The lazy mindreader : a humanities perspective on mindreading and multiple-order intentionality

Duijn, M.J. van

Chapter 5

Language and joint intentionality: reflecting on orders of intentionality is the exception, not the default when communicating[*]

5.1 Introduction

In section 1.2.2 of Chapter 1 I have distinguished the three main roles of language in relation to mindreading, as used in this thesis. Language can *represent* mindstates and the relationships between them (first role), either formally, in propositions of the form "A thinks that B intends that C...etc.", or naturally, using a mix of different linguistic elements capable of viewpoint coordination. Various types of these linguistic elements were discussed throughout the previous chapters. Going from the smallest to the largest level of analysis, these were: lexical items such as the viewpoint packages "allegedly", "accidentally", and "mistaken", grammatical constructions such as complementation and the inquit-construction, the patterns of focalisation and reported speech and thought (STR) that coordinated the different perspectives presented in Woolf's *Mrs Dalloway*, and the expository strategies of narrative that aided the audience in understanding the complex thoughtscape underlying Shakespeare's *Othello*. What all of these linguistic elements had in common was that they provided conventionalised ways in which speakers of English could represent mindstates and the relationships between those in discourse—from a single belief held by one person up to an entire thoughscape.

[*] Versions of this chapter, especially Sections 5.3 and 5.4, were presented at the 47th Annual Meeting of the Societas Linguistica Europaea (SLE) in Poznan, Poland, 11-14 September 2014, and at the Perspective Project Kick-off Meeting in Nijmegen, The Netherlands, 17 November 2014. See the Reading Guide for more details.

In this way, it was argued that natural languages support not only efficient communication of mindstates and thoughtscapes, but seem also capable of providing *support for cognition* (second role). When looking at one particular usage event, language can provide a form of short-term "online" support: I have argued that the way in which mindstates and their mutual relationships are represented linguistically can execute strong influence on the ease or difficulty with which they could be processed (think of the expository strategies making a complex thoughtscape manageable and viewpoint packages conveying multiple intentional relationships at once in a holistic way). When looking at the longer term, language users somehow internalise ways in which language makes mindstates and their relationships insightful, which can account for what some researchers have referred to as "implicit support" for mindreading: I have discussed research suggesting that various aspects of language usage, once mastered, work as scaffolding, conceptual underpinning, or training for our intentional reasoning skills. For example, experimental evidence suggests that children aged 3-4 who were for a while exposed more intensively to embedding constructions and perspective-shifting discourse, pass false-belief tests earlier, presumably because their general "thinking repertoire" got enhanced when they learned to master particular grammatical patterns (Lohmann and Tomasello, 2003; Milligan et al., 2007). Also it has been suggested that stories in all their different appearances, ranging from the day's latest gossip or a myth told around the campfire to an award-winning novel, help both children and adults to develop and sharpen up their mindreading skills over time (see Chapter 1, Section 1.2.2; Chapter).

Language was also argued to be itself heavily *dependent on* our mindreading abilities (third role). In the current chapter, this dependency will be investigated in more detail. According to researchers such as Sperber (1994; 2000) and Scott-Phillips (2015), it takes the capacity to reason at four or five levels of intentionality to exchange even very basic utterances.[67] This position is intuitively controversial: if language is naturally capable of representing

[67] I agree with both Scott-Phillips and Sperber on many points regarding language and meaning. However, there also is an important issue on which I disagree: the way they construe the relationship between linguistic interaction (or, more generally, "pragmatic competence") and multiple-order intentionality, which they refer to as "recursive mindreading" or "recursive metarepresentation". This issue will be central in this chapter.

complex thoughtscapes, does that mean that an addressee of a short story involving, say, four mutually linked perspectives essentially has to deal with a total of eight or nine intentional states, four or five from the communicative situation plus four from the story? And if so, how can this be unified with evidence that dealing with multiple intentional states is cognitively demanding, or, for that matter, with the claim that humans can deal with *at most five levels* of intentionality reliably (see Chapter 1, Section 1.2.1)? What is the role of linguistic and narrative elements capable of viewpoint coordination, such as the viewpoint packages discussed in previous chapters? And how does all of this affect our evolutionary story?

These questions lead up to the objectives of this chapter, which has two main parts. First, as announced in Chapter 1, I will contest the claim made by Sperber and Scott-Phillips. Using the concepts of *common ground* (Clark, 1996) and *joint cognition* (Hutchins, 2006; Verhagen, 2015) I will argue that only in exceptional cases do we need to bother about any layers of intentionality. Regarding some aspects I will be relatively brief in my analysis, and refer to existing work or point out opportunities for future research. Other aspects, however, will turn out to be closely tied to points made in the previous chapters, and be elaborated in full detail. An important role will be played by the *ratchet effect*: linguistic items "store" communicative experience of generations of language users on which every new generation can build.[68] This, then, leads to the second part of this chapter: providing an *integrated conceptual model* for analysing the particular class of linguistic elements central in this thesis so far, namely: elements capable of viewpoint coordination in discourse. After the model has been introduced and explained, I will briefly explore some of its consequences for our evolutionary story.

[68] The term "ratchet" is taken from Tomasello (1999). My usage of it here is compatible with his, however, I apply the idea more specifically to linguistic items whereas Tomasello speaks about cultural conventions more broadly.

5.2 Association, ostension, and shared intentionality

Throughout the literature, it is quite generally recognised that human communicative interaction "as we know it" requires some form of mindreading on behalf of both interlocutors, irrespective of whether we use language, gestures, facial expressions, or any other means to get our messages across (see e.g. Verhagen, 2005; Levinson, 2006; Tomasello, 2008). The usual argument is that signallers have to design their communicative behaviour such that their particular audience will be able to infer from it what they mean, and addressees have to reckon why a signaller picked out this particular behaviour—both these tasks entail a degree of understanding of the other's mindstate. However, on top of this, some researchers have made a case for why human communication cannot succeed just by virtue of basic mindreading competencies. Indeed, Scott-Phillips makes this point explicitly: what most linguists and philosophers of language have failed to appreciate, according to him, is that sophisticated intentional reasoning skills including "recursive mindreading" are a prerequisite not only for the successful execution of communication-as-we-know-it, but also for such communication to evolve and develop at all (2015: 68-69). The argument thus has two components: it deals with the question "Which mindreading skills enable interlocutors to take part in communicative interaction as we know it today?" ("synchronic"), and with the question "Which mindreading skills were necessary for the emergence of such a form of communicative interaction in the first place?" ("diachronic"). It should be noted that these two components are not always brought forward and supported separately by Scott-Phillips, but in this chapter I find it useful at several points to keep the synchronic and the diachronic stories apart.[69]

[69] Note that this is a different divide than the "classic" one between ontogeny and phylogeny (Tinbergen, 1963). Here I mean not "development over a lifetime" versus "development over evolutionary time", but "the working of communication as it is now" versus "the emergence of such communication over time". In fact, my notion of "synchronic" is closest to Tinbergen's question of the proximate mechanism, whereas what I call "diachronic" covers both his developmental and evolutionary questions.

5.2.1 Scott-Phillips' two models of communication

The starting point of the idea advocated by Scott-Phillips goes back for a large part to Sperber (1994; 2000) and Sperber and Wilson (1995; 2002), and ultimately has its roots in what could be called the "pragmatic turn" in linguistics and philosophy of language that began with the second half of the twentieth century.[70] It sets human communicative interaction, whether or not involving language, apart against other communication found in nature by arguing that it is, at its core, not a system of "coding-and-decoding" information. A coding-and-decoding system can be found in (to follow Scott-Phillips' example) grasshoppers producing six different signals associated with six different states of the grasshopper world: "I would like to make love", "You are trespassing my territory", "How nice to have made love!", and so on (2015: 5, citing Moles, 1963: 125-126). Various forms of code-system communication can be found throughout the primate world, including in humans, ranging from olfactory cues (smell) guiding behaviour of newborns, to spontaneous emotional vocalisations working as alarm calls, and, potentially, (Duchenne) laughter signalling social solidarity (see Scott-Phillips, 2015: 5-6; Burling, 2005). Even though such code systems need not be fully deterministic or, for that matter, trivial, the primary mechanism linking signals to messages is *association*: every signal type stands for one particular meaning type (or, if a code is probabilistic instead of deterministic: a class of meaning types). However, association falls short of explaining human communicative interaction, given that there are many ways in which we can express a particular meaning, while at the same time all of our expressions can have multiple different meanings. The example given in Section 1.2.2 of Chapter 1 was that of someone saying to a friend "hey, there is Ann", which could mean "all right, we can go inside", "let's go somewhere else", "what a coincidence", and so on, depending on the circumstances. The same goes for non-linguistic communication: if we raise a full glass of beer in the air while seeking eye contact with someone who also holds a full glass, this probably means "cheers!", while it can also mean

[70] Sperber has developed his insights on the basis of Grice (1957), although Grice was not interested in evolution. For overviews covering also the important contributions made by Austin (1962), Searle (e.g. 1969), and Wittgenstein (e.g. 2006 [1953]) see Hacker (1986: esp. chapter 6-11) and Keller (1995).

"thanks!" if the other person is the one who just paid for the round, or "do you want one as well?" if the other is holding an empty glass.

Neither the utterance "hey, there is Ann" nor the behaviour of raising a glass and seeking eye contact stand for all of these meanings in the sense of the code model: there are no one-to-one associations. There is a different system at work that forms the basis for the production of meanings, which Scott-Phillips describes as the "ostensive-inferential" model of communication (2015: 7-13; see also Sperber and Wilson, 1995). According to this model, signallers have the intention to alter an addressee's mindstate or behaviour in some way. They provide particularly designed evidence for this, thereby enabling the addressee to draw the right inferences. This evidence can take the form of a string of words, but could, depending on context and desired effects, just as well be a set of gestures, facial expressions, or any other behaviour, as long as it is in some way *ostensive*: it has to be possible for the addressee to infer not only *what* the signaller wants her to understand (referred to as the "informative intention"), but also *that* the signaller is trying to communicate this in the first place (called the "communicative intention"). As an illustration, consider the example he borrows from Sperber (2000):

> Mary is eating berries. She wants Peter to know that she thinks that the berries are very tasty, so she eats them in a somewhat exaggerated, stylized way, and pats her tummy as she does so. This reveals two things to Peter: (i) that Mary thinks the berries are tasty (this is the content of her informative intention); and (ii) that Mary wants to communicate this fact to Peter (this is the content of her communicative intention). If Mary simply ate the berries enthusiastically, but did not do so in a stylized or exaggerated way, Peter would still be able to infer that they are tasty, but not because Mary had expressed either an informative or a communicative intention. There would be no communication in that case. (Scott-Phillips, 2015: 9)

In other words, given that there is no fixed set of signals associated with particular messages in this case, Scott-Phillips (along with Sperber) suggests that each signal must in principle first be negotiated *qua signal*—a process that

is explained by the ostensive-inferential model of communication, but not by the code model (see also Stolk, 2014, for discussion and an experimental approach). Ostension and inference are thus the basis of human communication, according to Scott-Phillips. On top of this, he argues, there is also a code at work: the conventions of a language provide global links of linguistic forms to certain meanings. In this way, ostension and inference make human communication possible in the first place, and the linguistic code makes it even more powerful (2015: 15-17).

The two different models of communication require quite different skills on a cognitive level.[71] In principle, the code model only requires a "glossary" listing all signals and associated meanings (which can be as simple as with the grasshoppers, but also more complex). This can be a genetically inherited glossary, but the capability to develop such a glossary through associative learning can also do the job. The ostensive-inferential model, by contrast, requires a great deal of flexible reasoning abilities, including mindreading. In order to design the right evidence for their intended meaning, signallers need not only take into account the context (where and when the communication takes place, who is present, etc.), but also what their addressees (already) know and believe about the topic and context. Addressees, in turn, must factor in what they think the signaller believed about them, the topic, and the context when designing the signal, in order to make the right inferences. Both interlocutors must thus be able to reason about contextual factors, including the other's intentional states, for ostensive-inferential communication to be possible.

I support the distinction between the code model and the ostensive-inferential model and agree with the analysis that the requirements on the cognitive level are the ability to form associations in the case of the code model, whereas flexible reasoning abilities including mindreading are needed in the case of ostensive-inferential communication. Yet this is where the debate begins: I disagree with the amount and complexity of the mindreading Scott-

[71] My aim in this section is clearly not to provide *exhaustive* lists of what is required for communication on a cognitive level. Rather, I will highlight important differences between the kind of cognitive structure needed for the code and inferential models to work (see also Scott-Phillips, 2015: 64), and in 5.2.3 I will do the same for my alternative communication model.

Phillips and Sperber consider to be necessary. In what follows I will argue that they misconstrue the complexity needed *in theory*. On top of that, I will argue that *in practice* we hardly need any mindreading at all for successful communicative interaction, by discussing various mechanisms that save interlocutors from cognitively taxing mindreading efforts.

5.2.2 Cognitive requirements of ostensive-inferential communication

As said at the beginning of the previous section, Scott-Phillips explicitly makes the point that many others across the literature agree that *some* mindreading is needed for human communication, but that its exact role and complexity are rarely spelled out. In order to get a grasp on this, he sets up the following argument, using a series of different scenarios taken from Sperber (2000):[72]

> Scenario one. Mary is picking and eating berries. She does this because the berries are edible.
>
> *Scenario two.* Mary is picking and eating berries. Peter is watching her, and hence forms a belief about the edibility of the berries. Here, *Peter believes$_1$ that the berries are edible* (because otherwise Mary would not be eating them). Note that Mary may or may not know that Peter is watching. Whether she does or not, it makes no difference to her intentions or behaviour.
>
> *Scenario three*. Mary is picking and eating berries. Peter is watching her. Mary knows Peter is watching her, and she wants him to believe that the berries are edible. So: *Mary intends$_1$ that Peter believes$_2$ that the berries are edible*. Here, note that Mary's behaviour is identical to her behaviour in scenarios one and two. All that has changed is that in scenario two Mary informed Peter about the edibility of the berries only incidentally [...] whereas here she does so intentionally – and she can satisfy this intention (that Peter believes hat the berries are edible) simply by picking and eating berries. She need not and does not do anything more than this. Mary's intention is an informative intention.
>
> *Scenario four*. Mary is picking and eating berries. Peter is

[72] Note that Scott-Phillips uses numbers in subscript to indicate orders of intentionality: "Mary intends$_1$ that Peter believes$_2$ that...".

> watching her. Mary knows Peter is watching her, and she wants him to believe that the berries are edible. Furthermore, Peter knows that Mary knows that he is watching her and, for whatever reason, he has reason to believe that she would like him to believe that the berries are edible. Correspon-dingly, *he believes$_1$ that she intends$_2$ that he believes$_3$ that the berries are edible*. Mary, however, does not know that Peter believes this. After all, she has not yet made her intention manifest to Peter. Indeed, Mary's physically observable behaviour is the same as it is in scenarios one, two, and three. As yet, she has not picked berries in a way that signals to Peter that her behaviour is intended to be informative. She has not yet signalled signalhood. All that is different between this and scenario three is that here Peter believes, correctly, that Mary has an informative intention.
>
> *Scenario five*. Mary is picking and eating berries. Peter is watching her. Mary knows Peter is watching her, and she wants him to believe that the berries are edible. Furthermore, Peter knows that Mary knows that he is watching her, *and* Mary knows that Peter knows this. As such, when she eats the berries, *she intends$_1$ that he believes$_2$ that she intends$_3$ that he believes$_4$ that the berries are edible*.
> (Scott-Phillips, 2015: 65-66, based on Sperber, 2000, and Grice, 1982; italics and subscript numbering in original)

Scenario five embodies a significant leap according to Scott-Phillips: here Mary has reason to change her behaviour from regular, unremarkable picking to any degree of slightly stylized or exaggerated picking. She now has two intentions, the informative intention (labelled $_3$) she had earlier and the communicative intention ($_1$) to "signal signalhood", which is new to this scenario. However, only if Peter recognises both intentions, "ostensive-inferential communication proper" has emerged:

> *Scenario six*. As per scenario five, including the fact that Mary picks and eats berries in a particularly stylized, exaggerated manner. Because of this, Peter grasps both of Mary's intentions, informative and communicative, as laid out above. As such, *Peter believes$_1$ that Mary intends$_2$ that he believes$_3$ that she intends$_4$ that he believes$_5$ that the berries are edible*. (idem)

Scott-Phillips states that in a world with only the scenarios one to four, there would be no difference between doing things because you need or want to, and doing things in order to communicate with others, since "nobody would signal signalhood" (2015: 67). Only in scenario six is signalhood signalled *and* recognised. At this stage, a form of *interdepence* between signaller and addressee has emerged which Scott-Phillips considers to be a defining characteristic of human communicative interaction: this only obtains if there is mutual recognition of the communicative intention to exchange a particular informative intention, presupposing four and five orders of intentionality to be handled by the speaker and addressee respectively.

After having laid out this strand of reasoning, Scott-Phillips anticipates three types of critique (2015: 68-75): (i) scenario five and six look complicated and cognitively taxing, while we all know from experience that communicating in this way is not; (ii) experimental evidence suggests that children and patients suffering from autism spectrum disorders cannot reason at higher orders of intentionality, but they certainly can be communicatively competent; and (iii) experimental evidence suggests that the limit of orders of intentionality for normally developed human adults lies around five, suggesting that communicative interaction as such is already at the limit. With respect to (i) he points out that there is no a priori reason to assume that something we experience as simple, is also simple in formal terms. He draws a parallel with vision: a formal model of this skill will clearly not be as straightforward as the act of seeing itself feels to us (see Scott-Phillips, 2015: 10). I agree with this in principle, however, we should of course note that this does not work as an argument the other way around: the alleged discrepancy between how vision feels from experience and how complex it may be formally, does by no means entail that everyday communication, feeling easy, should be complicated in formal terms. Besides that, a reason why Scott-Phillips' parallel might not be a feasible one is that vision, being widely spread throughout nature, and pragmatic competence, being unique to humans, require explanations on very different evolutionary time scales. Without a priori excluding anything in the case of human evolution since the divide from the other great-ape lineages, the shortage of evolutionary time is an argument for looking at the most

economical hypothesis in terms of cognitive complexity first (see also Tomasello, 1999).

Regarding (ii) and (iii), Scott-Phillips explains that there are in principle two ways out of the seeming contradictions posed by these types of critique: either the analysis he (and Sperber) set out overcomplicates the matter, or the experimental evidence is wrong and dealing with multiple-order intentionality is much easier and less effortful compared to what is generally assumed. Clearly, Scott-Phillips sets out to argue for the latter. I agree at least partly with him on this point, and do also think that there are issues with the ways in which the experimental evidence has been produced and interpreted (Chapter 6 will deal with this in more detail). However, the two ways out of the seeming contradictions he suggests are not mutually exclusive: besides agreeing that there are some issues with the experimental evidence, I still think that his analysis overcomplicates the matter—in the next subsection I will explain why.

In short, his argument is thus that he sees no possibility to leave out any of the steps of recursive mindreading leading to the emergence of "ostensive-inferential communication proper", as cited above. Therefore, he states, experimental evidence must be at least partly wrong when suggesting that mindreading involving five orders of intentionality is highly cognitively demanding (see Chapter 1, Section 1.2.1), developing late in childhood or even adolescence (idem), and not available to people suffering from certain cognitive disorders (idem). He suggests that ever since mindreading has been investigated experimentally, starting from the late 1970s, the *actual* abilities of human test subjects have been masked by methodological shortcomings. For example, as soon as false-belief tests were carried out "implicitly", i.e. not using explicit questions of the type "Where does she think the sweets are hidden?", the age at which children were able to pass them could be brought down dramatically (from around 3-4 years of age to 12-18 months; see Baillargeon et al., 2010, but see also Apperly, 2011: 29-30 and Heyes, 2012). Similar arguments can be made for people suffering from several psychopathological conditions: different tests have led to better results.

These are indeed important points, which should be kept in mind for the next chapter. However, Scott-Phillips takes them too far in my view: in Chapter 6 I will analyse an "implicit" version of the mentalising test designed under his

supervision, which allegedly demonstrates that normally developed human adults are capable of handling up to eight or nine orders of intentionality effortlessly. His line of reasoning is the same here: he suggests that previous versions of the test have masked the actual performance, and that this is the first one being ecologically valid, thus providing insight into the capacity as it "really" is. He concludes: "There are good reasons, both theoretical and empirical, to conclude that recursive mindreading is not cognitively demanding. More likely, it is, like simple mindreading, something we do habitually and subconsciously, as part of our everyday, low-level perception of the world around us" (2015: 73). I will get back to this in the next chapter. In what follows here, I will take the other of the two suggested paths: instead of (only) criticising the existing experimental evidence, I will (also) scrutinise Scott-Phillips' theoretical analysis of communicative interaction and argue that it is misguided regarding the amount of mindreading complexity it presumes to be necessary.

5.2.3 Individual versus shared intentionality

Brought back to its core, the point I intend to make here can be summarised as follows: whereas the basic cognitive unit in Scott-Phillips' (and Sperber's) analysis is that of a human individual, I argue that the basic cognitive unit of human communicative interaction should rather be understood as *at least two people sharing a great deal of beliefs and intentions*. As a consequence, all the steps suggested to explain how two individuals reach a state of mutual recognition of communicative and informative intentions are rendered superfluous. In other words, where Scott-Phillips sees communicative interaction as a process in which signaller and addressee have separate sets of intentional states which they eventually seek to "pair", I suggest to see communicative interaction as a process in which interlocutors sharing a set of intentional states negotiate what is and what is not part of their shared intentionality.

My view relies for an important part on Clark's work on *common ground* (1996) and *joint projects* (2006), and on Verhagen (2015), who brings together insights from Tomasello's and Rakoczy's notion of *self-other equivalence* (2003; see also Tomasello, 2008; 2014), Searle's *we-intentionality* (1995), and Hutchins'

work on *group-level cognition* (1995; 2006). In the rest of this subsection I will provide some more details on the idea of shared intentionality. After that, I will explore what a communication model based on joint intentionality demands from interlocutors on a cognitive level, especially with respect to mindreading.

The core idea of shared intentionality underlying communicative interaction is that interlocutors consider a particular set of beliefs and intentions to be mutually shared or "common ground". Or perhaps rather: they *a priori* act as if these intentional states are shared, until they have evidence to the contrary. Which set of beliefs and intentions is considered common ground depends on the identity and situatedness of the interaction partner(s). As suggested by Clark and Marshall (1981) there are multiple types of "sources" of common ground. First of all, people can be in the same here-and-now, which is in linguistics generally referred to as sharing the same "Ground" (Langacker, 1990; Verhagen, 2005). In that case they can, for example, use deictic expressions (including pointing and eye gaze) under the assumption that the other can figure out what they mean: they *both believe* that "now" stands for the same moment in time, "here" for the same place, "the book over there" refers to a particular book of which they *both believe* it is that one rather than another, and so on. Another source can be one's personal relationship to someone, formed by a shared history of previous interactions. When speaking to a friend I can, for example, refer to a mutual acquaintance by just using her first name "Susanna", given that we are used to referring to her in this way—in other words, we *both know* who we mean. However, if I want to refer to the same person when speaking to my mother, I may have to say "Susanna Smith", since when my mother and I say "Susanna" we usually mean a different Susanna. In the case of strangers who do belong to the same cultural-linguistic group as we do, most of the common ground is less specific, but we can still assume that a great deal of beliefs are shared, most notably of course the conventional rules of our language. If I produce the sound "huis" (meaning "house" in Dutch) in front of any stranger in the Netherlands, there is a big chance that both our individual histories have assured that we *both think* of a fairly similar concept. In fact, within the Netherlands it is safe to assume this until encountering alongword

evidence to the contrary.[73] Finally, our common ground with all other humans who are strangers and not members of our cultural-linguistic community is still sufficient for some form of shared intentionality to support communicative interaction: this is what we rely on, for example, when looking upward to refer to the sky, or when referring to "food" or "eating" by miming that we take a bite.

Note that people who are in one's personal social network, often also share membership of a cultural-linguistic community, and that both people in one's network and strangers who are members of this community are humans. Therefore, my version of the sources for common ground, adapted from Clark and Marshall (1981) and Verhagen (2015), can be conceptualised as a series of concentric circles:

[73] Verhagen (2005) makes a categorical distinction at this point: (i) all linguistic signals, which rely on sounds (or signs, writing, etc.) being conventionally linked to particular functions within a linguistic community; and (ii) linguistic items that, on top of this, rely on particular knowledge shared between communicators, such as deictic expressions presupposing shared knowledge of the communicative situation or Ground. Tomasello and Rakoczy (2003) rather seem to suggest a continuum where shared knowledge can be very general within a linguistic community on the one side, and very specific between two interlocutors on the other side. This is what I suggest here too.

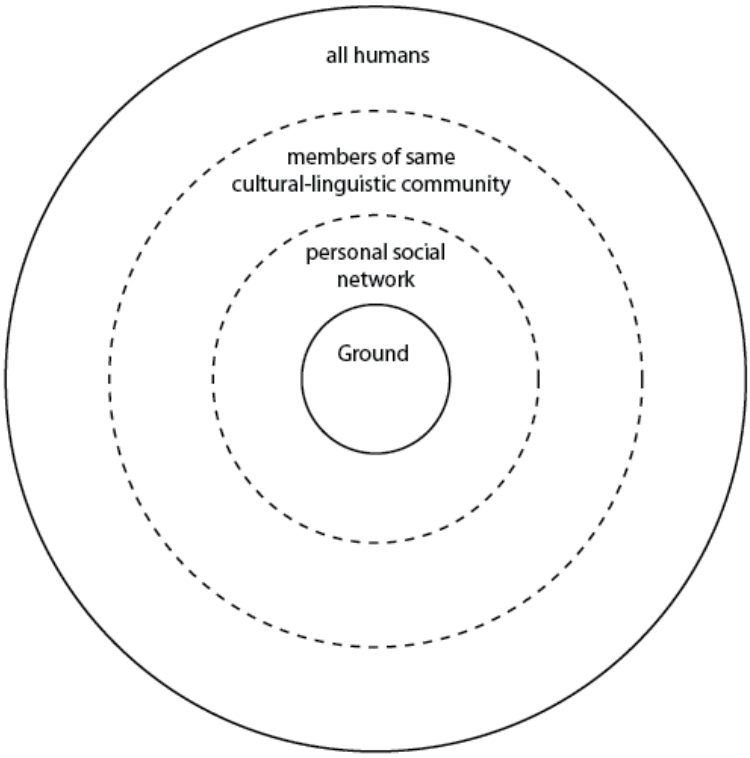

Figure 1 – Types of "sources of evidence" for common ground, adapted from Clark and Marshall (1981) and Verhagen (2015) (I have added the outer circle, slightly altered the categories, and introduced dashed lines for the second and third circles). Which knowledge is considered to be part of the common ground depends on who the interaction partners are. Are they part of the same interaction event, and do they thus share the same here-and-now, i.e. same Ground? Are they people with whom I have a history of interaction? Are they members of the same linguistic and cultural community? Are they humans (or perhaps the question should be: are they normally developed human adults)? The dashed circles indicate that it is possible that interlocutors sharing the Ground can but need not be part of each other's personal social network, and can but need not be members of the same linguistic community.

Groupings like "all Dutchmen" or "all Italians" may yield strong prototypical examples of cultural-linguistic communities, but the definition of such communities also extends to, say, Londoners, dentists, fans of The Police, Oxford students, cricket players, generative linguists, Jehovah's Witnesses, and so on. Whenever a Dutch dentist meets an American dentist, there will be particular knowledge they can consider to be shared on the basis of the community they take part in by virtue of their profession. This probably includes particular experiences and practices, but may also involve a specific

lexicon (sometimes also known as *jargon*, e.g. "endodontics" for root canal therapy) or certain behavioural conventions (e.g. never provide details about a patient's identity). Clark (1996; 2006a) points out that some communities are nested (e.g. Londoners, Brits, speakers of English) and others are cross-cutting (e.g. Oxford students, Police fans, speakers of English). Viewed this way, people are members of many different communities. When two people interact, they generally have a gradation of common ground, based on the amount of shared community memberships: for example, any two Oxford students can assume that the other knows what to wear when sitting exams, whereas Oxford students who are also members of a particular college can not only assume that the other knows about exam dress codes, but also about who used to live upstairs from the old kitchens.

Note that if shared knowledge is indeed a crucial basis for communication, one would expect that people interacting for the first time try to assess whether they share membership of one or more communities, potentially providing them with some common ground. According to Clark (2006a) this is indeed why most conversations with strangers begin by exchanging information about residences, interests, occupations, and so on. At the same time, accents, dressing style, or other aspects of people's overall "habitus", may work as cues (overtly or in disguise) for membership of particular communities. This is not just a matter of finding "something to talk about", but goes much deeper: it is about finding out what the conventional rules underlying the interaction are. To start with, common ground includes knowledge of what to consider as a meaningful signal. Cricket players may draw crucial inferences from gestures hardly even noticeable by outsiders. Or what is just a plate with some used cutlery on it for a member of one community, may to members of another community signal "I haven't finished eating yet". Broadly speaking, these examples are not very different from the fact that speakers of any spoken language recognise speech sounds as

meaningful signals, but only some also recognise "clicks" as such.[74] In all cases, membership of a particular community has, over time, assured that individuals take a particular bit of behaviour as being meaningful in a communicative setting.

Among members of these communities there is thus no need to negotiate these behaviours *qua signal*, to "signal signalhood" in Scott-Phillips' terms. Once they have identified an interaction partner as a member of the same community, hence established a basis for common ground, they can use a signal in the same way as this signal has been used towards them by members of this community. This is what Tomasello and Rakoczy (2003) have referred to as *self-other equivalence*, leading not only to community-wide consensus over what does and does not count as a meaningful signal, but assuring also that linguistic form/meaning pairs become intersubjectively shared within cultural-linguistic communities. In other words, if the principle of self-other equivalence is systematically adopted by members of a community towards other members of this community, this yields a mechanism through which conventions spread reliably. These conventions can be taken in the broadest sense, ranging from what to regard as a signal in the first place, or what to communicate about and what to leave implicit, to more specific conventional associations between forms and meanings such as the gesture "thumb up" signalling a positive attitude, the sound "bal" referring to a round object suited for playing particular games, or the word order "John hits Peter" meaning that Peter is at the receiving end of the action.

The crux in the case of cultural-linguistic communities is that no previous contact between two individuals within a community is needed for them to have a similar set of knowledge states "installed" on their individual cognitive systems. On top of or besides such communal common ground, personal interaction is another way in which shared knowledge can be built, updated, and extended. As soon as two people start interacting, they not only do this

[74] E.g. the Khoeid languages spoken by hunter-gatherers peoples in Namibia (see Voßen, 1997). The difference is of course that such sounds are elements constituting symbols through combination—a very powerful trait of human language—whereas cutlery arranged on a plate is a symbol by itself. There is clearly a lot more to say about how combinations of sounds become conventionally associated with particular meanings, both developmentally and evolutionarily, but that is not relevant for my purposes here.

"against the common ground they believe they already share [but also] as a way of adding to that common ground" (Clark, 2006a: 107, referring to Stalnaker, 1978). This can involve superficial updates (e.g. finding out about a mutual acquaintanceship with "Susanna Smith", enabling unique reference using her name), but also go as far as two people (or a small clique of a few people), who interact frequently, developing their own words, constructions, accent, behavioural rules, and so on.

In this way, we have specific common ground with all people in our personal social networks, besides or on top of the common ground we might have with them by virtue of shared membership of various cultural-linguistic communities, sharing the same "here-and-now" of the interaction ("Ground"), and/or, in the minimal case, being human. An important observation can now be made: once the degree and nature of common ground with a particular interaction partner have been established (possibly through visible cues, accent, conversations about interests, occupation, residence, etcetera), it is possible to derive inferences about this interaction partner's *individual* knowledge, *if need be*. For me as a speaker of Dutch it is possible infer about another speaker of Dutch that she will *know* that the sound "huis" can be used as a signal to draw the attention to some house. Also, I can draw the inference that she will *know* that I *know* this, given that she knows that I am a speaker of Dutch too. Theoretically, I can draw infinitely many inferences like this about what other speakers of Dutch know, what they know that I know, what they know that I know that they know, and so on (this point is also made in different forms by Clark, 2006a, and Verhagen, 2015, both referring to Lewis, 1969). However, this is not what I need to do *by default* before communicating with them, given that "as such" these inferences add nothing new: each of them is a *derivate from* the already existing common ground between all speakers of Dutch, not a step *towards* the emergence of such common ground. The same holds true for any form of common ground. When I sit behind my desk and my office mate has gone out, before going home I can leave a note on his desk saying that I won't be "in HQ" before next Tuesday, *knowing* that he will *know* that I *mean* our office by HQ ("Head Quarters"). And he will know that I know that he knows what I mean. Also, when we both sit behind our desks, I can point towards the windowsill, where our coffee machine is situated. He may

nod, which I can take as an indication that he indeed would fancy a cup. This works because I *know* that he *knows* where our coffee machine stands, and I *know* that he *knows* that I *know* this, and I *know* that he is a coffee drinker, and that he *knows* that I *know* that he is. All these inferences about who knows what about the meaning of "HQ", or the location of the coffee machine and the desire of it being put to use, can be *derived from* our personal common ground and the common ground provided by our co-presence in the same here-and-now. Most of the time, we never get around to drawing such inferences, although it is possible to think of contexts in which we might do so. For example, imagine him saying "no thanks" once I put the freshly brewed cup of coffee in front of him. We might enter a conversation about what went wrong in the previous communication: I could say that I *thought* he *wanted* coffee because he nodded when I pointed at the coffee machine, to which he might reply that he *understood* I *wanted* to lower the blinds and therefore pointed at the window. In this example, intentional reasoning seems to enter the stage only because of a misunderstanding inciting some reflection on differences in perspectives. Towards the end of this chapter I will follow up on this point of explicitly constructing the different perspectives involved in an interaction event, also in the light of the distinction between holistic and compositional complexity introduced in the previous chapter.

The analysis at this point closely resembles what I have covered in Chapter 1 by introducing the "Schelling mirror world" (following Levinson, 2006). Interlocutors toss into the Schelling mirror world a piece of behaviour (words, gestures, facial expressions, or otherwise) which they consider suitable for letting the other draw the desired inferences. "Meetings of the mind" (Levinson, 2006: 49) in this world can, as discussed, occur by virtue of having a shared sense of salience. We can now see that such a sense can be derived from common ground using the principle of self-other equivalence: I can pick the behaviour *I* find most appropriate in order to evoke a particular inference in my interlocutor, which is, given our common ground, by definition also the most appropriate inference in her eyes. Only if this goes wrong, may we need to figure out what happened asking ourselves what the other was thinking, and possibly what the other thought we were thinking, what the other thought we were thinking the other was thinking, and so on. Even without a previous

misunderstanding we may want to reflect on the communicative situation in such a way, perhaps for purposes of analysis or strategic planning ahead. Crucial, however, is that complex mindreading or intentional reasoning comes with such reflection, and is not relied on by default.[75]

Or is it after all? How often do we need such reflection? Are we not in need of reflection on the communicative situation all the time, either implicitly or explicitly? As also mentioned in Chapter 1, various mechanisms seem to be saving communicators from having to engage in cognitively demanding mindreading most of the time. First, following the idea of *relevance* as worked out by Sperber and Wilson (1995; 2002) most communication comes down to the signaller picking the first (i.e. the most relevant) expression that comes to mind and the addressee picking the first interpretation that comes to mind. Sperber and Wilson themselves argue that sophisticated mindreading skills are needed for this process. However, I agree with Apperly that they seem to overlook that especially their updated account of relevance (2002) renders mindreading almost entirely superfluous: given that interlocutors are "in complementary predicaments", it is "a good bet for [them] to follow their own individual cognitive paths of least resistance" (Apperly, 2011: 115). Both pick the most relevant expression or interpretation first; if this does not lead to a satisfactory result, they can try the second-most relevant expression or interpretation, thus working downwards on the gradient of relevance. This fits with what various psycholinguists studying "alignment" have found: Pickering and Garrod (2004) argue explicitly that due to these mechanisms (relevance, alignment) interlocutors can refrain from constantly making inferences about the other's mindstates (see also Apperly, 2011: 116). Besides this, Apperly makes another point that is relevant here: in everyday interaction, we do not have to go to the bottom of everything. Rather, we seem to work with representations

[75] Note that Tomasello seems to come to a similar conclusion in his 2014 book (see especially page 38). However, it is unclear from this passage, and from the parts of his 2008 book that he refers to here, what his exact position is in "diachronic" and "synchronic" terms. Does he see shared intentionality as a feature that emerges from and is conceptually underpinned by layers of embedded intentionality, but is in practice usually not decomposed into these constituting layers? In that case he would defend the same diachronic story as Scott-Phillips (2015) does, but a (somewhat) different synchronic one. Alternatively, his view could be that both the diachronic and synchronic stories can do without these layers, which is what I suggest in this chapter and in the Conclusion below.

that are "good enough" for the interaction to keep going, but no better (2011: 114-119 and personal communication). If required in a conversation, interlocutors can together work out a particular point in more detail, aiding and steering each other in the desired direction turn by turn. In Levinson's (2006) terms: many conversations do not have a "signal-response" structure, but rather one of "testing-adjusting-retesting".[76]

What is more, not only are signaller and addressee experienced in choosing the most relevant cues and interpretations, the linguistic tools they have available also contain a wealth of such accumulated "experience". After all, they have emerged in the course of generations of language users attempting to coordinate the perspectives of themselves, their interlocutors and possibly third-party referents. Verhagen (2005; 2015) shows, for instance, for deixis, negation, and particular discourse connectives how they work "argumentatively" in the process of negotiating how (potential) deviations from the common ground can be resolved, in order for the interaction to be able to continue. This, then, introduces the issue central in the second part of this chapter: starting from a set of shared intentional states as defined by the interlocutors' common ground, it is possible to single out and highlight differences between individuals and the non-shared part of each of their intentional states, thereby enabling negotiation about how the common ground should be updated. This is the domain of viewpoint coordination in discourse, for which language has a great number of specific tools, some of which have been discussed in the previous chapters. The next section discusses a conceptual model for analysing this class of linguistic tools.

The final question remaining for this subsection is what kind of structure is required on the cognitive level for this alternative model of interaction, starting from shared intentionality or common ground, to work. It is important to note that I am not suggesting that processing efforts needed to determine what is and what is not part of the common ground with an interaction partner

[76] This again fits well with Clark's work on conversations as *joint projects*, in which interlocutors implicitly commit themselves to particular goals (which could be anything from setting a coffee meeting for tomorrow to cooperatively completing a complex building task) and converse about how the common ground has to be updated in order to achieve these goals (see Clark, 2006).

are always little or insignificant compared to those needed for ostensive-inferential communication according to Scott-Phillips. However, I argue that such efforts start at or close to zero by default (given that common ground is assumed a priori), and are then scaled up if necessary. In contrast to this, Scott-Phillips suggests that we use "recursive mindreading up to level five" by default. At least regarding the "synchronic" part of the story (see Section 5.1 above), a list of requirements needed for my alternative model would look like this:

(i) quick and fairly accurate abilities of distinguishing between individuals belonging to our own social network and/or particular cultural-linguistic communities we participate in;

(ii) abilities to keep track of former interactions we had with others, and access these "records" during interaction;

(iii) the capability to apply the principle of self-other equivalence, as needed to operate in the "Schelling mirror world";

(iv) the ability to reflect on individual perspectives and how they deviate from the common ground when prompted to do so, especially in the light of misunderstandings or other special circumstances that require scaling up of processing efforts.

In the conclusion of this chapter I will get back to this list from a "diachronic" perspective, thus exploring consequences for our evolutionary story.

With all of this in mind, let us now briefly return to Scott-Phillips' and Sperber's example of Mary and Peter communicating about berries being edible. I suggest that Mary and Peter live in a world where particular beliefs and intentions can be considered to be shared by default. Therefore, they can under normal circumstances bypass any intentional reasoning about what the other believes, what the other believes that they believe, and so on, including what the other will believe to be a meaningful signal and what not. If Mary wants Peter to believe that the berries are edible, she can just carry on picking and eating. Only if she wants to *deviate* from this default do things become more

complicated: for example, if she wants to mislead him about the berries being edible, or if she realises that he should not be thinking they are.[77] In this case, she can reflect on the situation, deriving from the common ground what Peter will believe and reasoning about how this will change according to her behaviour. In other words, she can begin negotiations with Peter about how the common ground should be updated without further difficulties, and only if need be will she unpack the default situation, thereby scaling up processing effort.

5.3 Coordinating mindstates in discourse

In this section I will propose a conceptual model for analysing the broad and diverse class of linguistic items capable of viewpoint coordination. The purpose of my model is not primarily to introduce another practice of drawing schemas, next to for example Dancygier's (2012) narrative spaces framework (as used in Chapter 3 and 4) or Fauconnier's (1985) mental spaces. It *can* be used for schematically representing individual viewpoint configurations as prompted by particular linguistic items, and I consider it illuminating to do this a few times when explaining its details, but the main purpose of including it here is to make a point about the general structure and working of linguistic interaction. This point is, to provide an anticipatory summary, that many linguistic items not only work to draw attention to some object or concept in the world—a function often described as *reference*—but also, and mostly at the same time, to provide and manage *perspectives on* or *stances towards* these objects or concepts. In Chapter 3 and 4 I have referred to this as the polyphonic nature of discourse. As such, the point that most language usage also entails viewpoint coordination is recognised by many linguists and narratologists; however, focus has mostly been *either* on how signallers and addressees mutually coordinate their perspectives (see e.g. Langacker, 1990; Sweetser, 1990; Verhagen, 2005), *or* on

[77] In his original discussion of the example, Sperber (2000) does address the special case of misleading someone else. However, this does not alter his analysis that even in basic, straightforward cases intentional reasoning up to five orders is needed.

how third-party perspectives are represented (e.g. Fludernik, 1993; Bal, 2009; Hühn, Schmid and Schönert, 2009).[78] The model of linguistic interaction proposed here integrates these functions all at once, instead of approaching them as distinct phenomena, thereby capturing the polyphonic nature of discourse in a conceptual model—giving it a schematic "face", as it were.

5.3.1 Dyadic and triadic communication

Communication in non-human animals typically involves a sender producing some observable behaviour (the "signal") that increases the likelihood of a receiver responding, i.e. behaving, in some particular way—for example, a bird signalling to a competitor to stay away from his territory. When the benefits of such a pattern of linked behaviours outweigh the costs for both senders and receivers, a (relatively) stable communication system may emerge. Thus, most non-human communication is about "regulating and assessing the behavior of others" (Owings & Morton, 1998: i). At this very basic level, the conceptual space needed to characterise communication is one-dimensional: no other dimension than that of the sender-receiver relationship is necessarily relevant to characterise a signal and its causes and effects. In the words of Tomasello (2008: 23), animal communication is mostly "dyadic": by far the majority of cases can be explained in terms of regulating others' behaviours without having to take into account attention (let alone *joint attention*) to any objects of reference.

[78] An exception is Dancygier (e.g. 2012), and to some extent also Vandelanotte (e.g. 2009): their approaches also integrate insights from linguists' interaction models with narratological views on speech and thought representation.

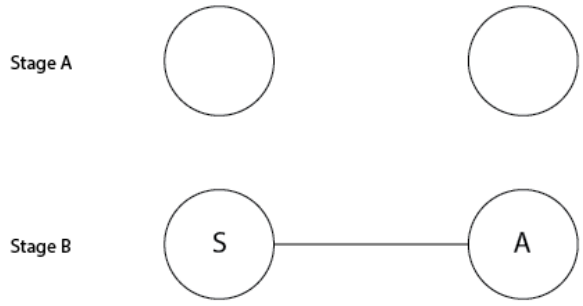

Figure 2 – Schematic depiction of "dyadic" communication as found in non-human animals. In Stage A there is no communication between the two depicted subjects (circles). Stage B shows the situation in which the sender (S) transmits a signal to the addressee (A) in order to influence the latter's behaviour, thus establishing one-dimensional communication. (Possibly, this induces a response signal: the subject on the right then becomes S and the subject on the left A.)

In contrast, human communication is prototypically "triadic" (Tomasello, 2008: 23) as it is by default *about* referents in the (shared) world outside of the communicators and their interaction. Following this idea, the conceptual framework needed to characterise normal human communication should thus at least be two-dimensional: apart from the relation between the communicators, the relation to the world must be taken into account to characterise signals and their causal connections. In other words, at the heart of interaction using language lies joint attention to some object of conceptualisation (person, event, relationship, etc.) and negotiating a particular stance towards this object.[79]

[79] Nonetheless, there are some instances of communication in non-human animals where functional reference to objects in the "outside world" does seem to play a role; a well known case is that of vervet monkey calls referring to different types of predators (see Seyfarth, Cheney, and Marler, 1980; Dennett, 1987: chapter 6). Conversely, humans also regularly engage in purely dyadic communication, such as greetings ("Hi!") or warnings ("Watch out!"). However, as Owings and Morton (1998: 211) argue, functional reference in animal communication should not be analysed as providing information about entities in the world, since it would confuse short term with long term causation; objects such as a snake in a snake alarm call should be seen as "long-term validators of the signal's utility", not as real-world correlates of signals which are causally involved in the receiver's response to the signal. In human communication, however, triadic communication *does* prototypically involve real-world objects of joint attention.

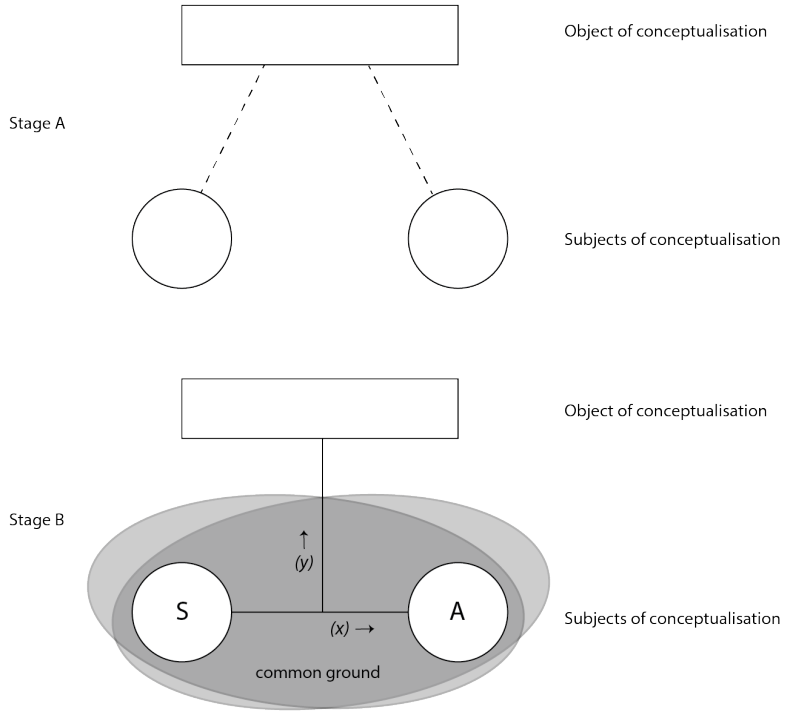

Figure 3 – In Stage A, where no communication takes place, the two subjects (circles) both have their own views (dashed lines) on some object (rectangle). Stage B depicts triadic communication: the signaller/speaker (S) and addressee (A) both assume a set of shared beliefs (the overlapping part of which is the *common ground* discussed in Section 5.2 above) and subsequently negotiate how the common ground should be updated with respect to the object. As will be discussed below, the used signals typically reflect both aspects of and operations on the relationship between S and A (the *(x)*-axis) and on the relationship between the common ground and the object (the *(y)*-axis).

In simple terms, the two-dimensional conceptual space depicted in Figure 3 makes a distinction between the "intentional" aspect of language, its capacity to be *about* some object in the world, and the "(inter)subjective" aspect of language, according to which sender and addressee negotiate a particular stance *towards* this object. These aspects are depicted by the *(y)*- and *(x)*-axes respectively.

The field of cognitive semantics also embraces the idea that the proper characterisation of language use requires a two-dimensional conceptual framework, but its background and history differ somewhat from the biological and psychological considerations made here. In fact, the cognitive view was explicitly designed by Langacker (1987) in opposition to so-called "objectivist"

approaches to semantics, which held that meaning in natural language could be fully characterised in terms of no more than its relation to the/some world (its contribution to "truth conditions"). Objectivist semanticists were thus precisely ignoring the "perspectival", "(inter)subjective" *(y)*-axis. This axis is indispensable in Langacker's view, since he claims that different "perspectival construals" are just as inherent components of linguistic meaning as reference is.[80]

What I will argue now is that a proper characterisation of viewpoint management in discourse (and of linguistic elements supporting viewpoint management), requires recognising a third dimension. I will begin with a single case that presents a problem for the two-dimensional model, and show how the addition of a separate dimension relating the present communicative situation to other ones, provides a straightforward solution. Subsequently, I will show that this new model also provides a very natural framework for the analysis of other items and viewpoint configurations.

5.3.2 Speaker commitment and viewpoint embedding: Dalabon and English

Consider the following utterance in the Australian language Dalabon and its English translation as suggested by Nicholas Evans (class lectures 2009, brackets in original):[81]

[80] Objectivist approaches to language thus in fact also assume a one-dimensional conceptual framework for the analysis of meaning in natural language, but highlight the *other* dimension (i.e. the *(y)*-axis) instead of the dimension I have suggested to be relevant for non-human communication (the *(x)*-axis). Verhagen (2005) extended Langacker's model by including a systematic distinction between the viewpoints of the sender and addressee, in order to bring out the fact that construal is not (just) a matter of a single viewpoint ("subjectivity") with respect to some object, but one of mental coordination between signaller and addressee with respect to an object of joint attention ("intersubjectivity"). In hindsight, we can say that the framework proposed by Verhagen (2005: 7) represents a merger of the biological and cognitive-semantic views of human communication.

[81] I thank Nicholas Evans for permission to use this example in this context. See Evans (2010) for more examples of elements for viewpoint coordination (esp. ch. 4), and for the glossing method used.

(1)	Ka-h-kangurdinjirrmi-nj	yangdjehneng	bûrra-h-marnû-dulu-djirdm-ey
	3SG-ASS-GET.ANGRY-PSTPERF	SUSPENDED-COMMITMENT	3DUHARM.SUBJ>3SGOBJ-ASS-BEN-SONG-STEAL-PSTPERF
	He got upset [because]	*[he thought that]*	*the two of them had stolen his song*

(2) He got upset [because] [he thought that] the two of them had stolen his song.

"Because" and "he thought that" are inserted in the paraphrase by Evans. I will briefly discuss the causal marker "because" in note 82, but focus on the insertion of "he thought that" in detail first. It is clear that the lexical unit *yangdjehneng*, glossed as "SuspendedCommitment" does not literally mean "he thought that", but rather conveys the message: "I, speaker, am not committed (to what I am going to say now)". A paraphrase closer to the original expression is thus:

(3) He got upset [because] [I, speaker, am not committed to this:] the two of them had stolen his song.

At first sight, it may seem remarkable that Evans renders the lexical unit that functions as a marker of suspended commitment with a complementation construction in English—are the two indeed equivalents? To illustrate that, in an important sense, they are, consider the differences between the more idiomatic translation in (2) and the more literal one in (3). The absence of "he thought that" in (3) does not mean that "he" no longer had the thought that "the two of them had stolen his song". In fact, awareness of the information in the second clause is equally implied in (2) and (3); if "he" had not had that thought, the stealing of the song could not have caused him to be upset. In both the Dalabon and English versions the speaker invites the addressee to view the information about "the two of them" having stolen the song from the perspective of a third party, namely "he" introduced at the beginning of the sentence. However, there are differences in the degree to which this is accentuated: the coordination of a third-party perspective is significantly more

pronounced in the idiomatic English translation in (2) compared to the Dalabon original (1) and its paraphrase in (3).

In a similar sense, the Dalabon element *yangdjehneng*, the English phrase "I, speaker, am not committed to this", and the idiomatic pattern of sentence complementation using the stance verb *to think* all three negotiate a(n epistemic) stance of the speaker towards parts of the presented content: in (1), (2), and (3), the speaker does not assert as true that "they had stolen the song". Yet the difference is again in the accentuation: in (1) and (3) the tempering of commitment by the speaker is realised "on stage", whereas in (2) this remains implicit.

In both the Dalabon and English versions the speaker thus invites the addressee to view the information about the two of them having stolen the song from the perspective of a third party, namely "he" introduced at the beginning of the sentence. However, there are differences in the degree to which this is *accentuated* or *profiled*: the embedding of the information in a third-party perspective is done explicitly by means of a particular syntactic construction (complementation) and a particular matrix predicate (*thought*) in English, with the speaker's reduced commitment remaining more implicit, while the latter is precisely being profiled by the Dalabon element *yangdjehneng*, with the third party's relatively higher degree of responsibility remaining more implicit. In short, these conventional ways, in these two languages, of distributing responsibility for a piece of information over the speaker and another party are each other's mirror image: what is explicitly "put on stage" and "what is left to inference" is so to speak reversed. But the totality of what is communicated with

these structurally very different expressions, is very much the same, in particular the connections between different relevant viewpoints.[82]

Can both the similarities and the differences between these expressions be stated in a single analytic framework? If we try to do so using the two-dimensional model of triadic communication in Figure 3, it soon becomes clear that this requirement cannot easily be satisfied. If the function of the element *yangdjehneng* ("I am not committed to this") is straightforwardly characterised as the speaker signalling to the addressee (*(x)*-axis) what his stance is towards (*(y)*-axis) the object of conceptualisation (i.e. "the two of them had stolen his song"), the associated heightened responsibility for this view of the third party ("he", the one who got upset), is necessarily left out. The reason is that this third party is only present in this model as an element of the situation being talked about, as an *object* of conceptualisation, and not as another *subject* taking a view on this situation.

Conversely, the representation of the complementation construction in (2) ("he thought that...") does not work very well in a two-dimensional framework either, conceptually. In (2) "he" is clearly not only an object of conceptualisation (we are presented with *what he thought*), but at the same time, "he" should not be seen as a subject of conceptualisation in the same sense as the speaker and addressee either. After all, the *negotiation* of a stance towards the object of conceptualisation takes place between speaker and addressee, meaning that the speaker can, as a part of this negotiation process, *invite* the addressee to consider the perspective of a third party on some aspect of the object of conceptualisation. However, perspective can never shift completely to this third party in the course of the modelled interaction event (cf. the way it can jump

[82] The phenomenon of marking explicitly only some aspects of what is to be conveyed is, of course, not limited to viewpoint expressions; on the contrary, it is quite general and well-documented for various conceptual relationships, including causality (see e.g. Verhagen, 2005). It should be noted that an analysis similar to the one given above applies to the pair (i) *He got upset; the two of them had stolen his song* and (ii) *He got upset because the two of them had stolen his song*. The conceptual representation of both (i) and (ii) contains a causal relationship (otherwise no coherent interpretation seems possible), but this is only marked explicitly, "on stage", in (ii). The difference between the Dalabon and English idiomatic ways of expressing both viewpoints and causal relations can be characterized as a difference in the available tools, and in the conventional rules for using them in the different languages. See also Wilkins' (1986) discussion of "particle/clitics" for criticism and complaints in Aranda, another Australian language, and his argument that these encapsulate "culture specific modes of thinking" that become clear when their use is explicated.

from one character to another in a novel). In other words, the view of the third party "he" can be instrumental in the speaker's and addressee's negotiation of a stance towards the object of conceptualisation, but "he" is himself not a participant in this negotiation process. All in all, the common problem when representing the sentences (1), (2), and (3) seems to be that in a two-dimensional conceptual model of communicative interaction, third-person conceptualisers can only be situated either at the level of the object of conceptualisation, or at that of the speaker and addressee, while in fact they normally belong to neither.

I therefore propose to treat other subjects of conceptualisation not as additional entities in the two-dimensional space, but as implying the addition of a third dimension, which links third parties exhibiting intentionality towards the relevant object of conceptualisation to the level of the negotiation process between speaker and addressee. The basic idea is captured in Figure 4:

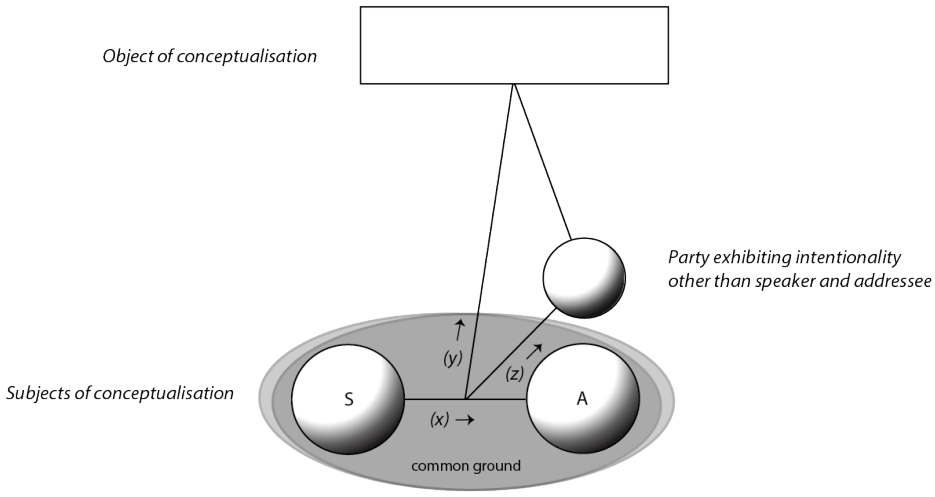

Figure 4 – The three-dimensional conceptual model of interaction featuring a non-speaker, non-addressee subject of conceptualisation.

We conceive of the third person represented in Figure 4 as a subject of conceptualisation in exactly the same way (i.e. with the same cognitive capabilities, including intentional reasoning) as the speaker and addressee. Moreover, the object of conceptualisation for this subject is (at least in part) the same as the one that the speaker is inviting the addressee to consider, capturing

the idea that the speaker presents the situation to the addressee *from a third-party perspective*.

This basic model provides the conceptual space to mark precisely the similarities and differences between the Dalabon and English viewpoint items discussed above, regardless of the fact that they belong to completely different language systems. A graphic representation of these forms can be found in Figure 5 and Figure 6, respectively.

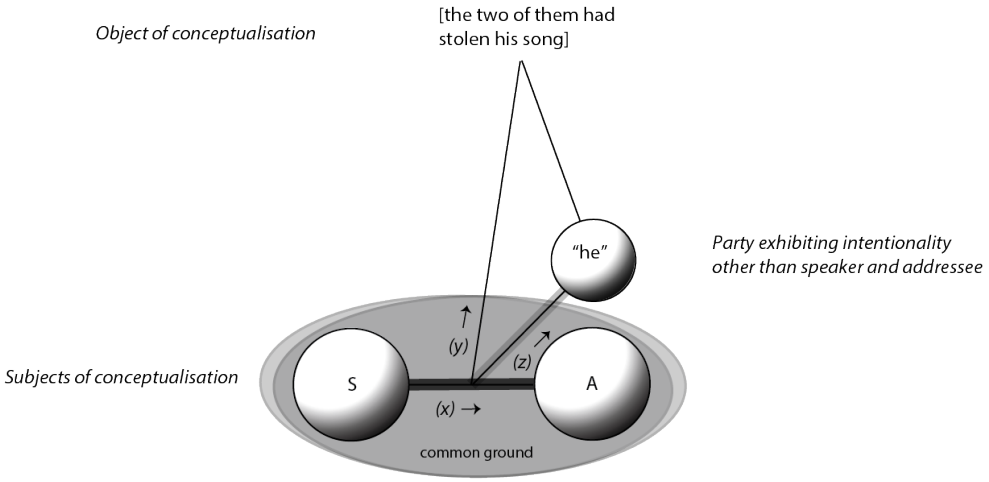

Figure 5 – Dalabon: *yangdjehneng*

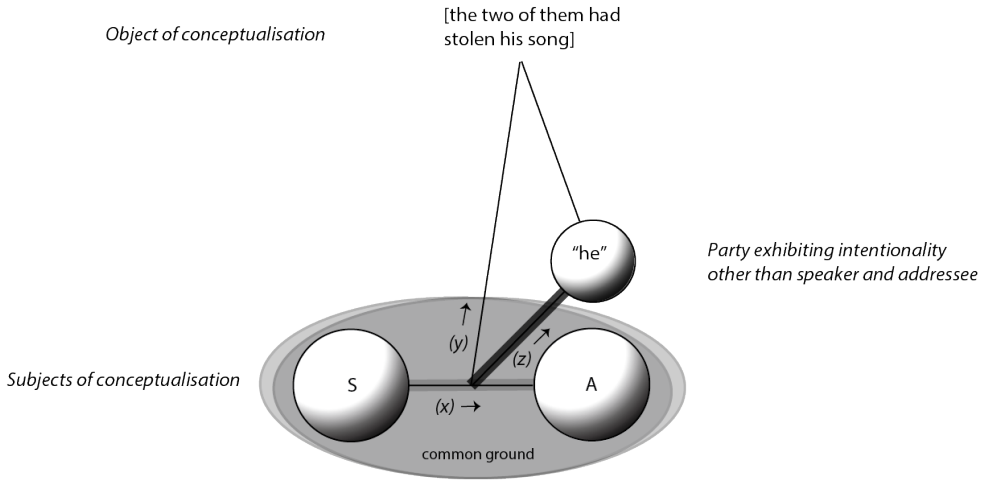

Figure 6 – English: *to think* + complement

As in the two-dimensional model, the negotiation of epistemic stance performed by both the Dalabon and English elements is highlighted along the *(x)*-axis between S(peaker) and A(dressee). In Figure 5, this axis is marked with a dark line, indicating foregrounding of the speaker's epistemic stance by the Dalabon element *yangdjehneng*, glossed as "I, speaker, am not committed". In Figure 6 this axis is marked with a light grey line, indicating that the English complementation construction with *to think* does impact upon the negotiation of epistemic stance between S and A, but in a less pronounced way than the Dalabon element does.

What is new in Figure 4, 5, and 6 compared to the two-dimensional version in Figure 3 is the *(z)*-axis connecting the *(x)*-axis to a third party, in this case the person referred to using "he" and "his" in Evans' translation in (2). On this axis, the reverse pattern obtains with regard to profiling: whereas in Dalabon this third-person perspective is only implied, indicated by a light grey line along the *(z)*-axis in Figure 5, in English it is explicitly realised "on stage", indicated by a dark grey line on the *(z)*-axis in Figure 6. Thus, thanks to the additional *(z)*-axis, we now have a single format for representing that both the Dalabon and the English versions of the utterance invite the addressee to consider the third person's perspective on the matter talked about, i.e. the (actual or imagined) stealing of the song by "the two of them", and that they do so in different ways, by highlighting what parts of the configuration are linguistically marked in each language, and which are implicit, but made inferable.

5.3.3. The general model

When I first introduced the three-dimensional model, I stated (below Figure 4) that the additional intentional party is a subject of conceptualisation whose perspective is instrumental in the speaker's and addressee's negotiation of how the common ground should be updated with respect to an object of joint attention, without himself being a participant in this negotiation process. However, this third party may himself be represented by the speaker as being involved in another communicative interaction event, and in fact, this party

may be talking or thinking about yet another interaction event. Thus, we may in principle expect to encounter more elaborate constellations of several subjects all in some way considering the same object of conceptualisation from different viewpoints, and affecting (more and less mediated through the viewpoints of others) the negotiation between S and A of epistemic stance, attitude, etcetera. Such a constellation is depicted in Figure 7:

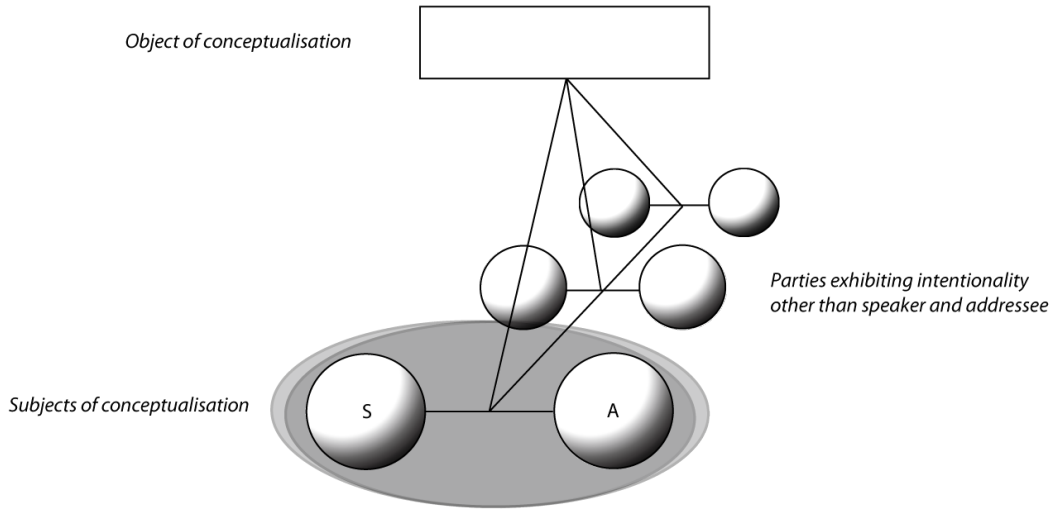

Figure 7 – While communicating about some object of joint attention, S and A may refer to other interaction events, each featuring *their* participants.

In the situation depicted in Figure 7, viewpoints from the other interaction events must, in one way or another, be relevant to how S and A assess their object of conceptualisation. As an example, imagine two people, Simon and Arran, waiting for a man named John to show up at their appointment. Simon has seen John the day before and when the appointment was mentioned, John's daughter Mary kindly reminded her father that he is *always* late. Now Simon says to Arran that "John assured Mary that he would be on time". Figure 8 depicts this situation schematically:

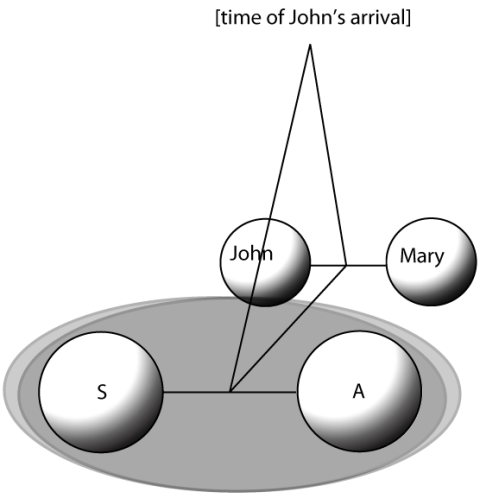

Figure 8 – "John assured Mary that he would be on time"

In this example, there is thus another interaction event being called up within the current interaction: Simon coordinates his perspective on John's expected time of arrival with Arran by referring to how John was coordinating *his* perspective on his arrival time with Mary the day before. He could have done this in numerous alternative ways, for example by saying "John said to Mary: 'I'll be on time'", "When I saw John and Mary, John thought he would be on time", "John will be on time. He promised Mary", and so on. All these alternatives feature a slightly different distribution of responsibility for what John said and the amount of commitment made by Simon to John being on time, given that some elements modify the nature of the relationship between third-person conceptualisers and the speaker and addressee in different ways. Thus, in this example, the use of indirect discourse and the choice, by the speaker, of the verb "assure" (unlikely to have been used by John himself), indicate some degree of co-responsibility of (and interpretation of John's utterance by) the speaker, higher than with the use of a neutral verb of speaking and direct discourse (e.g. "John said to Mary: 'I'll be on time'"). These differences are as such interesting from a semantic, grammatical, or narratological perspective, but go beyond the point I want to make here—which is that all alternatives feature different linguistic elements (words,

grammatical constructions, patterns of speech and thought representation) with different meanings, leading to a variety of overall interpretations, by operating on parameters within the same conceptual space: the relationship between the speaker and addressee (*(x)*-axis), the relationship of the communicative interaction with other interactions featuring third-parties (*(z)*-axis), and all of their perspectives on the Object of conceptualisation (*(y)*-axis).

5.3.3 Thoughtscapes and the model

Some of the examples provided in Chapter 4 were drawn from news sources reporting on the "Pistorius case", the tragic shooting of Reeva Steenkamp by athlete Oscar Pistorius. Recall that the difference between the competing versions of what happened during the night of the shooting completely depended on the construal of Pistorius' intentional state at the moment of pulling the trigger: did he *think* he was shooting at a burglar or did he *know* his girlfriend was behind the bathroom door? The news media not only reported the perspective of the athlete, but also of police detectives, spokespeople, journalists, witnesses, family members, and so on. The result was what I termed a "thoughtscape", a series of perspectives that are mutually connected and embedded in various ways. What could be found in the news reports was what I referred to as "polyphonic" discourse representing this thoughtscape: all kinds of linguistic elements were doing some part of the labour of coordinating the involved perspectives, including grammatical constructions (such as complementation and inquit-constructions), various patterns of reported speech and thought, lexical items (such as *allegedly* and *accidentally*), tense, modality, and more. One of the examples was the following opening sentence from a South-African press release:

> (4) Athlete Oscar Pistorius **allegedly accidentally** shot dead his girlfriend at his house in Pretoria on Thursday morning, *Beeld.com* reported.
> (SAPA, 'Oscar Pistorius shoots girlfriend: report', 14 February 2013)

As a whole, (4) fits a particular embedding pattern, termed an inquit-construction in Chapter 4, in which the reported clause precedes the reporting

clause (underlined). The inquit-construction does part of the viewpoint coordination: it attributes the claim that "Pistorius allegedly accidentally shot dead his girlfriend" to the perspective of newspaper *Beeld.com*. However, there are more viewpoints being coordinated. It is implied by the adverbs *allegedly* and *accidentally* (boldface) that some external source *claims* that Pistorius *did not intend* to shoot his girlfriend. In other words, already on the basis of one sentence, readers are confronted with a thoughtscape involving three viewpoints, without even counting the perspective of the speaker (i.e. the journalist who wrote the sentence).

I will first abstract from the reporting clause of the inquit-construction and concentrate on the reported content:

(5) Pistorius allegedly accidentally shot dead his girlfriend.

The words *allegedly* and *accidentally* are instantiations of what was in Chapter 4 described as viewpoint packages, words implying a topology that introduces one or several extra viewpoint layers. In the case of *accidentally*, it is given in this topology that an agent did not intend X, but it is known that the outcome is X. In actual usage this topology is assimilated (through *blending*; see Chapter 3) with details provided in the immediate context. For example, readers of (5) will blend their knowledge of the topology of *accidentally* with "Athlete Oscar Pistorius" and "shot dead his girlfriend", and take it that he shot her dead, but *did not intend* to do so. In this way, using *accidentally* the speaker invites his addressee to consider the perspective of a third party, in this case Pistorius. Since this is not highlighted explicitly, in the depiction below a light grey line is used along the *(z)*-axis:[83]

[83] The word *accidentally* clearly also negotiates a relationship to an object in the world on the *(y)*-axis, but in my discussion here I will abstract from these relationships and focus on those indicated on the *(x)*- and *(z)*-axes.

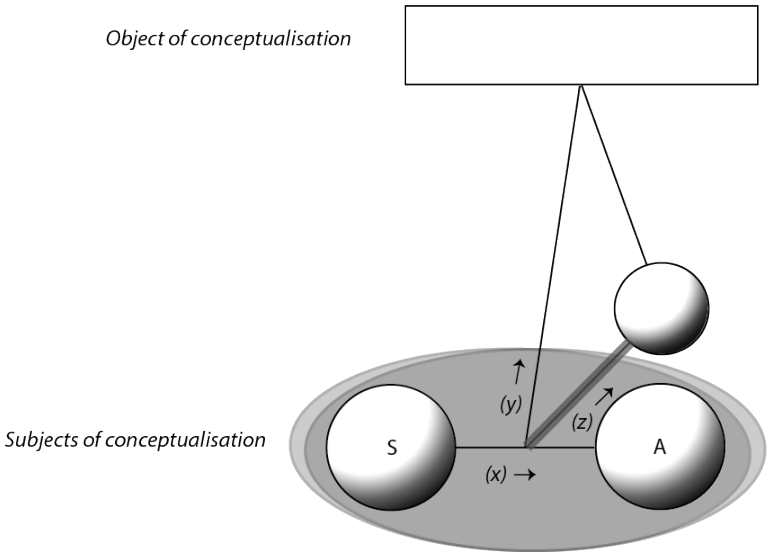

Figure 9 – accidentally

In a similar way, in the topology of *allegedly* it is given that some source X, not the speaker, asserts the content under the scope of this adverb. This topology can be elaborated to various degrees. The identity of source X can be given in the context, or left open, as is the case in (4): readers of this sentence will understand that some external source not specified here claims that Pistorius accidentally shot dead his girlfriend.[84] In that sense *allegedly* shows strong similarities to the Dalabon element *yangdjehneng* cited in (1) above. It suggests the presence of an extra viewpoint, lowers the epistemic commitment the speaker makes to the related content, and, indeed, could also be "translated" using a complementation construction:

(6) It is claimed that Pistorius accidentally shot dead his girlfriend.

[84] Recall that in Chapter 4, Section 4.3.3, an alternative reading of (5) is discussed next to the one given here. However, distinguishing between these two options is not relevant here.

In terms of the present model, *allegedly* is thus a linguistic cue that negotiates a particular epistemic stance of the speaker, while at the same time inviting the addressee to consider the perspective of a third, in this case unspecified, party. It operates along the *(x)*- and *(z)*-axes, albeit without a particular emphasis on either. Consider the schematic depiction in Figure 10:

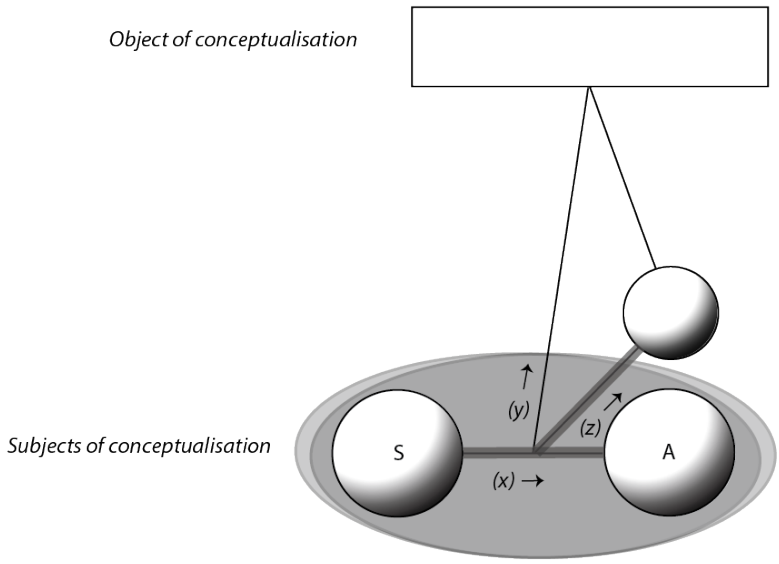

Figure 10 – *allegedly*

A schematic rendering of (5), involving at the same time the viewpoint coordination effected by *accidentally* (i.e. the athlete *not intending* to shoot his girlfriend), is also possible in the proposed conceptual space. This involves the inclusion of one more viewpoint along the *(z)*-axis, which can be done as follows:

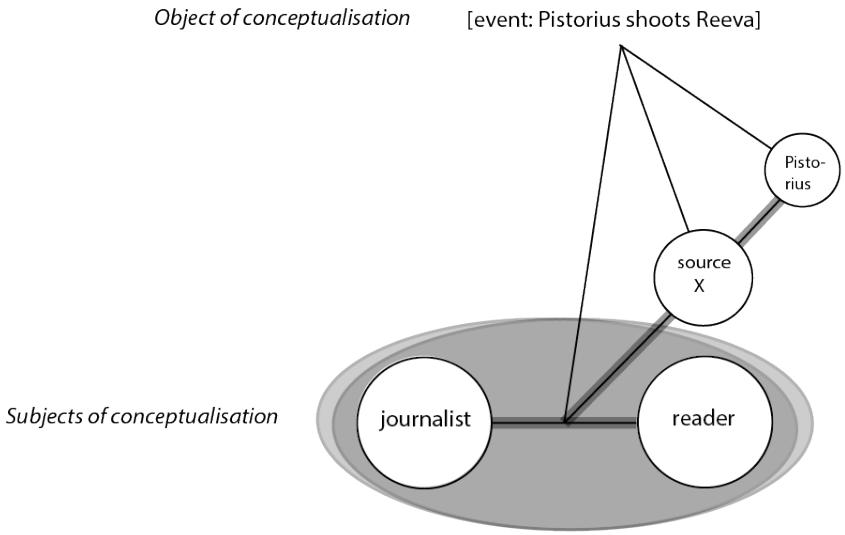

Figure 11 – Sentence (5)

Finally, the perspective of *Beeld.com*, which is coordinated with respect to the reported content using the inquit-construction in (4), can be added to the picture. Given that the introduction of the perspective of *Beeld.com* takes place explicitly, "on stage", a dark grey line is used here along the *(z)*-axis. The introduction of the two additional perspectives (Source X and Pistorius) as well as the negotiation of epistemic commitment is done implicitly, "off stage", hence the light grey lines:

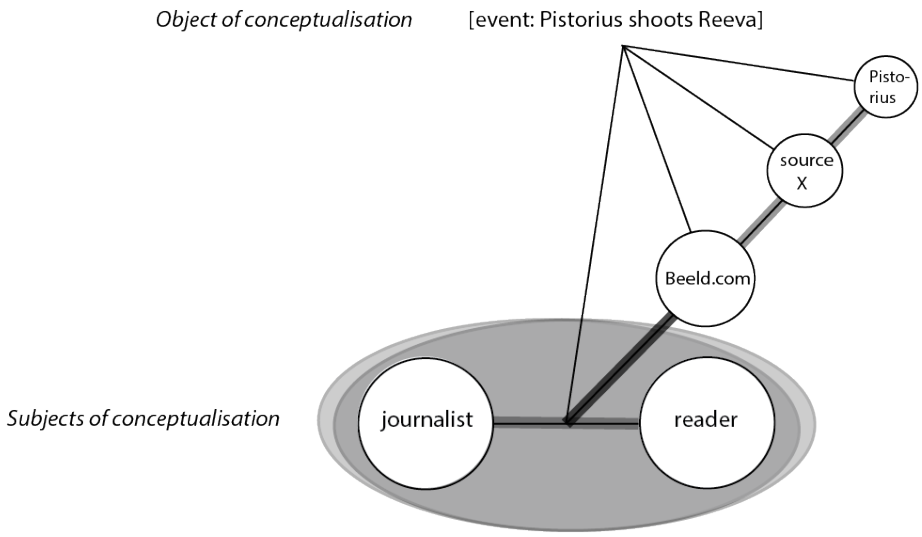

Figure 12 – Sentence (4)

All in all, using the conceptual framework suggested here, and depicted schematically in Figure 4-12, I argue that linguistic elements across different categories, levels of analysis, and languages (lexical units, grammatical and narratological patterns, English, Dalabon) operate along three dimensions: speaker and addressee negotiate ((*x*)-axis) how the common ground should be updated with respect to a particular object of conceptualisation ((*y*)-axis), potentially by inviting the other to view this object of conceptualisation (in part) from the perspective of third parties ((*z*)-axis). In the case of some interaction events this process of updating the common ground involves no third-party perspectives at all (to those interaction events only the first two dimensions are relevant), whereas in other cases a complex thoughtscape is conjured up in the course of this process. Sometimes, the perspectives in these thoughtscapes are embedded into one another (cf. Figure 11 and 12) and sometimes they are related in different ways. For example, third parties can themselves be represented as being interlocutors in a different interaction event (cf. Figure 8), or their viewpoints can form meaningful conjunctions or exhibit causal relations from the perspective of the speaker and addressee. These latter two options have not been explored using examples in this chapter (however, Chapter 6 will feature

several examples). In conclusion, it is worth noting that the suggested conceptual space can accommodate such thoughtscapes comprising conjunct, causally related, or otherwise linked third-party perspectives:

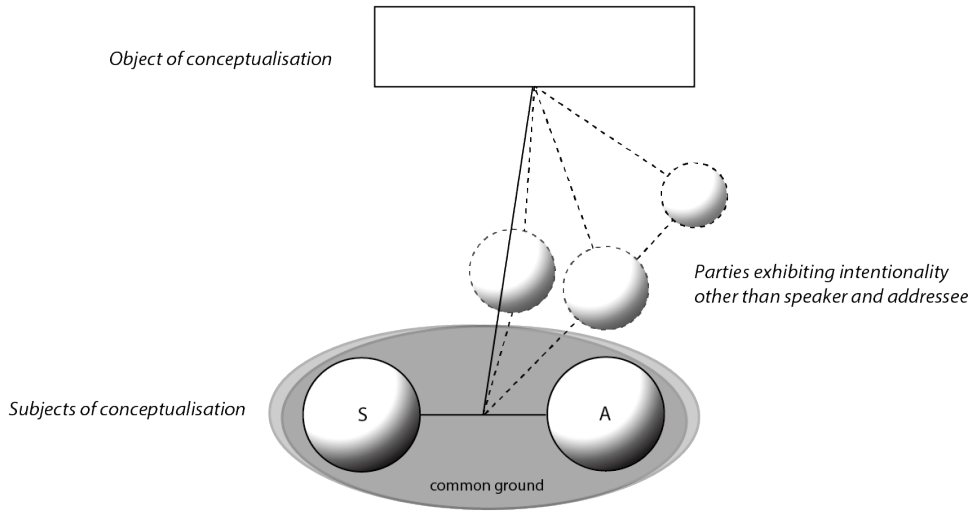

Figure 13 – In the course of some interaction events, a thoughtscape is conjured up taking the form of a network of perspectives related in various kinds of ways. This has been discussion in more detail in Chapter 3 and will be further discussed in Chapter 6. The dashed lines and circles suggest a conjunction between one single perspective and two perspectives exhibiting a form of embedding (e.g. "John believes that X while Mary thinks that Peter doesn't want that X").

5.3.4 Updating the common ground

In this final subsection I will introduce the view of viewpoint coordination as a matter of highlighting, negotiating, and anticipating how individual perspectives *deviate* from the common ground (see again also Clark, 1996). This view will be important throughout the next chapter and be built on in the Conclusion. As one last example, consider the following excerpt of a recorded conversation from the Corpus of Spoken Dutch (CGN), followed by my English translation:

(7) A: oh dan is vandaag Allerheiligen.
B: 't is vandaag Allerheiligen ja. [...]
A: oh oh dan heeft Ella zich denk ik vergist.
want ze dacht dat morgen Allerheiligen was en dan waren de winkels beperkt open
B: ja.
A: nee dat is dat is uh dat is vandaag.

A: oh in that case today is All Saints.
B: today is All Saints indeed. [...]
A: oh oh then I think Ella was mistaken.
because she thought that tomorrow was All Saints and then the shops were only open for a limited period of time
B: yes.
A: no that is that is uh that is today.

Interlocutor A finds out that Ella, a third-party subject not present in the current interaction event, falsely believed that All Saint's Day was tomorrow. The underlying assumption is that knowledge of when this is, is part of the common ground within the cultural-linguistic community of which A, B, and Ella are apparently members. Interlocutor A first opposes Ella's false belief-state to the common ground using the viewpoint package "mistaken" (i.e. *holistically* in terms of Chapter 4), and then further elaborates using the complementation construction "she thought that tomorrow was All Saints" (*compositionally* in terms of Chapter 4). Just as in the example given above of my office mate and me having a misunderstanding (note the viewpoint package!) over whether pointing was directed at the coffee machine or the window blinds, the working out of different perspectives enters the stage in order to figure out a *deviation* from the common ground.

In fact, this is not different with the Pistorius case: the entire thoughtscape hinges on the fact that there are two competing versions of the story (i.e. a crucial discrepancy in common ground) distributed over various third-party subjects. What the interaction as depicted in Figure 11 boils down to is a

journalist negotiating the exact nature of this difference in front of a reader. In Figure 8, depicting the situation of Simon saying to Arran "John assured Mary he would be on time", we see how Simon singles out John's perspective, which is contrasted to both interlocutors' (and Mary's) expectation that he will be late. Recall also once again the case of Shakespeare's *Othello* discussed in detail in Chapter 2: the plot of this play combines multiple scenarios (revenge, a scheming plan, suspected adultery) that imply crucial knowledge differences between the involved parties, inducing a long sequence of negotiations about how various character mindstates deviate from a common ground. I will say a few more words about this point in the Conclusion, after it has been applied in the context of testing intentional-reasoning competence experimentally in Chapter 6.

5.4 Discussion and concluding remarks

Verhagen (2005: 4) argues that "mental coordination" is an essential part of linguistic interaction, and therefore it is to be expected that languages have developed, over the course of their history, special conventionalised signals to support this function, in line with Du Bois' (1985) claim that grammars code best what speakers do most. Verhagen (2005) focuses specifically on words and constructions (besides gestures, facial expressions, and other meaningful elements) which support mental coordination between speaker and addressee, but here I have cast the net wider and also included the marking and coordination of the mindstates of third parties, who may or may not be present at the time of speaking, or who may even exist only in the imagined worlds of thought and fiction. What I have argued is that linguistic items capable of viewpoint coordination serve to highlight and negotiate how individual perspectives *deviate* from the common ground. This reflects an important characteristic of human interaction: instead of starting from individual intentional systems that seek to become "paired", the default is that interlocutors take part in a system of shared intentionality or common ground and negotiate how individual perspectives relate to this.

An important remaining question is how all of this affects our evolutionary story. As pointed out in Chapter 1, it is generally assumed that our ancestors had to reach a certain threshold of intentional reasoning capacity before communication "as we know it" could begin. Indeed, according to Sperber (2000) and Scott-Phillips (2015), the capacity to reason at five orders of intentionality had to predate "proper" ostensive-inferential communication. After all, individuals had to mutually recognise communicative and informative intentions, understanding that the other intends one to see that the other intends one to understand that something is the case. However, the view advocated in this chapter allows for an alternative: I suggest that our ancestors in some way first started to establish forms of common ground, and then developed increasingly sophisticated ways of singling out individual perspectives and ways in which they differed.[85] This process is presently reflected in all kinds of linguistic items being capable of highlighting and negotiating how the perspectives of signallers, addressees, and third-party subjects relate to and, indeed, deviate from the common ground.

Finally, note that this, on an abstract level, is a similar kind of theoretical "move" as the one made by Shultz et al. (2011) regarding early primate social life. Their evidence seems to support a scenario in which individuals *first* started living (c.q. foraging) in groups, and *then* developed increasingly profound dyadic bonds and relationships. In the Conclusion I will integrate this point in Dunbar's framework as set our in Chapter 1. However, before getting there I will apply the developed views to the practice of assessing multiple-order intentionality experimentally in the next chapter.

[85] Note that this is much in the fashion of what Moll and Tomasello (2007) term the "Vygotskian intelligence hypothesis" (cf. Vygotsky, 1978).