On common knowledge in conversation

Piotr Labenz

20th May 2004

Abstract

This working paper aims towards a coherent account of the sort of common knowledge that is necessary for making felicitous utterances in a conversation. The hypothesis to that effect, originating from Lewis [29] and Schiffer [37] is presented in Section 1 along with the standard definition of common knowledge. However, as Clark and Marshall [10] observed, that definition leads to a paradox; in Section 2 I discuss the possibilities of avoiding it. I argue (contra Clark) in favour of a fixed-point approach a la Barwise [7], which, relies on the notion of coordination devices. Since the argument about the common knowledge paradox resorts to the cognitive plausibility of the notion, in Section 3 I try to clarify its empirical status by considering some relevant data.

This being done, the outlook of this paper is sketched in Section 4, being a genetic account of coordination devices in terms of some simple games. Its purpose is to supplement the fixed-point approach to common knowledge in parallel with Clark's [9] psychological account. This concludes with an attempt at a cognitively plausible definition of common knowledge.

It has to be stressed that as it stands, this paper is an informal, working sketch meant as a basis for further work.

1 Introduction. The hypothesis

Having a conversation, we quite ordinarily assume our interlocutor to share some information with us. Hadn't we such an assumption, it would be reasonable to be entirely and thoroughly explicit in our utterances. But not only would this be bizarre, but impossible as well. For instance, in saying:

(1) The only recipe I have to avoid fatigue is not to do too much work.

I assume that my interlocutor knows what work, fatigue and a recipe are, what constitutes doing work, having a recipe and so forth. It would hardly be possible for me to be more explicit; indeed, any attempt to would be ridiculous (and so infelicitous). The interlocutor, in her turn, must know that I know what a recipe is etc., in order to be sure that I meant what (1) means rather than something else. After all, had I not known the correct meanings of the words in (1), I might have mistakenly used them to express, for instance

(2) I like kids, but I don't think I could eat a whole one.

or something even more preposterous. Thirdly, I must know that she knows that I know what a paper is etc., because otherwise I would not be sure that in having uttered (1) I had conveyed the information I had intended to. Indeed, has she not known the meanings of words constituting (1), uttering it would not be felicitous. But, by the same token, she must know that I knew – and so on ad infinitum. So common knowledge of a language's meaning postulates is a prerequisite of using that language[1].

Indeed, this phenomenon is even more conspicuous in the case of definite reference, be it by deixis, by anaphora or by proper name. Taking an instance of the first:

(3) I've met that blonde before.

presupposes some proposition identifying the individual referred to like, taking the expression on the left-hand side to stand for a token expression (uttered as a part of (3) by a certain agent at certain time), and that on the right-hand side to be a name uniquely referring to that blonde:

(4) that blonde = Lydia.

Had it not been for these presuppositions, the reference of (3) would not be (uniquely) determined. In uttering these sentences I assume that my interlocutor identifies what is referred to – namely, Lydia – the same way as I do. Furthermore, as Clark and Marshall [10] noticed, these presuppositions must be common knowledge. To utter (3) felicitously, I must not only presuppose (4), but know that my interlocutor presupposes it as well – thus identifying that blonde the same way I do. To understand (3) the way I meant, she must know that I know that she presupposes (4). And I must know that she

[1] Already Ajdukiewicz [1] observed that these postulates must be known. That common knowledge is required has been first noted by Lewis [29]. About common knowledge and presuppositions cf. Stalnaker [40].

will understand it thus, hence I must know..., etc. If I was unsure about any level of this regress, then I could not be sure that my utterance will be felicitous.[2]

Therefore it has been claimed, first by Lewis [29] and Schiffer [37] that common knowledge is a prerequisite of linguistic communication, because it is indispensable for making felicitous utterances. Common knowledge can be defined thus:

Definition 1 (Iterated common knowledge). [3] A proposition p is common knowledge in a set C of agents iff
$\forall x \in C \, \Box_x p$ and
$\forall x \in C \, \forall y \in C \, \Box_x \Box_y \, p$ and
$\forall x \in C \, \forall y \in C \, \forall z \in C \, \Box_x \Box_y \Box_z \, p$ and
etc. ad infinitum.

Then if a proposition q presupposes propositions $p_1, p_2 \ldots p_n$, these must be common knowledge in a set of agents containing the speaker and the hearer in order for p to be uttered felicitously. This view, which may be called "common knowledge hypothesis" (CKH), is the received view[4].

Of course, it is somewhat unsettling to postulate an infinite series of knowledge attributions in order to explain how is it possible to utter sentences felicitously; after all it is not *that* difficult. Thus some have suggested that the maximum required knowledge-operator depth be reduced from infinite to three or four[5]. While doubtless more plausible psychologically, this would, however, be logically inadmissible. The consecutive steps in the infinite regress yielded by definition 1 form, as Lewis [29, p. 53] remarks, a chain of implications. Cutting it at any point would, by a sort of domino effect, invalidate the initial elements as well: for instance, should depth n fail to obtain of (4), then so would depth $n-1$, etc. Moreover, it is possible to contrive examples showing that common knowledge of any arbitrary depth can be explicitly required for the felicity of some utterances; cf. [37], [10], [41].

[2] Cf. also[19], [9].

[3] This definition was first suggested by Schiffer [37] and is standard in epistemic logic and AI; cf. [14], [31]. On definitions of common knowledge generally, see [7], [14], [42].

[4] Even though there is a considerable amount of conceptual confusion: "common", "mutual" and "shared"; "knowledge", "belief" and "ground". See [28], [26, p. 29], [9, p. 99].

[5] For references, see [9, p. 100], [10]. A major attempt to dispense with the notion of common knowledge altogether was Sperber and Wilson's relevance theory [38], which, however, seems to have some drawbacks – in particular to be rather underdeveloped on the formal side.

2 The paradox and the definitions

Clark and Marshall [10] argue that Definition 1 leads to a paradox. Namely, even if C is finite, operators for its every member can be iterated along the formula, for instance thus: $\Box_x\Box_y\Box_x\Box_y\Box_x\Box_y\ldots$, so that the definiens is an infinite conjunction. Hence to check whether p is common knowledge, one has to check whether each of the infinite number of conjuncts is true. Now, on the computational conception of the mind, checking whether a sentence is true takes some time – perhaps small, yet nonzero. Therefore to learn whether p is common knowledge takes an infinite time. But, by CKH, in order to make felicitous utterances, we need to know that some presuppositions are common knowledge. Yet we make felicitous utterances in finite time (notwithstanding that one admittedly might take quite some time pondering sentences like (3)). A contradiction.

To begin with, note that Definition 1 can be equivalently rephrased in a semantic manner:

Definition 2 (Iterated common knowledge). Let R^* be the ancestral closure of the accessibility relations of all agents in C. Then p is common knowledge in C at the world w (assume w is the actual world) iff for all possible worlds v, $R^*(w,v) \rightarrow v \models p$.

The problem on which the paradox hinges is of syntactic nature; it could be easily alleviated if there was a finite method of checking whether p is common knowledge (CK-checking for brevity) semantically. But there is not. Namely, by Definition 2, p must either be true at all worlds in the model, or false only at such 'solitary' worlds that no agent sees them from any other world. Then, trivially, for every agent x, $\Box_x p$ is true at each world except solitary worlds and, by the same token, $\Box_x\Box_y p$ etc., so eventually p is common knowledge in C. Thus for CK-checking one must check that p is true at all non-solitary worlds. But unless the model is finite, this requires an infinite number of steps (note, however, that to check that something is not common knowledge, always a finite number of steps suffices). If the model is finite and p true at all non-solitary worlds, and the model doesn't split into submodels such that no world belonging to one submodel can be seen from the other and conversely, then p is common knowledge. These conditions are finitely verifiable; therefore one way to avoid the paradox is to assume the iterated definition limited to finite models.

Whether the restriction to finite models is a plausible one is philosophically debatable. On the one hand, the physical world seems to be finite; on the other, there are infinitely many numbers, counterfactual possibilities

and so on. However, regardless of that there are other reasons to reject the iterate approach. Firstly, even restricted to finite models it is cognitively untenable, because CK-checking on the above lines requires processing an immense search space. It seems most unlikely that processing every utterance we check whether each of its (numerous) presuppositions is true at each of (very numerous) non-solitary possible worlds. Secondly, as McCarthy et al. [30] observed, the iterate account does not entail that common knowledge of p should always be common knowledge itself. That is a drawback insofar it seems reasonable to expect a sort of group introspection: it should be common knowledge among the members of C what is the common knowledge they share.

Therefore an alternative definition of common knowledge would be most desirable. One alternative is to explicitly mention shared basis, the set of beliefs that give rise to the common knowledge in a given group:

Definition 3 (Shared-basis common knowledge). [6] The proposition p is common knowledge among the agents in C iff there is a basis B such that $\forall x \in C\ \forall p \in B\ \Box_x p$, where

$$B \models (p \wedge \forall x \in C\ \forall p \in B\ \Box_x p).$$

This definition nicely corresponds with Clark's psychological principle of justification: "in practice, people take a proposition to be common ground in community only when they believe they have a proper shared basis for the proposition in that community" [9, p.96]. Clark also claims that it avoids the paradoxical regress.

Or does it? Let $\Box_x p \in B$; then by Definition 3 for any $y \in C$, $\Box_x \Box_y p \in B$. One can iterate this step obtaining in B box strings of any length: $\Box_x \Box_y \Box_x \Box_y \Box_x \Box_y \ldots$, just like on the iterate approach. Belonging to B, propositions starting with these strings must be known to all members of C; therefore syntactically there is the same sort of regress as previously. However, semantically Definition 3 does much better: it suffices to supply an example of shared basis B, which is doable in a single step, and common knowledge follows. However, as Barwise [7] remarks, there can be many different bases B that yield the same common knowledge (are "informationally equivalent"). Indeed, there may be "unintended" ones that we would not think of, that do not represent any psychologically plausible shared basis. And there may be infinitely many such unintended bases; loosely speaking, the shared-basis approach is not categorical. Therefore one may check in finite time that p is common knowledge, but not that it is not common

[6] First proposed by Lewis [29] and Aumann [4].

knowledge, because one cannot rule out that there is some unintended B, on which p would be common knowledge. That seems to be unacceptable, since on CKH we should be able to rule infelicitous utterances out by discovering that their presuppositions are not common knowledge in finite time. Therefore in fact Definition 3 does not help to avoid the paradox.

Fortunately enough, there is another alternative, namely the fixed-point, or reflexive, approach.

Definition 4 (Fixed-point common knowledge). [7] A proposition p is common knowledge among the agents in C iff some q holds such that

$$q \rightarrow \forall x \in C \, \Box_x (p \wedge q).$$

Here, q is a non-well-founded proposition that is not itself part of common knowledge, but describes some situation giving rise to it. For instance, recall the example (3), where in order to definitely refer to Lydia the presupposition (4) identifying her spatio-temporally must have been common knowledge. Then the corresponding q would be something like:

(5) All agents belonging to C see Lydia and that they are seeing her.

understood to entail that, firstly, that is the person referred to, and secondly, that (5)[8]. The latter entailment is explained because Lydia is perceptually *jointly salient* for C. Joint salience is a sort of coordination device, a signal prerequisite for coordinating the action (in this case – doxastic) of several agents. How coordination devices work falls outside the scope of this paper; Clark [9] provides a detailed psychological study[9].

Definition 4 being non-well-founded, it requires Barwise's [7] circular situation semantics for CK-checking. Situations are sets of "infons", that is, roughly speaking, well- or non-well-founded facts:

Definition 5 (Situations). Let x, y be individuals, H a well-founded and S a non-well-founded relation symbol. Situations and infons are elements of the largest classes SIT and INF, respectively, such that $s \in INF$ iff s is a triple, either of the form $\langle H, x, y \rangle$, or of the form $\langle S, x, s \rangle$, where $s \in SIT$, and $s \in SIT$ iff $s \subseteq INF$.

[7] First suggested by Harman [22], developed by Barwise [6], [7].

[8] Recall Schiffer's [37] example of two persons looking at each other and a candle standing on the table between them.

[9] Based on the work of Lewis [29] and Schelling [36]. Other coordination devices besides joint perceptual salience include convention, precedents, explicit agreement etc.

Of course this generalises straightforwardly to include conjunctive situations and multiple individuals. Thus for instance what (5) expresses is the circular situation s such that, taking S as the relation of seeing, $C = \{c_1, c_2, \ldots, c_n\}$ and x=Lydia:

$$s = \{\langle S, c_i, (s \wedge x)\rangle | i = 1, 2 \ldots, n\}$$

This permits to check whether p is common knowledge in a finite number of steps, by finding a true q describing a coordination device that gives rise to p on the lines of Definition 4. Yet, just like in the case of the shared-basis approach, the negative case seems not to be finitely checkable. Namely, if p is not common knowledge, then one must run through the entire model to ensure that; which, if the model is infinite, of course takes infinitely many steps. Thus it would seem that the fixed-point approach is in the same position as the shared-basis one. Moreover, they both can avoid the paradox by restriction to finite models, but even then require a cognitively untenable search across the entire model. In this respect they remain not much different from the original iterate approach.

However, on the fixed-point approach there is a way to solve the negative case more concisely. Recall that on this approach a proposition can only be common knowledge if there is a suitable coordination device. Now, while a shared basis can in principle be any set of propositions, the stock of coordination devices seems to be rather limited. Descriptions of perceptually salient objects are one class of these; other include assumptions about human nature, cultural facts and so on; Clark [9, Ch. 4] gives a thorough classification of (linguistic) coordination devices. Of course since this classification is based on psychological data, the limitation is of empirical rather than logical nature. Nevertheless, it makes it reasonable to claim that there is a defined finite subset of SIT that contains all relevant coordination devices – thus restricting the search space to a cognitively plausible size and solving the paradox.

To sum up, it seems that both the iterate nor the shared-basis accounts are ruled out by the common knowledge paradox. They could be saved by the restriction to finite models, but even so would remain cognitively implausible in requiring a search across the whole model. Only the fixed-point account can avoid this (which confirms Barwise's [7, p. 218] remark that it is the correct analysis of common knowledge) by limiting the processing to the set of coordination devices. Thus what should be done now is to spell out the content of this set in terms more precise than psychological. Of course this content depends very much on the context; thus a convenient account of it would be genetic – providing a general mechanism, by which coordination device tokens and, consequently, common grounds arise. An attempt at such

explanation will be discussed in Section 4.[10]

3 Some data, some scepticism

Even though the argument underlying CKH, as discussed in Section 1, is of rather formal nature, it seems desirable that the hypothesis should conform to the empirical data about language use. One attempt at gathering such data was by Clark et al. [11], who had people shown a photograph of two men, one well-known (namely, Ronald Reagan) and the other rather obscure and asked either of the two questions:

(6) You know who this man is, don't you?

(7) Do you have any idea at all who this man is?

Of course the answers to (6) tended to be about Reagan and to (7) – about the unknown man. This shows that the answerers reasoned about the interrogator's presuppositions: whoever utters (6), must presuppose that the referent of "this man" is well-known, and that if the interrogator had intended the utterance as felicitous, he must have had assumed that the answerer will also have that presumption – and so on. However, in a sense this experiment showed no more that the plain observation that people do communicate felicitously. Definitely, it could not show that they entertain higher-level knowledge attributions consciously for that purpose – since they do not. Nor could it *verify* the CKH: in fact, being concerned with tacit knowledge, it cannot be verified (and thus might better be called an explanatory assumption rather than a hypothesis).

Then perhaps it can be falsified? Some data has been interpreted against the CKH, attempting to show that people do not in fact rely on common knowledge when making felicitous utterances. Pickering and Garrod [33, 4.1.] (cf. also [17]) summarise it thus:

> For example, Horton and Keysar [24] found that speakers under time pressure did not produce descriptions that took advantage of what they knew about the listener's view of the relevant

[10] There is another formal difficulty: as Halpern and Moses [21] have proven, the common knowledge of p requires that all agents in C learn p simultaneously, which is rather doubtful for human agents. However, it can be overcome by, on the one hand, treating time not densely, but rather coarsely enough for whatever is learnt by the agents during the conversation to be learnt simultaneously; cf. [15]. On the other hand, an assumption has to be made that whatever have the agents learnt before the conversation is taken to have been learnt at some arbitrary time 0.

> scene. In other words, the descriptions were formulated with respect to the speaker's current knowledge of the scene rather than with respect to the speaker and listener's common ground. Keysar, Barr, Balin, and Paek [25] found that in visually searching for a referent for a description listeners are just as likely to initially look at things that are not part of the common ground as things that are. ... Additionally, speakers will sometimes use definite descriptions ... when the referent is visible to them, even when they know it is not available to their interlocutor (Anderson and Boyle [2]).[11]

Yet when not pressed for time, speakers do consider common knowledge in formulating utterances [24], [25]. On the whole,

> these results suggest that performing inferences about common ground is an optional strategy that interlocutors employ only when resources allow (see Clark & Wasow [12]). Critically, such strategies need not always be used, and most 'simple' (e.g., dyadic, non-didactic, non-deceptive) conversation works without them most of the time.

In other words, we do not resort to common knowledge in ordinary dialogue, but only when we have excess time or are forced to correct some misunderstanding.[12] In run-of-the-mill dialogue we only take our knowledge and the input of our interlocutor's utterances into account. Should these turn out to be insufficient, we can resort to considering her knowledge. And should this be yet insufficient, perhaps higher levels of mutual knowledge may come into play, but no higher than fourth, as Brown's [8] and Lee's [27] investigations

[11] Cf. also [5]

[12] This also explains some data that has been cited in favour of CKH can well be explained without resorting to it. For instance, Fischer [16] has explored what is the content of common ground in dialogue, by examining how people try to communicate with a (mock) robot in natural language, by typing utterances into a terminal. Not knowing what it would understand, they try to establish certain pieces of common ground. However, such a setting is by no means an ordinary dialogue setting. Rather, it is its peculiarity that forces the interlocutors to consider what the robot might know. This is analogous to above-mentioned result that we resort to common ground when time allows (and apparently time pressure in a typed dialogue with a machine is smaller than in a spoken one with a human) or when misunderstandings arise (which they did, since Fischer's robot only responded with an 'unspecified error message'). Even so, there is no reason to conclude that the interlocutors considered common knowledge rather than merely conjectures on what would the robot know (along with what syntax would it understand etc.).

evince. Therefore, in normal dialogue it is likely that knowledge attributions of depth one or two (as in processing 6 and 7) will suffice. The risk that some higher level of iterated knowledge hierarchy will fail is normally so small as not to be worth the effort of considering them. (And the examples where depth, say, four knowledge is essential – given by Schiffer among others [37] [10] [41] – must be too intricate to be realistic.) One might thus claim that in general CKH does not hold.

That, however, would be too rash a conclusion. The data cited above in favour of that sort of scepticism shows no more that there are cases where speakers fail to consider common knowledge. In doing so, though, they always run the risk of an infelicitous utterance (e.g. in they would refer to something visible to them and not to their interlocutors [2]). Thus this sort of data cannot refute CKH, just as the positive data could not prove it. Rather, it points to understanding CKH as a conditional: if the speaker wants to be certain that her utterance will be felicitous, she must consider common knowledge. However, as the usual philosophical scepticism has it, complete certainty is unattainable, and one reason for that is the feebleness of human reasoning. So in particular sometimes we neglect to take the requirement imposed by CKH into account, and do say something infelicitous.

Now, one could argue that this in fact is an argument against CKH. Namely, since – as has been admitted – we do run the risk of infelicity, a sufficient condition is that we should keep the risk reasonably small. Now, in most cases it is enough to consider the first few levels of the knowledge-attributions regress to minimise the risk satisfyingly; thus no common knowledge is needed. However, recall the domino-effect argument from Section 1. Common knowledge hypothesis is not about explicit beliefs, but about requirements for a consistent theoretical account of felicity and of presuppositions – so indeed, for pragmatics.

But then why do bother at all about defining common knowledge in a cognitively plausible way, if the theory is not meant to describe conscious mental processes or even unconscious, but fallible actual mental processes? Well, it still seems desirable that it should strive at being plausibly a model of a computational mind, even if a model idealised by assuming infallibility – rather than a purely formal model with no cognitive justification. One could say that it is desirable to approximate the mind, even if some idealising assumptions still must be made. The same holds for representing other aspects of language and reasoning – for instance, formal theories of grammar: people do utter ungrammatical sentences, but that is no argument against constructing a grammar that would predict the way a perfect speaker would speak. Indeed, this seems to be the only reasonable way of making progress

in general; and in particular, discarding CKH would topple much of the theory of pragmatic presuppositions – think of Stalnaker's [39] definition of presupposition as that which is part of common ground.

To sum up, it can be said that common knowledge shares the lot of knowledge as such. Knowledge is a very strong notion, not only in view of the usual epistemological scepticism, but in particular of the logical omniscience paradox. The latter stems from the observation that knowing one tautology, we should (by the standard epistemic logic S5 [23]) know all the other tautologies as equivalent to it; but, sadly – being the weak reasoners we are – we do not (cf. e.g. [31], [43]). Similarly, CKH has it that uttering a proposition felicitously, we should have common knowledge of all its presuppositions; but – for similar reasons – we may fail to have it. Now, that can be a motivation for developing a more realistic epistemic (or rather doxastic; cf. [43]) logic that would also allow for a fallibilistic account of common knowledge. But this is not to say that completing the account of the notion as it stands is not useful – quite the contrary; thus, it will be the purpose of the following section.

4 Evolution through coordination

It has been argued in the preceding sections that common knowledge of a proposition's presuppositions is indispensable for uttering it felicitously (CKH), and that the relevant notion of common knowledge is captured by the fixed-point approach (Definition 4). In order to establish what is common knowledge and what is not, that approach relies on a set of circular situations that are coordination devices (in the sense of Clark [9], Lewis [29] and Schelling [36]). Because the content of that set is thoroughly context-dependent, it will be most convenient to define it procedurally – by explaining how may the coordination devices arise.

Psychologically speaking, they do so basing on what we observe to share with the other members of some group that our interlocutors belong to. "Group" is to be understood widely; on the one hand it may be based on a general feature, like nationality, residence, education, occupation, hobby, religion etc. – and thus give rise to what Clark [9, Ch. 4] calls communal common ground, containing presuppositions about human nature, meaning of expressions in communal lexicons and cultural facts and norms. On the other hand, group in a narrow sense of people sharing some perceptual situation – e.g. present in the same place and time – so that salient perceptual events give rise to personal common ground, including presuppositions simi-

lar to (4) discussed above.

Thus, elaborating upon the fixed-point approach discussed in the previous section, the following definition can be established:

Definition 6 (Coordination devices). A circular situation s is a coordination device, for $C = \{c_1, c_2, \ldots, c_n\}$:

$$s = \{\langle R, c_i, (s \wedge x)\rangle | i = 1, 2 \ldots, n\}$$

where x is some object and R a relation such that $R(c_i, s)$ entails $\Box_{c_i} Obtain(s)$ and $\Box_{c_i} R(c_i, x)$, where P is some predicate.

For instance, in the example discussed in Section 2, R was the relation of seeing, and x was a person seen in the vicinity, and P a predicate of being seen. In general, there seem to be few relations that actually fulfil these conditions and thus give rise to coordination devices. One class relations of *perception* like that of seeing; another class, for communal common ground, are relations of *comprehension*, on the lines of, taking L to be a set of interpreted words:

(8) All agents belonging to C understand the words in L and that they are understanding them.

where "understanding a word" means using it according with its interpretation, and "understanding that c_i is understanding a word" means assuming that c_i uses it in the same way. Thus, taking R to be such a relation of comprehension, P would be "is understood" and $x = L$.

The question is, therefore, how do such groups C emerge? Answering it would amount to accounting for which of the many possible groupings of agents are relevant in being based on a relation that gives rise to a coordination device. The above indicates roughly what sort of relations these may be, but there is nothing logically compelling for relations of perception and understanding to be the ones giving rise to coordination devices. Rather, they must emerge with the emergence of the relevant groups – which occurs by evolution. The groups evolve as a means of improving linguistic communication by establishing common grounds; without them, speakers would always run risk of infelicity, with them, it may be minimised.

Now, when I meet a stranger b and have a conversation with her, uttering p I must decide which presuppositions to utter explicitly and which leave as presuppositions sensu stricto. Thus I face a series of decision problems of the form: utter the presupposition q explicitly or not? Let these actions be written, respectively, $\mathbf{q}$ and $\bar{\mathbf{q}}$. What is their utility U?[13] If b does not know

[13] If these terms sound unfamiliar, consult some basic introduction to game theory, for instance [32] and the references therein.

that q, $U(\bar{\mathbf{q}})$ is negative, because it is likely to prevent her from understanding p, thus spoiling the purpose of my uttering it.[14] Conversely, if she knows it, then $U(\mathbf{q})$ is negative, because to say what is known anyway entails a risk of losing face – and also because it may generally be costly to speak (cf. [35], [34]). Thus, for some utilities $\alpha, \beta, \gamma \in \mathbb{R}_+$:

	$\Box_b q$	$\neg\Box_b q$
$\mathbf{q}$	$-\alpha$	β
$\bar{\mathbf{q}}$	β	$-\gamma$

How do I decide, which strategy to take? That depends on how probable which state of b is. If there is something perceptually salient around, b will likely see it, and thus know the relevant presupposition. If some observations about b give evidence that she is an English speaker (e.g., she utters sounds resembling these of English words), then she will likely know the presuppositions stating the meanings of some set of English expressions. That is to say, given the evidence about b and the context, I cast hypotheses about what b is likely to know. Then, writing P for probability, the expected utilities are:

$$EU(\mathbf{q}) = P(\Box_b q) \cdot -\alpha + (1 - P(\Box_b q)) \cdot \beta \;=\; \beta - P(\Box_b q)(\beta + \alpha)$$

$$EU(\bar{\mathbf{q}}) = P(\Box_b q) \cdot \beta + (1 - P(\Box_b q)) \cdot \gamma \;=\; P(\Box_b q)(\beta + \gamma) - \gamma$$

The optimal strategy is to take the decision yielding the greater expected utility. It follows from the above that

$$EU(\mathbf{q}) > EU(\bar{\mathbf{q}}) \text{ iff } P(\Box_b q) < \frac{\beta + \gamma}{2\beta + \gamma + \alpha}$$

which the long run amounts to a mixed strategy, depending on the values of α, β and γ – namely that of choosing $\mathbf{q}$ whenever the inequality on the right-hand side above is satisfied; call this strategy $\mathbf{m}$. (For instance, assuming that $\gamma = \alpha$, $\mathbf{m}$ will amount to uttering q explicitly whenever I judge the hearer to be less likely to know that q than not.) In fact, following $\mathbf{m}$ amounts to obeying the Gricean [20] principle of Quantity: say neither too much nor too little.

Now, of course my interlocutor is in a symmetric situation. Over the course of our conversation(s), either of us follow $\mathbf{m}$, or some other strategy. One follows $\mathbf{m}$ whenever two conditions are met: firstly, one strives for maximising utility (which can be granted by assumption) and secondly, one correctly judges the interlocutor's likelihood of knowing the relevant presuppositions. Therefore whenever both of us can correctly judge what the other knows, we will certainly both follow $\mathbf{m}$; if either of us does not follow $\mathbf{m}$,

[14] Cf. Zeevat's [45] conditions on common ground for the successfulness of an assertion.

but some other strategy $\bar{\mathbf{m}}$ they must be unable to correctly judge what the other knows[15]. Then, for some $\delta > \eta$ the utilities are:

	$\mathbf{m}$	$\bar{\mathbf{m}}$
$\mathbf{m}$	δ, δ	δ, η
$\bar{\mathbf{m}}$	η, η	η, δ

so that $\mathbf{m}, \mathbf{m}$ is a Nash equilibrium. That agrees with the above point: whenever we can play $\mathbf{m}$, we will; so one only plays $\bar{\mathbf{m}}$ if one cannot rightly judge the interlocutor's knowledge.

Moreover, the utilities given above for both agents can be summed up and thus thought of not as particular agents', but the group's – which is reasonable, since not only the agents benefit from successful communication, but the group as well. Then, taking U as this summary utility, $\mathbf{m}$ fulfils the following two conditions, for all $\bar{\mathbf{m}} \neq \mathbf{m}$:

$$U(\mathbf{m}, \bar{\mathbf{m}}) \leq U(\mathbf{m}, \mathbf{m})$$

$$U(\bar{\mathbf{m}}, \mathbf{m}) = U(\mathbf{m}, \mathbf{m}) \rightarrow U(\bar{\mathbf{m}}, \bar{\mathbf{m}}) < U(\mathbf{m}, \bar{\mathbf{m}})$$

Therefore $\mathbf{m}$ is an evolutionarily stable strategy (ESS), since these conditions form the definition of ESS (cf. [44]). Thus it is also self-preserving: once a group of agents uses an ESS, there (provably) exists a uniform invasion barrier. It is a fraction ϵ of the group such that even if that fraction tired to change $\mathbf{m}$ to some other strategy (i.e. "mutated"), it would be forced back to the original $\mathbf{m}$.

This gives an answer to the question posed above: which of all the possible groupings of agents these ones are relevant in being based on a relation that gives rise to a coordination device? Namely these, the members of which employ the same ESS. Recall: the ESS amounts to systematically correctly judging whether the other members of the group know a presupposition q (or, slightly generalising, a set thereof). That, in turn, can be thought of as co-occurring with the circular relation of comprehension or perception in the sense of Definition 6. For instance, when I see something and see that both myself and my interlocutor see it – I can judge her to know the relevant presupposition. Doing that systematically and by all members of the group constitutes common knowledge.

In other words, by accumulating decision problems on whether others might know some presuppositions, an ESS can emerge – and with it, a corresponding coordination device. Therefore the set of coordination devices

[15]Strictly speaking, it is also required that the utilities are the same for both of us. But that can be granted by assuming that generally utilities are objective – which seems to be implicit in any evolutionary approach.

can be defined as the set of relevant ESS-s that exist among (human) agents ("relevant" meaning such that pertain to uttering or not some sets of presuppositions). This, in turn, allows to upgrade Definition 4 on the following lines:

Definition 7 (Common knowledge, more plausibly). A proposition p is common knowledge among the agents in C iff

$$q \rightarrow \forall x \in C \, \Box_x (p \wedge q).$$

where q holds because of an ESS employed by the members of C.

This version is more plausible, because it restricts the search space to a finite and, moreover, quite limited set of the ESS-s used in C pertaining to the particular presupposition. This set is finite, because any *actual* group of agents can only employ a finite number of ESS; it furthermore cannot be too large, since there seem to be no more than a few relations of perception or comprehension available (that can give rise to the coordination devices corresponding to the ESS).

Therefore extending the fixed-point approach in this direction – that is, providing a genetic account of coordination devices in terms of evolutionary games – finally provides a satisfactory, cognitively plausible solution to the common knowledge paradox. Of course the formulation of Definition 7 needs refinement; in particular, it should be made explicit, what can "relations of perception or comprehension" exactly be. On a more general level, it might be interesting to use a more realistic – fallibilistic – notion of knowledge (cf. [43], [13]). Nevertheless, I think that this approach is promising in integrating the formal and psychological aspects of common knowledge in conversation.[16]

References

[1] Kazimierz Ajdukiewicz. Sprache und Sinn. *Erkentniss*, 4, 1934.

[2] A. Anderson and E. Boyle. Forms of introduction in dialogue: their discourse contexts and communicative consequences. *Language and Cognitive Processes*, 9, 1993.

[16] One direction of extending this work would surely be to combine it with the dynamic accounts of how information present in the conversation becomes common ground; cf. [45], [3], [18, Ch. 6]. Thanks go to Henk Zeevat and Barteld Kooi for comments on earlier versions of parts of this working paper. Most of it has been presented at the Youth Philosophy Forum in May 2004 in Lublin, Poland.

[3] N. Asher and A. Gillies. Common ground, corrections and coordination. *Argumentation*, 17, 2003.

[4] Robert Aumann. Agreeing to disagree. *Annals of Statistics*, 4, 1976.

[5] E.G. Bard, A.H. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon, and A. Newlands. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42, 2000.

[6] Jon Barwise. Three views of common knowledge. In Moshe Y. Vardi, editor, *Proceedings of the second conference on theoretical aspects of reasoning about knowledge*, San Francisco, 1988.

[7] Jon Barwise. On the model theory of common knowledge. In *The situation in logic*. CSLI, Stanford, 1989.

[8] G. Brown. *Speakers, listeners and communication. Explorations in discourse analysis*. Cambridge UP, 1995.

[9] Herbert H. Clark. *Using language*. Cambridge UP, 1996.

[10] Herbert H. Clark and C.R. Marshall. Definite reference and mutual knowledge. In A.K. Joshi, B.L. Webber, and I.A. Sag, editors, *Elements of discourse understanding*. Cambridge UP, 1981.

[11] H.H. Clark, R. Schreuder, and S. Buttrick. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22, 1983.

[12] H.H. Clark and T. Wasow. Repeating words in spontaneous speech. *Cognitive Psychology*, 37, 1998.

[13] R. Fagin and J.Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34, 1988.

[14] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about knowledge*. MIT Press, 1995.

[15] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. Common knowledge revisited. *Annals of Pure and Applied Logic*, 96, 1999.

[16] K. Fischer. How much common ground do we need for speaking? In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *BI-DIALOG 2001*, Bielefeld, 2001.

[17] S. Garrod and M. Pickering. Toward a mechanistic psychology of dialogue: the interactive alignment model. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *BI-DIALOG 2001*, Bielefeld, 2001.

[18] Jelle Gerbrandy. *Bisimulations on planet Kripke*. PhD thesis, University of Amsterdam, 1999.

[19] R.W. Gibbs. Mutual knowledge and the psychology of conversational inference. *Journal of Pragmatics*, 11, 1987.

[20] H. Paul Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Syntax and semantics*, volume 3. Academic Press, New York, 1975.

[21] Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3), 1990.

[22] Gilbert Harman. Review of *Linguistic behaviour*. *Language*, 53, 1977.

[23] J. Hintikka. *Knowledge and belief. The logic of two notions*. Cornell UP, 1962.

[24] W.S. Horton and B. Keysar. When do speakers take into account common ground? *Cognition*, 59, 1996.

[25] B. Keysar, D.J. Barr, J.A. Balin, and T.S. Paek. Definite reference and mutual knowledge: process models of common ground in comprehension. *Journal of Memory and Language*, 39, 1998.

[26] Marga Kreckel. *Communicative acts and shared knowledge in natural discourse*. Academic Press, London, 1981.

[27] B.P.H. Lee. *Establishing common ground in written correspondence*. PhD thesis, University of Cambridge, 1998.

[28] B.P.H. Lee. Mutual knowledge, background knowledge and shared beliefs: their roles in establishing common ground. *Journal of Pragmatics*, 33, 2001.

[29] David Lewis. *Convention*. Harvard UP, 1969.

[30] John McCarthy, M. Sato, S. Igarashi, and T. Hayashi. On the model theory of knowledge. In *Proceedings of IJCAI-1977*. MIT Press, 1977.

[31] J.-J. Ch. Meyer and W. van der Hoek. *Epistemic logic for AI and computer science*. Cambridge UP, 1995.

[32] Marc Pauly. Some game theory for logicians. Course notes, ILLC, Amsterdam, 2003.

[33] M. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, to appear.

[34] Robert van Rooy. Quality and quantity of information exchange. *Journal of Logic, Language and Information*, 12, 2003.

[35] Robert van Rooy. Evolution of conventional meaning and conversational principles. *Synthese*, to appear, 2004.

[36] T.C. Schelling. *The strategy of conflict.* Harvard UP, 1960.

[37] Stephen Schiffer. *Meaning.* Clarendon, 1972.

[38] Dan Sperber and Deirdre Wilson. *Relevance. Communication and cognition.* Basil Blackwell, 1986.

[39] R. Stalnaker. Pragmatic presuppositions. In M.K. Munitz and P.K. Unger, editors, *Semantics and philosophy.* New York UP, 1974.

[40] R. Stalnaker. Common ground. *Linguistics and Philosophy*, 25, 2002.

[41] P.F. Strawson. Intention and convention in speech acts. *Philosophical Review*, 75, 1964.

[42] P. Vanderschraaf. Common knowledge. *Stanford Encyclopedia of Philosophy*, 2002. `http://plato.stanford.edu/entries/common-knowledge/`.

[43] Frans Voorbraak. *As far as I know. Epistemic logic and uncertainty.* PhD thesis, Utrecht University, 1993.

[44] J. Weibull. *Evolutionary game theory.* MIT Press, 1995.

[45] Henk Zeevat. The common ground as a dialogue parameter, 1997.