutch the analogous constructions *denkt te* and *meent te* + infinitive are ok, as witness the number of Google hits on `"denkt * te *"`

de se by fronting them with a λ at LF and binding the free variable, thereby type-shifting the propositional complement into a property. The variable binder has no discernible surface realization, so the ambiguity of (4) boils down to the existence or non-existence of this variable binder at LF. The argument then continues with recalling that infinitival complements of (5) have independently been argued to come with such a variable binder of their own. In Chierchia's theory this is further linked up with Chomsky's PRO, the invisible subject of such infinitival complements: the surface structures of (5) have an invisible subject NP, called PRO, which is nothing less than the surface realization of the λ-abstractor.

The next step is Schlenker (2003) who adds cross-linguistic data of 'monstrous' behavior of indexicals embedded in beliefs, that is, constructions like (6) where a first person indexical refers, not to the speaker, but to the subject of the reported belief. A literal gloss of a true Amharic belief report into English would for instance look like this (Pseudo-Amharic):

(6) Noriko believes that I_{Amh} am on fire

Somehow, it must be possible to interpret the embedded first person as a first person *with respect to the belief*. Moreover, on this reading, the truth-conditions are conjectured to be *de se* (Schlenker 2003, p.38), i.e. it patterns with the English PRO construction in (5). Note that I_{Amh} *can* be read *de re* (wide-scope), but then it patterns with English *I*, ascribing Noriko a belief about *me*, Emar. Schlenker eventually arrives at a theory which assigns 1st person features to *I*, I_{Amh} and PRO, and postulating a typology of 1st person pronouns: *I* must be evaluated with respect to the actual context, I_{Amh} can take the actual context or the belief context (in which case the belief is *de se*)[5], and PRO can only take the belief context.

The problem for reductionist theories of *de re*/*de se* can now be singled out to be the fact that they in effect scope the subject of the attitude complement out of the belief operator. This is can be seen in (3): the embedded NP *Ortcutt* is evaluated in the actual world w, and Ralph is supposed to be R-related to this Ortcutt in that world w too. In this way we can get the right result for English embedded *I*, but how about I_{Amh}? Surely, that is just as much a 1st person, just not always the *actual* 1st person. Once we accept that, we might as well follow Schlenker in analysing PRO as a first person too, since both are used to express first person thoughts. In the next sections I propose a way to incorporate these observations in a reductionist framework based on acquaintance resolution. The account differs from Schlenker's in that it uses a form of scoping, and from Von Stechow's (2001, 2002) in that the embedded subject's surface features are straightforwardly interpreted (no *feature deletion*).[6]

[5]There's an obvious correspondence between properties and sets of contexts as complements of a belief: self-ascribing the property of being on fire is the same as 'believing' the set of contexts whose agent (or *center*) is on fire.

[6]With respect to the third person, Schlenker also needs a morphological *agreement* mechanism, whereas the proposal developed below requires nothing of the sort.

5 Acquaintance Resolution

My proposal, *Acquaintance Resolution*, tries to give a formal semantic treatment of *de re* and *de se* belief reports by implementing an enhanced version of the relational attitude semantics exemplified in (3), in the framework of DRT with presuppositions. I assume some familiarity with basic DRT (Kamp and Reyle 1993), presupposition-as-anaphora (van der Sandt 1992), and possible worlds semantics. The aims are (i) to be weakly reductionistic in the sense that there be no syntactic ambiguity in the simple belief reports of (4)-(6) (contra (Chierchia 1989)), and (ii) pronouns are interpreted according to their surface features (contra e.g. (von Stechow 2002)). And of course we need to get the right truth-conditions for the reports in (4)-(6) in a systematic way, but that's just the definition of natural language semantics.

Just representing adequate *de re* and *de se* truth-conditions already necessitates some additions to standard DRT. I will here simplify a bit with respect to the formal semantics, focusing more on mapping sentences to representations. In the appendix or (Maier 2004) the interested reader will find the tedious semantic details, i.e. a 2-layered fragment of LDRT to account properly for direct reference and indexicality. For simplicity then, we now proceed with a 1-dimensional toy version, keeping in mind that certain uniqueness and rigidity facts require the machinery of the appendix.

First, add to the DRS language a predicate 'believe' with interpretation $\mathcal{B}el \in [\mathcal{D} \times \mathcal{W} \rightarrow \wp\mathcal{W}]$:

(7) $[\![\text{believe(x):}\varphi]\!]^f = \left\{ w \in \mathcal{W} \middle| [\![\varphi]\!]^f \supseteq \mathcal{B}el(f(\text{x}), w) \right\}$.

Now we can represent things like:

(8) a. Noriko believes that there's an arsonist in the house

b. $\left[\text{x} \;\middle|\; \text{Noriko(x), believe(x):} \left[\text{y} \;\middle|\; \text{arsonist(y), in\_the\_house(y)} \right] \right]$

c. $[\![\text{(8b)}]\!]^f_w = 1$ iff there is an individual a, called 'Noriko', in w, all of whose belief worlds $w' \in \mathcal{B}el(a, w)$ feature some arsonist who is in the house (at w')

Next, we need a predicate to represent the first person, i.e. to refer to the current speaker. Since we don't care about rigidity, we may simply take the predicate 'speaker', assuming implicitly that the worlds of evaluation are more like contexts: centered worlds with a unique speaker.[7] However, as the discussion of (Pseudo-)Amharic and PRO shows, a first person pronoun may also refer to the 'I' of a thought, the 'speaker' of an interior monologue. I propose a predicate 'center' to represent the first person in this somewhat generalized sense. With the 'center' predicate we can represent first person pronouns, (9a-b), and consequently *de se* ascriptions, (9c-d):

(9) a. $\left[\text{y} \;\middle|\; \text{center(y), on\_fire(y)} \right]$

b. $[\![\text{(9a)}]\!]^f_w = 1$ iff w has a center (speaker) who is on fire in w

[7] This is worked out more precisely in the 2D version, see appendix

c. $\left[\text{x} \;\middle|\; \text{Noriko(x), believe(x):} \left[\text{y} \;\middle|\; \text{center(y), on\_fire(y)} \right] \right]$

d. $[\![(9c)]\!]^f_w = 1$ iff a certain Noriko in w has a belief set in which each world has a center (experiencer) who is on fire

Given the simplifying assumptions discussed above, (9c) correctly represents the *de se* truth-conditions, but we have not said how to get at such a representation, given a sentence like (4a), (5a) or (6), all of which have the truthconditions of (9d). This process is often described as a two-stage procedure: first the sentence is parsed and compositionally transformed into a preliminary DRS, then (presupposition) resolution merges the preliminary DRS with the context (input) DRS and takes care of context-dependencies by binding or accommodating presuppositions, yielding the final (output) DRS representing the new context. My aim is to give an analysis of belief reports that assigns them all a single uniform preliminary *de re* DRS and in that sense unifying *de re* and *de se* reports. Note that my analysis is thus only weakly reductionistic because although the preliminary sentence representations of say (4a) and (4b) are uniform, after resolution the final representations differ, which is as it should be given the differing truth-conditions for *de re* and *de se* (readings of) reports.

To sketch the workings of acquaintance resolution, consider the 3rd person reports about Noriko (4a), and Tujiko (4b), in the mistaken identity context. In our dynamic framework we must first represent this input context, in which it is common ground (among the reporter and her audience, that is, Tujiko of course is clueless) that there are two girls, called Tujiko and Noriko, the first of whom is looking a mirror but not recognizing herself. This is represented as:

(10) $\left[\text{x y} \;\middle|\; \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \right]$

Now, the preliminary DRS of (4a) is:

(11) $\left[\begin{array}{l} \partial\left[\text{z} \;\middle|\; \text{Noriko(z)} \right] \\ \text{R(z,w)}\doteq ? \\ \text{believe(z):} \left[\text{u v} \;\middle|\; \begin{array}{l} \text{center(u), R(u,v), on\_fire(v)} \\ \partial\left[\text{w} \;\middle|\; \text{fem.3.sg.(w)} \right] \end{array} \right] \end{array} \right]$

This represents a sort of LF based on the relational analysis of *de re* sketched in section 2. The proper name *Noriko* and the pronoun *she* have triggered presuppositions, denoted by the ∂DRS, but there is also another kind of underspecification in (11), viz. R, a 2nd order free variable (ranging over 2-place relations), which is supposed to hold of z and w in the main DRS (corresponding to the real world). This R further serves as the descriptive content under which Noriko has the *de re* belief, as represented in the complement DRS which says 'there is a v that the belief center is R-acquainted with, and that v is on fire' in accordance with the *de re* reduction of (1), p.3.

After merging (11) and (10), we resolve the regular presuppositions, binding z and w to x (Noriko), and get:

(12)
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{R(x,x)}\doteq? \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), R(u,v), on\_fire(v)} \end{array}\right] \end{array}\right]$$

The resolution algorithm must perform a 'second order binding' to determine R, given that must be a two-place relation that holds in the context between x and x. This 2nd order binding is done by means of 2nd order matching, a special case of higher order unification (Dalrymple, Shieber and Pereira 1991): we look for a substitution for R that verifies the equation R(x,x)$\doteq\ldots$, the $\doteq$ representing $\alpha\beta\eta$-interconvertability of lambda terms, and the dots are to be replaced by a contextually salient relation relating x to x. By default we take x=x, which is not explicitly written in the context DRS, but can be thought of as always implicitly there, since it adds nothing to the truth-conditions. This gets us (13a). Then there are 4 possible unifying substitutions, of which (13b) is the one we want, the non-trivial one that resolves R to the relation of equality. Applying it to the whole gives (13c), which is equivalent to (13d-e):

(13) a.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{R(x,x)}\doteq\text{x=x} \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), R(u,v), on\_fire(v)} \end{array}\right] \end{array}\right]$$

b. $\text{R}\mapsto\lambda\text{s}\lambda\text{t.s=t}$

c.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & (\lambda\text{s}\lambda\text{t.s=t})\text{(x,x)}\doteq\text{x=x} \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), }(\lambda\text{s}\lambda\text{t.s=t})\text{(u,v), on\_fire(v)} \end{array}\right] \end{array}\right]$$

d.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{x=x}\doteq\text{x=x} \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), u=v, on\_fire(v)} \end{array}\right] \end{array}\right]$$

e.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u} & \text{center(u), on\_fire(u)} \end{array}\right] \end{array}\right]$$

We have succeeded in assigning a *de se* output DRS, equivalent to our earlier (9c), to an underspecified input. Now, a *de se* output for Tujiko (4b) would be false, contradicting our judgments, so let's see what happens if we add the same preliminary structure (11), except for the proper name, to the same context (10). After merging and resolving presuppositions, we're at (14a). If now we were to choose the default resolution, y=y for the question mark position and to consequently bind R to equality, we'd get *de se* which the context falsifies. But we can choose a different route, since now there is a salient contextual relation between y and herself: the looking in the mirror, the derivation of the *de re* reading we get from that is shown in (14). One of the main selling points of this kind of analysis is that we can view the deviation from the default equality acquaintance, and the associated pragmatic backtracking described above, as an explanation of the awkwardness many people feel with (4b)'s way of reporting the situation.

(14) a.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{R(y,y)}\doteq? \\ & \text{believe(y):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), R(u,v), on\_fire(v)} \end{array}\right] \end{array}\right]$$

b. $\text{R}\mapsto\lambda\text{s}\lambda\text{t.see\_in\_mirror(s,t)}$

c.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{believe(y):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), see\_in\_mirror(u,v), on\_fire(v)} \end{array}\right] \end{array}\right]$$

d. $[\![\text{(14c)}]\!]^f_w = 1$ iff in w . . . [context] . . . and all of Noriko's belief worlds have a center who sees someone in a mirror being on fire

Note in conclusion that the third person feature of the syntactically embedded *she* is straightforwardly interpreted as a semantic condition in the presupposition. This means that in the resolution *she*'s presupposition floats up to the main DRS, reminiscent of the 'wide-scope property' of (1) that was shown to cause trouble with PRO and shifted I_{Amh}. It remains to be seen if we can do better than the classical reductionist account.

6 Embedded first person and unambiguous *de se*

Now let's see what happens if we apply our analysis to the unambiguously *de se* (5) and the Pseudo-Amharic (6). But first, consider an English first person report. Picture a different scene, featuring me and my friend Noriko, me uttering (15).

(15) Noriko believes I'm on fire

The preliminary DRS for (15) is the same as (11) except for the pronoun's features which are now *1.sg* instead of *3.sg*. The context has two individuals, of which I am the speaker (center), so if we merge context and preliminary structure and resolve the proper name presupposition to its obvious antecedent, we're at:

(16)
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Emar(y), friend(x,y), center(y)} \\ & \text{R(x,w)}\doteq? \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), R(u,v), on\_fire(v)} \\ & \partial\left[\begin{array}{l|l} \text{w} & \text{1.sg(w)} \end{array}\right] \end{array}\right] \end{array}\right]$$

The remaining presupposition w can float up to top-level and be resolved to y, since obviously it's centers that can bind first person presupposition. Resolution then proceeds as follows:

(17) a.
$$\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Emar(y), friend(x,y), center(y)} \\ & \text{R(x,y)}\doteq\text{(Emar(y), friend(x,y))} \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), R(u,v), on\_fire(v)} \end{array}\right] \end{array}\right]$$

b. $\text{R}\mapsto\lambda\text{s}\lambda\text{t.(Emar(t), friend(s,t))}$

c. $\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Emar(y), friend(x,y), center(y)} \\ & \text{believe(x):}\left[\text{u v} \mid \text{center(u), Emar(v), friend(u,v), on\_fire(v)}\right] \end{array}\right]$

That is, Noriko believes something of the form "My friend Emar is on fire": the reading we want. So much for the wide-scope resolution of the *1.sg* presupposition, that gave the desired result: a real *de re* reading of *I*. But why can't we bind the presupposition of (16) to the local center? Well, in principle we *can*, but for English we'd get pathological readings, so we should really stipulate that English *I* always takes widest scope, a reformulation of (a corollary of) Kaplan's (1989) *Principle 2* which states that indexicals are directly referential.

However, as Schlenker (1999, 2003) shows, I_{Amh} behaves rather differently. As it happens, we can characterize this difference exactly by giving up the wide-scope stipulation for I_{Amh}. To see this, let us see what happens in our old context (11) with the Pseudo-Amharic (6), whose preliminary DRS is the same as for its English counterpart (15), so after merge we get:

(18) $\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{R(x,w)}\doteq\text{?} \\ & \text{believe(x):}\left[\begin{array}{l|l} \text{u v} & \text{center(u), R(u,v), on\_fire(v)} \\ & \partial\left[\text{w} \mid \text{1.sg(w)}\right] \end{array}\right] \end{array}\right]$

Schlenker suggests that the English-type wide-scope resolution is possible, which we account for by adding a representation of me as speaker to the context and resolving w to it, proceeding as sketched in (17a). Now, we consider the alternative, narrow scope resolution w↦u:

(19) $\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{R(x,u)}\doteq\text{?} \\ & \text{believe(x):}\left[\text{u v} \mid \text{center(u), R(u,v), on\_fire(v)}\right] \end{array}\right]$

What can R(x,u) be bound to? At first sight this seems strange since u has become unbound, but that need not be a problem since the main DRS does not *claim* that R(x,u) is the case, but rather asks for a part of a DRS with some conditions that involve the (free) variables x and u. In the current DRS there is a salient relation between x (Noriko) and u (Noriko's belief-self): *being the person you believe to be*, in fact this is explicitly present as the smallest subpart of the DRS containing both x and u:

(20) a. $\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{R(x,u)}\doteq\text{believe(x):}\left[\ \mid \text{center(u)}\right] \\ & \text{believe(x):}\left[\text{u v} \mid \text{center(u), R(u,v), on\_fire(v)}\right] \end{array}\right]$

b. $\text{R}\mapsto\lambda\text{s}\lambda\text{t.believe(s):}\left[\ \mid \text{center(t)}\right]$

c. $\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{believe(x):}\left[\text{u v} \mid \text{center(u), believe(u):}\left[\ \mid \text{center(v)}\right], \text{on\_fire(v)}\right] \end{array}\right]$

And then we're stuck, since what we want is *de se* truth-conditions, as represented by (13e) (≈(9c)); (20c) seems to attribute to Noriko a belief about a belief, rather than merely a belief about the self being on fire. Is there a way to deduce the *de se* belief from (20c)?

What's missing is some kind of introspection principle: if you believe to believe φ, you believe φ. Such principles have been studied in doxastic modal logic, i.e. modal systems with an operator $\Box$ interpretable as 'I believe that'. For example, in a well-known standard system for belief, **KD45**, as in **S5**, we have both *positive introspection* ($\Box\varphi \rightarrow \Box\Box\varphi$) and *negative introspection* ($\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$), semantically corresponding to the frame properties of transitivity and Euclidicity. Without first going into modal logic proofs, note that our semantics differs from these classical logics in that we make heavy use of the fact that our beliefs have centers, which makes the belief objects more like self-ascribed context sets, or equivalently Lewisian properties, than classical propositions. I will therefore posit a generalized introspection principle for centered belief, and since our representations are kind of 'heavy', I will also give a neat semantic formulation and show how it helps us get what we want.

First, syntactically, what we want is to reduce the double belief embedding to a single one. Classically that would be $\Box\Box\varphi \rightarrow \Box\varphi$ ("If I believe that I believe something, then I believe it"), which is indeed a theorem of **KD45** provable from the axioms of *consistency* ($\neg\Box(\varphi \wedge \neg\varphi)$) and negative introspection. The analogon of this theorem for centered belief would state that if x has a belief in which the center has a belief, then that second belief is actually belief of x's. As a sort of axiom it would look roughly like (21), where $U(\varphi)$ denotes the set of discourse referents, the *universe*, of the main DRS of φ, and $Con(\varphi)$ the set of conditions in φ.

(21) $\left[\ldots \mid \ldots\text{believe(x):}\left[\text{u} \ldots \mid \text{center(u), believe(u):}\varphi, \ldots\right] \ldots\right]$

$\Rightarrow \left[\ldots \mid \ldots\text{believe(x):}\left[\text{u } U(\varphi)\ldots \mid \text{center(u), } Con(\varphi), \ldots\right] \ldots\right]$

If we accept this, on grounds of its roots in classical **KD45** or its own intuitive appeal as a principle of a logic of belief, we see that (20c) is now equivalent to:

(22) $\left[\begin{array}{l|l} \text{x y} & \text{Noriko(x), Tujiko(y), see\_in\_mirror(y,y)} \\ & \text{believe(x):}\left[\text{u v} \mid \text{center(u), center(v), on\_fire(v)}\right] \end{array}\right]$

One additional stipulation is needed to arrive at the *de se* truth-conditions, and it's one we have more or less assumed all along: there can be but one center per belief alternative. In other words, you cannot believe yourself to be two people at once.[8] Syntactically:

[8] A very uncontroversial assumption, probably not even falsified by people with severe multiple personality syndrome.

(23) $\left[\ldots \;\middle|\; \ldots \text{believe(x):} \left[\text{u v}\ldots \;\middle|\; \text{center(u), center(v), } \ldots \right] \ldots \right]$

$\Rightarrow \left[\ldots \;\middle|\; \ldots \text{believe(x):} \left[\text{u v}\ldots \;\middle|\; \text{center(u), center(v), u=v, } \ldots \right] \ldots \right]$

An application of axiom (23) and consequent unification of u and v finishes the proof: Acquaintance resolution with axioms (21) and (23) predicts two readings for the Pseudo-Amharic (6), one wide-scope, same as for the English I in (15), viz. (17c); the other *de se*, i.e. same as the *de se* reading that was preferred for the English third person report (4a), viz. (13e). The same Amharic sentence now with *Tujiko* as subject is therefore predicted to be false in our scenario, because Tujiko doesn't recognize herself and so does not belief that the belief-center is on fire. Our prediction with respect to I_{Amh} are thus completely in line with Schlenker's discussion of the data (Schlenker 2003, p.38,74-76).

How about Chierchia's (1989) PRO and infinitival reports like the (Pseudo-)English ones in (5)? We simply assume a PRO with 1.sg features, like *I* and I_{Amh}, but adding the stipulation that this type of first person must take narrowest scope. We then predict only the above derivation of *de se* (using our new axioms), which is as it should be. Note that we have not got rid of Schlenker's stipulative typology of indexicals (Schlenker 2003, p.38,74-76), we merely replaced it with a reformulation more appropriate to our representational framework, i.e. in terms of scope: PRO must take narrow scope, *I* must take wide scope, and I_{Amh} can take either.

We have added two axioms, whose working is clear, but whose formulation is a bit hairy. A truly semantic formulation may be cleaner and more insightful or even easier to swallow, so that's why I offer (24):

(24) a. Every belief-alternative w has exactly one center
b. If $w' \in \mathcal{B}el(a, w)$ and b is the center of w', then $\mathcal{B}el(a, w) = \mathcal{B}el(b, w')$

In other words, well, (24a), the replacement for (23), speaks for itself, and (24b) says that the person you believe yourself to be, has, in the world you believe to inhabit, the same beliefs as you. This last is a bit stronger than (21) (it verifies an axiom of positive introspection as well), but in any case it will not be too hard to see that in all models in which (24) holds, (20c) is equivalent to (9c), i.e. we can derive *de se* readings for reports with embedded first persons.

Conclusion

In this paper I argued for a reductionist account of *de se* reports, based on the relational analysis of *de re* belief, according to which *de se* belief is *de re* belief about oneself, under the acquaintance relation of equality (or under the description *the person I am*, if you will). Reductionist accounts of belief *reports* typically predict that the type of acquaintance relation is not conveyed by linguistic means, a belief report is *de re* if the subject has a *de re* belief under *some* acquaintance relation. This prediction is borne out in third person reports of the form *x_i believes that she$_i$…* which are true if the belief in question is *de se* or merely *de re* (someone referring to herself without realizing it). However, data involving PRO and shiftable indexicals have cast doubt on the reductionist

endeavor, since it seems that some reports are not in the same way underspecified for (or *ambiguous between*, depending on your framework) *de se* and *de re*, but exclusively *de se*(on the co-referential reading).

My own framework is reductionist in the sense that it assigns a uniform preliminary structure to all reports of the form *NP believes that NP VP*, in which definite NPs are all interpreted as presuppositions with their content drawn straight from the surface features, and in which the acquaintance relation is left underspecified. A mechanism is provided by which the presuppositions and acquaintance relation are resolved in context, so it's really pragmatics that disambiguates between *de re* and *de se*, not syntax. As a bonus we get a pragmatic explanation for the fact that non-linguist/philosophers often find it hard to accept a co-referential third person pronoun report in a mistaken self-identity scenario.

The basic setup of Acquaintance Resolution is then extended with a logical principle of introspection to overcome difficulties encountered with a narrow-scope resolution of an embedded first person, something that we wouldn't need for 'well-behaved' indexicals like English *I*, but is the key to our treatment of shifted I_{Amh} and PRO. The principle is adapted from standard modal logic treatments of belief, and if we add it, we can make the right predictions, not only for reports with third person, but also with *I*, I_{Amh} and PRO. To capture the difference between these last three, some stipulation could not be avoided, so I simply reformulated Schlenker's typology of first person pronouns in terms of scope. The differences between my pragmasemantic account on the one hand, and the competing recent accounts of Schlenker and Von Stechow, who heavily rely on morphological agreement, on on the other, lie more in the third person realm: I maintained that the 'ambiguous' report's embedded *she* is interpreted wide scope and as a third person, whereas Schlenker and Von Stechow would need to interpret it as a first person for the *de se* reading, and as a third person for the *de re* reading.

References

Boër, S. and Lycan, W.: 1980, Who, me?, *The Philosophical Review* **89**, 427–466.

Chierchia, G.: 1989, Anaphora and attitudes *de se*, *in* R. Bartsch, J. van Benthem and P. van Emde Boas (eds), *Semantics and Contextual Expression*, Vol. 11 of *Groningen-Amsterdam Studies in Semantics*, Foris.

Cresswell, M. and von Stechow, A.: 1982, *De Re* belief generalized, *Linguistics and Philosophy* **5**(4), 503–535.

Dalrymple, M., Shieber, S. M. and Pereira, F. C. N.: 1991, Ellipsis and higher-order unification, *Linguistics and Philosophy* **14**(4), 399–452.

Geurts, B. and Maier, E.: ms, Layered DRT, `http://www.ru.nl/phil/tfl/bart` (2003).

Kamp, H. and Reyle, U.: 1993, *From Discourse to Logic, Vol. 1*, Kluwer Academic Publishers, Dordrecht.

Kaplan, D.: 1969, Quantifying in, *in* D. Davidson and J. Hintikka (eds), *Words and Objections*, D. Reidel Publishing Company.

Kaplan, D.: 1989, Demonstratives, *in* J. Almog, J. Perry and H. Wettstein (eds), *Themes from Kaplan*, Oxford University Press, New York, pp. 481–614. [Versions of this paper began circulating in 1971, page references to the 1989 publication].

Lewis, D.: 1979, Attitudes *de dicto* and *de se*, *The Philosophical Review* **88**, 513–543.

Maier, E.: 2004, Acquaintance resolution and belief *de re*, *in* L. Alonso i Alemany and P. Égré (eds), *Proceedings of the 9th ESSLLI Student Session*. http://lingua.fil.ub.es/~lalonso/stusESSLLI04/programme.

Maier, E.: to appear, *De re* and *de se* in quantified belief reports, *Proceedings of ConSOLE XIII*, Tromsø. http://www.sole.leidenuniv.nl/.

Percus, O. and Sauerland, U.: 2003, On the LFs of attitude reports, *in* M. Weisgerber (ed.), *Proceedings of SUB 7*, Konstanz. http://ling.uni-konstanz.de/pages/conferences/sub7/proceedings/download/sub7_percusSauerland.pdf.

Quine, W. V. O.: 1956, Quantifiers and propositional attitudes, *Journal of Philosophy* **53**, 101–111.

Reinhart, T.: 1990, Self-representation. Lecture delivered at Princeton conference on anaphora, October 1990. Ms. http://www.let.uu.nl/~tanya.reinhart/personal/Papers/De_se_91.wp.pdf.

van der Sandt, R.: 1992, Presupposition projection as anaphora resolution, *Journal of Semantics* **9**, 333–377.

Schlenker, P.: 1999, *Propositional attitudes and indexicality*, PhD thesis, MIT.

Schlenker, P.: 2003, A plea for monsters, *Linguistics and Philosophy* **26**, 29–120.

von Stechow, A.: 1982, Structured propositions, *Technical report*, Universität Konstanz, Konstanz. http://vivaldi.sfs.nphil.uni-tuebingen.de/~arnim10/Aufsaetze/Structured%20Prop%201.pdf.

von Stechow, A.: 2001, Schlenker's monsters. handout http://vivaldi.sfs.nphil.uni-tuebingen.de/~arnim10/Handouts/Schlenkers.Monster.pdf.

von Stechow, A.: 2002, Binding by verbs: Tense, person and mood under attitudes, *in* M. Kadowaki and S. Kawahara (eds), *Proceedings of NELS 33*, GLSA, Amherst, MA, pp. 379–403. http://vivaldi.sfs.nphil.uni-tuebingen.de/~arnim10/.

Zimmermann, T. E.: 1991, Kontextabhängigkeit, *in* A. von Stechow and D. Wunderlich (eds), *Semantik/Semantics: Ein internationales Handbuch der zeitgenössischen Forschung*, Walter de Gruyter, Berlin/New York, pp. 156–229.

Appendix: Maier's (2004) 2-layered fragment of Geurts and Maier's (ms) LDRT

Syntax: An LDRS is a set of *fr*(egean) and/or *k*(ripkean) labeled discourse markers, paired with a set of labeled conditions, e.g. $\left[\text{x}_{fr}\ \text{y}_{k} \mid \text{love}_{k}(\text{x,y}) \right]$

Semantics: Let φ be an LDRS, $m, l \in \{k, fr\}$, $\langle \mathcal{D}, \mathcal{W}, \mathcal{I} \rangle$ a model, $w \in \mathcal{W}$, and f a variable assignment:

a. definedness:
   - $[\![\varphi]\!]^f_{l,w}$ is defined iff there is an embedding g, $Dom(g) = Dom(f) \cup \{\text{x} \mid \text{x}_l \in U(\varphi)\}$ and for all $\psi \in Con(\varphi)$: $[\![\psi]\!]^g_{l,w}$ is defined
   - $[\![P_m(x^1, \ldots, x^n)]\!]^f_{l,w}$ is defined iff $\{x^1, \ldots, x^n\} \subseteq Dom(f)$
   - $[\![\neg_m \varphi]\!]^f_{l,w}$ is defined iff $[\![\varphi]\!]^f_{l,w}$ is defined

b. If defined, the semantic values of conditions and LDRSs are;
   - $[\![\varphi]\!]^f_{l,w} = \left\{ g \,\middle|\, Dom(g) = Dom(f) \cup \{x \mid x_l \in U(\varphi)\} \text{ and for all } \psi \in Con(\varphi) : [\![\psi]\!]^g_{l,w} = 1 \right\}$
   - $[\![P_m(x^1, \ldots, x^n)]\!]^f_{l,w} = 1$ iff $m \neq l$ or $\langle f(x^1), \ldots, f(x^n) \rangle \in \mathcal{I}(P)(w)$
   - $[\![\neg_m \varphi]\!]^f_{l,w} = 1$ iff $m \neq l$ or $[\![\varphi]\!]^f_{l,w} = \emptyset$

c. $[\![\varphi]\!]^f_l = \left\{ w \in \mathcal{W} \,\middle|\, [\![\varphi]\!]^f_{l,w} \neq \emptyset \right\}$ if $[\![\varphi]\!]^f_{l,w}$ is defined for some w (otherwise $[\![\varphi]\!]^f_l$ is undefined).

d. Contexts: $\mathcal{C} = \{w \in \mathcal{W} \mid \mathcal{I}(\text{center})(w) \text{ is a singleton}\}$

e. Truth-conditional content: if $[\![\varphi]\!]^f_{k,c}$ is a singleton, $[\![\varphi]\!]^{f,c} = [\![\varphi]\!]^g_{fr}$ where g is the unique element of $[\![\varphi]\!]^f_{k,c}$. Otherwise undefined.

f. Diagonal proposition: $\Delta^f(\varphi) = \left\{ c \in \mathcal{W} \,\middle|\, [\![\varphi]\!]^{f,c} \text{ is defined and } c \in [\![\varphi]\!]^{f,c} \right\}$

g. Belief set: $\mathcal{B}el \in [\mathcal{D} \times \mathcal{W} \to \wp\mathcal{C}]$

h. If $x \in Dom(f)$ and $[\![\varphi]\!]^f_{m,w}$ is defined, $[\![\text{believe}_l(x){:}\varphi]\!]^f_{m,w} = 1$ iff $m \neq l$ or $\Delta^f(\varphi) \supseteq \mathcal{B}el(f(x), w)$

Example: (18) $\rightsquigarrow$

$$\left[\begin{array}{l|l} \text{x}_k\ \text{y}_k & \text{Noriko}_k(\text{x}),\ \text{Tujiko}_k(\text{y}),\ \text{see\_in\_mirror}_{fr}(\text{y,y}) \\ & \text{R(x,w)}\doteq ? \\ & \text{believe}_{fr}(\text{x}){:} \left[\begin{array}{l|l} \text{u}_k\ \text{v}_{fr} & \text{center}_k(\text{u}),\ \text{R}_{fr}(\text{u,v}),\ \text{on\_fire}_{fr}(\text{v}) \\ & \partial\left[\text{w}_k \mid 1.\text{sg}_k(\text{w}) \right] \end{array} \right] \end{array} \right]$$