Economy in embodied utterances[1]

Matthew Stone

Rutgers University

In natural language generation (NLG)—the branch of computational linguistics concerned with constructing utterances to communicate specified information to users—brevity has always been a central concern. Naive designs for NLG systems often yield output utterances that are distractingly verbose; we'll see examples in Section 1. Doing better requires surprisingly sophisticated models of planning and coordination in conversation. In developing this computational perspective on pragmatics, NLG research suggests ways to account more generally for the efficiency of natural utterances in conversation.

In this chapter, I draw on the perspective of NLG to explore one kind of brevity, TEXTUAL ECONOMY, or, more generally, COMMUNICATIVE ECONOMY, which I introduce in Section 1. In communicative economy, speakers organize their actions to contribute to multiple communicative goals simultaneously, rather than pursuing each communicative goal separately. This makes their utterances more efficient and more concise they would be otherwise. The project of this paper is to present a computational sketch of the reasoning required to achieve successful communicative economy in conversation, and to use this account to better understand the possibilities and limits of this kind of brevity.

I begin in Section 2 by invoking general accounts of how agents plan efficient strategies for action in the world—the theory of planning in artificial intelligence

(AI). The key reasoning problem here is to recognize opportunities for efficient action as you flesh out what to do; the solution is maintaining a rich network of intentions to guide your deliberation. Rich communicative intentions, in particular, allow agents to recognize opportunities to communicate efficiently.

However, brief utterances aren't just a matter of speakers' plans; in communication, hearers must also recognize efficiencies. Otherwise their understanding is incomplete. I explain this in Section 3 in terms of the special requirements of COHERENCE that follow from the collaborative organization of conversation. Coherence limits efficiency. In many cases, it seems, we have to keep our thoughts separate in utterances, and work to express them overtly and distinctly, in order to ensure that we coordinate successfully on the process and content of communication. I close in Section 4 by exploring cases where interlocutors can achieve coherence only by elaborating compact, bare-bones utterances with explicit signals of the relationships among their ideas.

1 Communicative economy and overloaded intentions

The challenge of brevity arises as soon as we consider the problem of generating fluid and natural instructions for physical action. Imagine hearing (1) in a situation like the one depicted in Figure 1, for example.

(1) Take the rabbit out of the hat.

In this case, you immediately identify which rabbit you have to grab: it's the one in the hat. And you immediately identify which hat is at issue: it's the one with the rabbit.

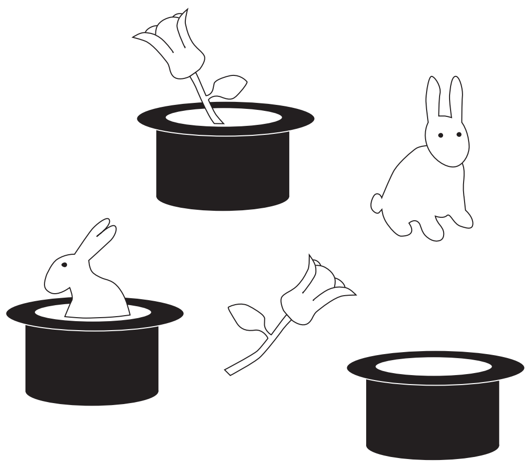

Figure 1 A situation for action, in which (1) would be appropriate.

The success of this brief instruction depends on interactions between the descriptions it contains. To see this, suppose we operationalize a theory of communicative actions and communicative intentions, along the lines explored by Grice (1957, 1975) and Searle (1969, 1975). The speaker needs to accomplish a series of speech acts. In (1), the speaker must refer to the rabbit that needs to be moved. The speaker must refer to the location from which the rabbit needs to be extracted. And the speaker must specify an action that the hearer is supposed to carry out on these two arguments.

You'd naturally expect that you could plan each of these speech acts separately, and then assemble the results into a larger utterance that achieves all of them together.[2] Indeed, since the referents are drawn directly from the speech situation, you might naturally abstract away from the intrasentential dynamics of context change, and formulate each noun phrase so that it directly specifies its referent in the visual context. But if you do this, you get (2) rather than (1).

(2) Take the rabbit that's in a hat out of the hat that holds a rabbit.

When you plan a speech act to refer to that hat, you can't just call it *the hat*. You have to distinguish it from the other hat by what makes it different: it has a rabbit in it. Conversely, there are lots of rabbits, so if you're planning a speech act to refer to that rabbit, you have describe it as *the rabbit in a hat*.

You can improve things by adding a treatment of the evolving context, planning each reference in turn, and exploiting the effects of earlier ones to bring entities to salience. Now you might get (3).

(3) Take the rabbit that's in a hat out of it.

Even (3) still strikes me as somewhat awkward in comparison to (1). In any case, we want our system to be able to learn the best choice between (3) and (1). Something's wrong if our approach cannot even entertain (1) as a possible utterance in this situation.

The real problem here is that communicative goals are not independent. Utterance planning has to recognize when material that is provided in the service of one communicative goal helps indirectly to achieve other ones. Such indirect relationships have been called OVERLOADING (Pollack 1991). Overloading an intention means using an intended action not just for its contribution to the goal for which you originally committed to it, but also for its contribution to other goals that you hope to achieve. So (1) exhibits PRAGMATIC OVERLOADING (Di Eugenio &

Webber 1996) in that information provided across its different constituents makes it possible to achieve multiple goals in an efficient way. This paper explores pragmatic overloading as one explanation for cases of brevity.

Overloading is a general feature of human action—one that often makes our everyday actions more efficient. For example, we often combine errands together into efficient routines by opportunistically planning efficient routes (Hayes-Roth & Hayes-Roth 1979). Suppose you have to go to the pharmacy to pick up some medication and go to the dry cleaner to pick up some laundry. Suppose further that the pharmacy is close to the dry cleaner. Then you might start by forming an intention to go the pharmacy, and then overload that intention. You can take advantage of the fact that the pharmacy is most of the way to the dry cleaner, and plan to go to the dry cleaner immediately afterwards.

Such decisions are sometimes conscious choices. But overloading is just as evident in patterns of behavior that may be routine or automatic. In such cases, in talking of an agent's choices, commitments, or intentions, we are giving a rational reconstruction the phenomena. That is, we are developing intuitions and metaphors to make sense of the underlying cognitive mechanisms that determine behavior—whether people have introspective access to their operation or not. See Agre (1997) for more on this way of reading computational models of efficient, opportunistic action. This methodological perspective is natural in computational approaches to cognitive science, but is common in philosophy as well. Grice (1957) himself took talk of communicative intentions to involve this kind of rational reconstruction.

Overloading is evident not just in patterns of practical action but in discourse as well. Consider example (4) from a children's craft book, as analyzed by Di Eugenio & Webber (1996).

(4) Cut the square in half to create two triangles.

In this setting—where you're really going to carry out the author's instructions and build a craft figure according to this recipe—you need to recover a detailed specification of each step of action. Here, you need to know that you should use scissors, cut along the diagonal, and make two pieces, that the two pieces will be the same size and shape, and so forth. The author is getting some of this information across explicitly. But the author is also relying indirectly on your ability to draw inference and the implicit connections between different parts of the instruction. When you find you are *cutting* the paper square, you naturally expect to use scissors. The inference to cutting on the diagonal is more complicated. But if you know you're starting from a square, and you know that the purpose or result of the action is to create triangles, you know you can't cut parallel to the sides of the square. Di Eugenio and Webber argue that it's useful to think of these inferences as cases of pragmatic overloading. The speaker decides to describe one aspect of the situation, say, that the goal of the action is to create triangles. But this information (and the audience's inferential interpretation) also contributes towards the other goals the speaker needs to achieve to give a complete instruction. Exploiting this connection gives the final utterance its economical expression.

It's crucial that overloading a practical intention, as in efficiently running errands, is really the same thing as overloading a communicative intention, as in (4). Our joint activities mix linguistic utterances with other kinds of signals and with ongoing practical activities. The interactions among these diverse actions often exhibit pragmatic overloading, which can be accounted for only through a common model of intentions and deliberation.

Consider first the coordination between speech and coverbal gesture. Lascarides & Stone (2009: 168) discuss an embodied explanation of the solution of a physics problem, excerpted in (5) below.

(5) If you see this larger ball as ten small balls like that...
The speaker gets down off his chair to match his interlocutor, who sits across from him on the floor. His right arm now extends out in front of him at shoulder height, with his fingers curled and his index finger touching his thumb, as though holding a pen. During the gesture, the hand sweeps along a horizontal line further to the right.

The extended discourse from which (5) is extracted evokes a hypothetical Galilean experiment to explain why Newton's law of gravity—which says that the gravitational force between two objects is proportional to the product of their masses—actually predicts that all objects fall to Earth at the same speed. The interlocutors have agreed to frame their explanation in terms of a particular puzzle: a ten-unit mass experiences ten times the force of a one-unit mass, yet the two fall identically. In (5), the speaker's words introduce a perspective that encodes the key insight in the explanation: Newton's law describes quantities of mass, rather than common-sense objects. Meanwhile, the speaker uses his gestures to act out the setup of a hypothetical experiment in which a series of masses are laid out at a level height and dropped simultaneously to measure gravity's effect. See Lascarides &

Stone (2009) for an extensive discussion of form and meaning in this example.

In (5), much as in (4), we find evidence of the overloading of intentions for efficient action. The speaker's words include the deictic expression *like that*, which indicates that an accompanying communicative action demonstrates the spatial relationship among the balls. Since this link is so fundamental to the organization of the communicative action, we might naturally expect that it arises early on in the process of planning. However, this link requires only that there be a gesture that indicates the location of the balls. A simple gesture might be sufficient for this, such as an unmarked pointing gesture with index finger extended. The gesture here, however, not only indicates the location of the balls, but embodies the action of a scientist placing the balls into an experiment. This extra depiction enriches the message that the speaker is able to convey with the utterance. This suggests that we should think of the speaker's gesture as a case of pragmatic overloading.

Language use in everyday joint activity also showcases connections between practical intentions and communicative ones. Consider (6), from Clark & Krych (2004: after Figure 6, p. 72).

(6) a. Sam: 'kay now get .. a-uh [eight piece green]
 b. Ted: [*reaches for block*] *exhibits block*
 c. Sam: and join the two . so it's all sym[metric]
 d. Ted: [*poises block*]
 e. Sam: yeah, right [in the center]
 f. Ted: [*attaches block*]

Example (6) gives a nine-second transcript of an interaction from a human-subjects experiment in which Sam guides Ted in the step-by-step assembly of a Lego model. Clark & Krych (2004) principally aim to document the closely-timed interaction they discover in such dialogues. For example, when Ted poises the block in its intended place at (6d), Sam can confidently wrap up his description of the target location and can immediately proceed to acknowledge Ted's correct placement.

Such fluid interactions seem to depend on integrated representations of practical and communicative intentions and integrated processes of deliberation towards practical and communicative goals. In (6b), Ted intends to use a specific block to comply with Sam's instruction; but first, he must get the block. Eventually, he must also confirm that the block is correct. As it turns out, getting the block—to do what's required in the world—naturally affords a streamlined opportunity to check the identity of the block with Sam. Ted can move the block into a prominent position and direct his attention to the block and then to Sam. Clark & Krych call this EXHIBITING the block. If this is right, Ted's action of exhibiting the block—a communicative action—overloads his intention to get the block in the first place—the practical action that the task calls for. More generally, it shows the importance of a general theory of intentions in deliberation as an account of efficient action, both communicative and practical alike.

2 Intentions, deliberation and efficient action

A general theory of intentions in deliberation must conceptualize intentions as complex mental attitudes, following Bratman (1987) and Pollack (1990). A good way to see this is by considering the famous BLOCKS-WORLD domain from AI research. The problem is to control a robot arm and move blocks around on a

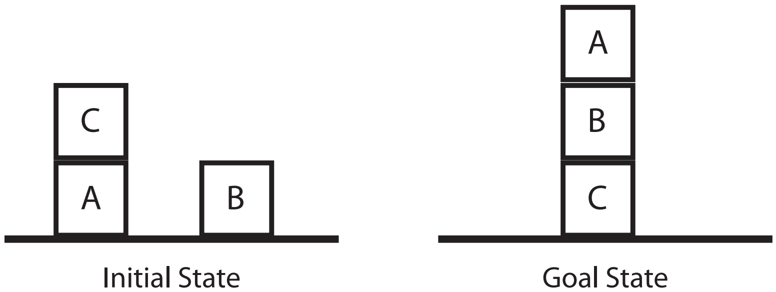

Figure 2 A blocks-world challenge: efficiently starting from the situation at left and building the tower shown at right.

tabletop. When a block has nothing on top of it—when the block is CLEAR in the argot of the blocks world—the robot can pick the block up and put it on another clear block. There's also a tabletop that can accommodate any number of blocks. The task of the robot is to build towers.

The blocks world turns out to enable a simple demonstration of the necessity of overloading in planning. One way to realize multiple goals is to sequence together programs of action that achieve them one after another. When realized straightforwardly, however, this strategy turns out not to lead to efficient programs of action in the blocks world. Instead, planners need to track the relationships among actions and their contributions to multiple goals.

The key case is the so-called SUSSMAN ANOMALY illustrated in Figure 2. The robot starts with block *C* on top of block *A*, and block *B* on the table. The robot needs to build a tower with *A* on top of *B* and *B* on top of *C*. To a human observer, it's obvious what the robot needs to do in this case. First it moves *C* to the table, then it puts *B* on top of *C*, and finally it puts *A* on top of *B*. The problem is to synthesize this plan by algorithmic reasoning.

One general framework for such reasoning is the intuitive idea of means–end analysis (Newell & Simon 1961). At each step of planning, you identify something that needs to change, find an action that will accomplish the change, and then adopt the goals necessary to use the action in this way. Here the robot's problem requires achieving two goals simultaneously. You might have expected that the robot could meet these goals by first achieving one goal efficiently, then moving on to accomplish the other goal. But this divide–and–conquer strategy turns out not to work here. If you try to achieve either of the goals on its own in an efficient way, you wind up having to undo your work as soon as you turn to accomplish the other goal. If your goal is to have *B* on top of *C*, you can move *B* there right away. But your next step will require you to move *A*, which you've just buried. If your goal is to have *A* on top of *B*, you can put *A* there as soon as you get *C* out of the way. But again, you then discover that *B* is in the wrong place, and you've just buried it.

The correct plan actually INTERLEAVES the actions involved in achieving the two goals. When you move *C* on to the table, you're working towards clearing *A* so you can put it at the top of the tower. You then switch gears to put *B* on *C*, achieving the other goal. Only then do you return to the goal of putting *A* on *B* that you were working towards with your first action.

This reasoning involves recognizing an opportunity for efficient action. Suppose the planner starts with the goal of having *A* on *B* and works backwards. The planner infers that it must put *C* on the table then put *A* on *B*. Now the planner considers the goal of getting *B* on *C*. Given the plan so far, we can't move *B* after we put *A* on *B*. That's one of the goals we've committed to. We can't put *B* on *C* before we put *C* on the table either. We're committed to keeping *C* clear till we move it away. However, if we move *B* to *C* in the middle while *A*, *B* and *C* are all clear, we can achieve our new goal while maintaining all the effects we've already planned for. In effect, the interleaved plan overloads the intention of moving *C* to the table. The action gets *A* clear and also moves *C* into position to put *B* there.

This deliberative strategy requires rich models and representations of intentions. A simple idea is that an intention is just a special kind of goal—perhaps, a goal that an agent is committed to. As Cohen & Levesque (1990) show, this idea leads to a useful theory of intentions for modeling agents' mental states. However, we need a stronger theory of intentions to explain how agents overload their intentions to act efficiently. In particular, all the commitments that an agent makes in the course of its means–end analysis must become part of its intentions. Only then can the agent assess whether new actions respect decisions already made.

If communicative intentions exhibit the same kinds of overloading, they are going to need the same rich structure. Communicative intentions have to commit to words and their meanings. But they also have to track changes to the evolving context of the conversation. And they have to show how the speaker is committed to aspects of that evolving context that their interpretation depends on, and track how those contextual requirements are met through aspects of the nonlinguistic environment or through the contextual contributions of previous communicative action.

Programmatically, this insight about intention allows us to refine our intuitive understanding of examples (1), (4), (5) or (6), As a simple case study, let's consider how the account applies to (1).[3] An instruction like this actually requires the addressee to perform a specific action involving specific objects. So we can start from the assumption that the instruction involves SPEAKER REFERENCE to these objects, in the sense of Kripke (1977). Moreover, the situation makes certain entities salient, and we naturally prefer to resolve speaker reference to these entities rather than any others. Here the noun phrases show where the sentence potentially refers to things. So understanding the utterance will involve identifying some object as the speaker reference R of the direct object noun phrase and some object for the speaker reference H of the source noun phrase.

To plan an utterance involves anticipating how linguistic action contributes to such communicative goals. Here the job is to track the evidence uses of words give about the properties of the corresponding referents. Saying *the rabbit* gives evidence that R is a rabbit. Saying *the hat* gives evidence that H is a hat. Crucially, here, we also get constraints on these referents from our common-sense knowledge, which allows us to recognize certain additional commitments that the speaker must be making in using the instruction—the SPEAKER PRESUPPOSITIONS in the sense of Stalnaker (1973). In particular, when we say that we're *taking R out of H*, we discover that R must be in H. It's generally true, when you take one thing out of another, that the thing that you take out starts off inside the thing that you take it out of. These additional relationships establish interconnections among communicative actions and make it possible to overload them.

Concretely, if you're picking this rabbit and this hat, then the constraints that R is a rabbit, that H is a hat and that R is in H are all simultaneously satisfied. Indeed, there's no other comparable tuple of salient individuals that satisfy these constraints—so there's nothing else one could take as alternative interpretations of this utterance in this context. That's enough to explain how it is that the utterance *take the rabbit out of the hat* manages to achieve all three of the speaker's communicative goals.

Now imagine planning the instruction using a process of intention-based deliberation on par with what we described for the blocks-world domain. We model the process within a framework that takes into account the referential requirements of practical instructions. Initially, we start from the goal of identifying R, identifying H, and characterizing the action that the addressee must carry out on them. The variables initially have an unconstrained interpretation but are linked privately with their intended referents.

Planning proceeds by adding content step-by-step to a provisional communicative intention, corresponding to the partially-constructed utterance. Each step commits to an open-class item, perhaps a noun or a verb or a longer, multiword expression with an idiomatic meaning, together with any functional elements (determiners, prepositions and the like) that may be syntactically required to fit that item into the incomplete utterance. You start with an empty description, so that R and H could be anything. Then you might add *the rabbit* and factor in its effects on interpretation: it lets you zero in on particular values for R. Then you might add *the hat*, and factor in its effects on interpretation: it lets you zero in on particular values for H. Finally, you might characterize the action with *take out of*. This has the interpretive effect of allowing the hearer to see that R must be in H, solving the overall problem for communication.

This process is guided by the communicative intention at each step. This representation tracks the intended interpretation of each unit, the requirements that this interpretation places on the context, and the effects that the unit has on the addressee's interpretation of the instruction. These varied representations of the effects of utterances make it possible to recognize, and to exploit, their potential to contribute to multiple communicative goals: to plan brief utterances.

3 Intentions and the principles of collaboration

While this account of intentions suggests how speakers might convey information more economically by recognizing opportunities to overload their communicative intentions, the account offers important insights into the limits of brevity as well. Communicative intentions are prototypically collaborative. In communication, interlocutors use utterances to contribute propositions to conversation, and thereby to address and resolve open questions, as part of a joint process of inquiry. This section argues that collaborative intentions generally, and communicative intentions in particular, are subject to constraints of COHERENCE that limit how tightly overloaded they can be.

Researchers such as Cohen & Levesque (1991) have argued that intentions have a distinctive role to play in the deliberations of agents that work together, because teamwork requires agents to coordinate with one another (Lewis 1969). A prototypical coordination problem arises when two people *A* and *B* want to meet together face-to-face for a discussion. They might choose to meet in *A*'s office, or in *B*'s office, or in a favorite café. They might be indifferent about the venue for the meeting, as there might be no particular advantage to any of the places they are considering. Nevertheless, the meeting will only happen if *A* and *B* actually wind up in the same place.

In coordination problems, agents can't just reason about the effects of their own actions. They also have to align their decisions to each other, and keep their decisions aligned with one another throughout the execution of the plan. This requires special work if the world is unpredictable, if communication is difficult, or if agents have partial information (Cohen & Levesque 1991). For example, people normally schedule meetings by having a brief conversation, perhaps by phone or email, in which they settle the arrangements that will govern their joint action together.

To achieve coordination, people arrive collaboratively at appropriate shared commitments that describe what they will do as a group. These commitments guide deliberation like ordinary intentions. But they also cement the teamwork agents do to understand each other, to meet their obligations to the group, and to track the contributions they themselves make, as they deal with the vicissitudes of carrying out their agreements. I'll follow standard practice in AI and refer to these commitments as collaborative intentions. A key strategy that people use to identify progress towards their goals and deal with opportunities or obstacles is to manifest and track their collaborative intentions.

We can point to pervasive analogies between teamwork as it applies in a cooperative conversation, and teamwork in pursuit of shared practical goals. Let's start with understanding. To understand your teammates' actions, you have to recognize the intentions with which they act. These intentions involve commitments not only to action, but also to relevant facts about the circumstances in which the action is being carried out and about the contributions which the action is going to make. Recognizing an intention is a process of explanatory inference that can start from background information about the agent's action, knowledge, preferences and goals, but that can also make assumptions to fill in new information about the agent as well. Reasoning about preferences is particularly important when agents maintain an open-ended collaborative relationship with one another (Cadilhac, Asher, Benamara & Lascarides 2011).

Imagine, for example, you are part of a team that's catering a party. You see one of your colleagues carrying a full tray of drinks towards a closed door. You probably conclude that your colleague intends to distribute the drinks to party-goers in the next room. You've used what you already know about your colleague's beliefs: you see your colleague moving, you see the full drinks, and it's obvious that your colleague is moving purposefully and is aware of the surroundings. You've also used what you already know about your shared goals: drinks must be distributed if the party is to be a success. At the same time, you've made additional assumptions. Perhaps you were previously unaware that the next room was open to guests, or that it even existed. But given the intentions you've recognized, your colleague must know about these guests and have the particular goal of serving them.

It's crucial that intention recognition gives you this new understanding of your colleague and the ongoing activity. You'll need it to track the state of the collaboration and to plan your own contributions to it. Your engagement with each other means that your colleague's continued action, like carrying the drinks here, provides the shared evidence you need to keep coordinating, and relieves you of the need to explicitly discuss each step of progress in the task. Thomason, Stone & DeVault (2006) explore this reasoning in more detail. They argue that collaborative reasoning always involves a shared presumption that team members act cooperatively and are engaged in tracking each other's contributions. This recalls the famous Cooperative Principle and Maxims of Grice (1975), of course.

To use language collaboratively, agents need to recognize the intentions behind utterances in much the same way. Suppose you're doing some shopping on the way home. Your spouse messages you with (7).

(7) Can you get me some goat cheese for the frittata I'm making tonight?

Part of understanding this utterance is recognizing the background goals and beliefs that make this utterance a useful request for your ongoing collaboration. This again involves appeal to relevant background knowledge. You may know, for example, that your spouse prefers firm but creamy, slightly aged goat cheeses for frittatas; you'll assume that you're expected to get one of those, rather than a soft fresh cheese or something exotic like a Norwegian brunost. You may also be led to new assumptions about your spouse, just as in the practical case. In this case, you may have been previously unaware that your spouse planned to make this frittata. It's nevertheless unproblematic to recognize this aspect of the plan from what your spouse says.

In teamwork, deliberation depends on the network of intentions that you and your collaborators have already committed to. Two cases are worth noting. You can choose to ACCOMMODATE your collaborators (Thomason 1990), and act in support of their intentions, by taking steps to assist their activities or defuse obstacles they may face. For example, when you see your colleague headed for the closed door with a tray of drinks, you can choose to accommodate them by opening the door. That way they don't have to orchestrate the problematic maneuver by themselves.

Another case, which Cohen & Levesque (1991) emphasize, is that you can warn your collaborators about obstacles in their plans that cannot be corrected. Such flaws suggest that you and your collaborators will have to reconsider the activities that are underway. For example, you may know that your catering colleague with the drinks is lost. That door leads to the closet, not to a room with party-goers. In this case, cooperative rationality dictates that you should say something, and get your colleague reoriented. More generally, any mismatch you detect between an agent's actions, preferences and circumstances may require your intervention.

Exactly the same kinds of options suggest themselves in language use. In the case of a request like (7), the obvious way to accommodate your spouse is to comply with the request. Your spouse has a plan that requires goat cheese, you can get the cheese, and thereby help make sure that the plan is carried off. You may also be in a position to recognize and correct other potential obstacles to this plan as well. You might follow up with (8), for example, if you suspect that your spouse is unaware of the cooking you've been doing yourself lately and hasn't looked in the egg carton.

(8) Do you need me to get eggs as well? I used a couple in the naan dough earlier this week.

And, of course, you may need to point out obstacles that cannot be so easily corrected.

(9) Wait a minute—Kim still has our frittata pan.

All of these count as cooperative responses, and meet the obligations required of collaborators, because all engage with and further the process of joint activity that's underway.

Finally, consider the requirements on intentions in your own collaborative actions. It's not enough for you just to do something that helps the team. You have to do it in a way that allows your collaborators to identify the assumptions and goals that you have, and the contribution that you expect to make to the team. After all, your collaborators will be working to recognize and respond to your intentions, just as you have been working to recognize and respond to theirs.

Suppose, for example, that you have decided to assist your colleague at the party by opening the door. You should probably call attention to yourself as you do it, perhaps by calling out "I've got that for you", or perhaps just with a ceremonial flourish of your arm as you reach for the handle and push. That way, your colleague can see that the door will be open and can coordinate with you to go through the door. It's easy to imagine other ways of opening a door that wouldn't make things so easy. Your colleague might be in the course of turning around, expecting to back into the door to push it open, and keep holding the tray clear. You could then slip past and open the door undetected. Your colleague, of course, will keep backing up slowly, waiting in vain for the resistance of the door. At some point, there will be a crisis: your colleague may manage to readjust to an environment that doesn't match the plan, but may instead come off balance and fall.

Language generation also requires us to be explicit, and to make the plans behind our utterances recognizable. Let's return to (8). We can contrast the full explanation with other, briefer possible reminders, as in (10), (11) or (12).

(10) What about eggs?

(11) Do you need me to get eggs as well?

(12) I used a couple eggs in the naan dough last week.

All are grammatical utterances that might, with luck, prompt your spouse to reconsider whether there are enough eggs at home and ask for more if necessary. However, in some sense, they all require your spouse already to be alert to the possibility of the obstacle that you aim to point out. Your spouse has just asked you to get cheese but not eggs, and so presumably thinks there are enough. Because the reminders in (10) and (11) leave out your specific reasons why your spouse should reconsider, they might not work. On the other hand, with (12), you give those reasons, but you don't suggest a course of action. You force your spouse to work out the possibility that you could get the eggs, and to ask you to do so. Perhaps your spouse will instead decide to get the eggs some different and less efficient way—without even telling you about it. It seems that (8) may be the briefest possible way to reliably point out this obstacle to your spouse's plan. Brevity is at odds with the coordination required for effective teamwork.

The constraints of collaboration thus place real limits on how economical our interactions with one another can be. One way to get a handle on these limits is through Clark & Schaefer's (1989) PRINCIPLE OF JOINT CLOSURE. CLOSURE is Norman's (1988) term for an agent's judgment about whether a course of action has been completed and whether it has achieved its goals. We've seen that collaborators have to achieve joint closure by agreeing to their satisfaction about the outcome of everything they do. Their coordination depends on it. Clark and Schaefer argue that human collaborators regiment their joint activities to make it easier to achieve joint closure. In particular, people organize their activities into a hierarchy of tasks and subtasks. Their joint work involves an orderly navigation through this hierarchy. They proceed through subtasks one at a time, and make sure that they have achieved all the goals associated with each subtask before they move on to the next one. For Clark and Schaefer, the PRINCIPLE OF JOINT CLOSURE refers to this special discipline for agreeing about progress in collaboration.

The Principle of Joint Closure is, in a sense, a balancing force against overloading. According to the Principle of Joint Closure, once collaborators organize a task into a sequence of subgoals, they proceed with a sequence of more fine-grained interactions that bring about each of these subgoals in turn. On the face of it, this discipline does not allow for overloaded actions that are chosen to make contributions to multiple subgoals. The group is planning for and monitoring its progress on one subgoal at a time. Of course, the organization of tasks into parts and the selection of subgoals to achieve at each stage is a matter of coordination among the group. So it's more accurate to say that overloading action requires collaborators to adopt a distinctive conceptualization of their joint activity—one that specifically anticipates the kinds of efficiencies they hope to achieve together.

Consider, for example, carrying out a series of errands as a group. We've seen that overloading intentions is quite natural in this setting. You go to the pharmacy, then to the dry cleaners, then to the grocery store, in that order, partly because you have a set of tasks including picking up medication, laundry, and goat cheese, but partly because that tour makes for a particularly short tour of the places you need to visit. If you think of the activity simply as carrying out your errands, one after the other, you haven't left room to coordinate the efficiencies you can get by considering your goal of having a short tour as you choose each step of travel. Imagine you arrive at the pharmacy and it turns out to be closed unexpectedly. If you're really just working on getting your medications, it'll be natural for you to proceed to the nearest pharmacy and worry later about how to proceed with the rest of your tour. This is the sense in which the organization of joint activity and the Principle of Joint Closure provides a new constraint on the possibilities for overloading.

Overloading may be important enough and frequent enough in these tasks that you make room for it: This means conceptualizing the activity in a different way. At each step, your subgoal is choosing and carrying out an errand that you need to do and that makes progress on a short tour. As you coordinate steps in this new activity, you naturally track and assess the possibilities for overloading.

Suppose, to start, that the trip to the pharmacy is successful. You've organized the activity to track your progress running errands: You have the medication. But you've also organized the activity to explicitly address the progress you've made provisionally towards future errands. You're in a position to efficiently move on from here to the dry cleaners. Suppose by contrast that the trip to the pharmacy is a bust: the pharmacy is unexpectedly closed. Your new, overloaded task invites you to coordinate an overloaded response. You now need to find and carry out your next errand—given where you are and where you have to go—in a way that makes progress on a short tour. You might, for example, decide to go to the dry cleaners next now anyway, because you know of another pharmacy past the grocery store that you can visit more easily later in your tour.

In short, overloading in collaboration is subject to a constraint of COHERENCE. As our example shows, applying the Principle of Joint Closure for overloaded actions requires special acknowledgment of the possibility of efficient action in the organization of the task itself. Collaborators must conceptualize the multiple effects of overloaded actions as a recognizable and repeated pattern that establishes a new pattern of subgoals for the overall task. This conceptualization is needed to support the use of overloaded actions in collaboration. It primes interlocutors to recognize the multiple contributions of an overloaded action to their joint activity; it gives them the conceptual vocabulary to track the effects of an overloaded action; and it prompts them to consider possible overloading when they replan and accommodate one another's actions.

Coherence constraints are particularly significant for communicative intentions because we normally don't explain our communicative strategies explicitly. Doing so can easily cancel out any savings we get from efficient communication itself. Anyway, there is no guarantee that explaining things in full will actually work. We face the danger of a regress, where we might have to work indefinitely to disambiguate language we have used in attempts to disambiguate. Since we can only hint at the principles that organize our communicative intentions, we must work harder to make them comprehensible to start with. Thus, the need for coherence imposes strong constraints on communicative intentions. Any overloading we do has to be predictable and recognizable.

4 Overloading and the limits of economy

The requirement of coherence helps make sense of the way we naturally organize discourse that seems to combine multiple goals. When we have multiple points that go together, and they can naturally be seen as making a single overarching argument, then it's quite easy to create brief utterances that merge that information together. Take (13).

(13) The book was long and boring.

This is an unproblematic case of overloading. The speaker here has a number of related ideas about the book to get across. An initial formulation—*the book was long*, let's say—conveys key bits of this information. But it also afford the opportunity to fold in additional material—*and boring*—which exploits the existing syntactic structure and conveys a related idea. We readily recognize this material as developing a coherent picture in two parts. The length of a book is often an indicator of the excessive detail or repetitive style that leads to boredom.

On the other hand, when we have information that goes together, syntactically and perhaps conceptually, but the connections are unexpected or difficult for interlocutors to recognize, we prefer to flag those connections explicitly. We see this, for example, in the preference for (15) over (14).

(14) The book was short and boring.

(15) The book was short but boring nonetheless.

A speaker might well be interested in saying of a short, boring book that it had those properties. But it seems there's no coherent subtask that presents just these two ideas. The reader wonders what the point is of the attempted overloading of (14). Is there some unexpected reason to expect short books to be boring?

We recognize a different organization to the speaker's argument in (15). The speaker chooses to draw a contrast, and signals it explicitly with *but* and *nonetheless*. Given contrast as an organizing principle, we can see how the different ingredients of the speaker's utterance contribute to the speaker's overall plan. But because the speaker has chosen to mark the organization by an explicit signal, the utterance can be only so economical.

We find similar constraints when we combine speech with gesture or with practical embodied action. For example, as we developed the analysis of gesture in our (2009), Alex Lascarides observed the contrast between cases like (16) and (17).

(16) The cable is elastic.
The speaker repeatedly brings two fists together and separates them apart horizontally.

(17) The cable runs horizontally.
The speaker repeatedly brings two fists together and separates them apart horizontally.

The gesture in (16) seems to depict the potential of the cable to stretch, while the same gesture in (17) seems to depict the layout of the cable in space. A speaker might hope to overload a description of an elastic cable in a horizontal configuration by combining bare-bones fragments of speech and gesture, but it won't work. These messages don't have a coherent organization, and the audience naturally appeals to coherence in understanding what the gesture contributes.

In retrospect, then, we can interpret the examples with which we opened this chapter in light of the constraints of coherence. In (1), repeated as (18) below, there is a natural way in which all the information fits together.

(18) Take the rabbit out of the hat.

The utterance offers a description of the next action that completely identifies its parameters. Identifying the action itself provides a natural conceptual framework to recognize the different contributions of the different components of the speaker's overloaded communicative intention. In the description of the Galilean experiment in (5), the speech and gesture together introduce a hypothetical condition where a scientist drops ten identical balls from a level height. In the Lego assembly of (6), speech and actions combine in the natural subtask of getting and checking a piece that matches the director's instructions. These natural summaries speak to the coherence that supports the economy of these examples, and reinforce the overall message of this chapter. Communicative economy depends on imposing a coherent organization on a set of contributions to conversation, and finding opportunities within this organization to deploy communicative actions efficiently in pursuit of multiple goals.

References

Agre, Philip E. 1997. *Computation and human experience*. Cambridge, UK: Cambridge University Press.

Allen, James F. & C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence* 15. 143–178.

Appelt, Douglas. 1985. *Planning English sentences*. Cambridge, UK: Cambridge University Press.

Bratman, Michael E. 1987. *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Cadilhac, Anaïs, Nicholas Asher, Farah Benamara & Alex Lascarides. 2011. Commitments to preferences in dialogue. In *Proceedings of the sigdial 2011 conference*, 204–215. Association for Computational Linguistics.

Cassell, Justine, Matthew Stone & Hao Yan. 2000. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the first international conference on natural language generation (INLG)*, 171–178.

Clark, Herbert H. & Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50. 62–81.

Clark, Herbert H. & Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science* 13. 259–294.

Cohen, Philip R. & Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence* 42. 213–261.

Cohen, Philip R. & Hector J. Levesque. 1991. Teamwork. *Nous* 24(4). 487–512.

Cohen, Philip R. & C. Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science* 3(3). 177–212.

Di Eugenio, Barbara & Bonnie Webber. 1996. Pragmatic overloading in natural language instructions. *Internationl Journal of Expert Systems* 9(2). 53–84.

Grice, H. P. 1957. Meaning. *The Philosophical Review* 66(3). 377–388.

Grice, H. P. 1975. Logic and conversation. In P. Cole & J. Morgan (eds.), *Syntax and semantics III: Speech acts*, 41–58. New York: Academic Press.

Hayes-Roth, Barbara & Frederick Hayes-Roth. 1979. A cognitive model of planning. *Cognitive Science* 3. 275–310.

Koller, Alexander & Matthew Stone. 2007. Sentence generation as a planning problem. In *Proceedings of the 45th annual meeting of the association for computational linguistics*, 336–343.

Kripke, Saul. 1977. Speaker's reference and semantic reference. In Peter A. French, Thomas Uehling, Jr. & Howard K. Wettstein (eds.), *Midwest studies in philosophy, volume II*, 255–276. Minneapolis: University of Minnesota Press. Volume 2: Studies in Semantics.

Lascarides, Alex & Matthew Stone. 2009. Discourse coherence and gesture interpretation. *Gesture* 9(2). 147–180.

Lewis, David K. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Newell, Allen & Herbert A. Simon. 1961. *GPS, a program that simulates human thought*. Santa Monica, CA: Rand Corporation.

Norman, Donald. 1988. *The psychology of everyday things*. New York: Basic Books.

Perrault, C. Raymond & James F. Allen. 1980. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics* 6. 167–182.

Pollack, Martha. 1991. Overloading intentions for efficient practical reasoning. *Noûs* 25. 513–536.

Pollack, Martha E. 1990. Plans as complex mental attitudes. In Philip Cohen, Jerry Morgan & Martha Pollack (eds.), *Intentions in communication*, 77–103. Cambridge, MA: MIT Press.

Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*.

Cambridge, UK: Cambridge University Press.

Searle, John R. 1975. Indirect speech acts. In P. Cole & J. Morgan (eds.), *Syntax and semantics iii: Speech acts*, 59–82. New York: Academic Press.

Stalnaker, Robert. 1973. Presuppositions. *Journal of Philosophical Logic* 2(4). 447–457.

Stone, Matthew. 2000. Towards a computational account of knowledge, action and inference in instructions. *Journal of Language and Computation* 1. 231–246.

Stone, Matthew. 2004. Communicative intentions and conversational processes in human-human and human-computer dialogue. In John Trueswell & Michael Tanenhaus (eds.), *Approaches to studying world-situated language use*, 39–70. MIT.

Stone, Matthew, Christine Doran, Bonnie Webber, Tonia Bleam & Martha Palmer. 2003. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence* 19(4). 311–381.

Thomason, Richmond H. 1990. Accommodation, meaning and implicature: interdisciplinary foundations for pragmatics. In Philip R. Cohen, Jerry Morgan & Martha E. Pollack (eds.), *Intentions in communication*, 325–363. Cambridge, MA: MIT Press.

Thomason, Richmond H., Matthew Stone & David DeVault. 2006. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. Target Manuscript for Workshop on Presupposition Accommodation, held at The Ohio State University.

Notes

[1]Preparation of this chapter was supported in part by NSF IIS-1017811 and in part by the Leverhulme Network on Context and Communication. Thanks to Laurence Goldstein, Alex Lascarides, Luisa Marti, Jeff Pelletier, and audiences at Leverhulme Workshops 3 and 4.

[2]There is a tradition of doing just this going back to the work of Allen, Cohen and Perrault (Cohen & Perrault 1979, Allen & Perrault 1980, Perrault & Allen 1980). In particular, Appelt (1985) used Kripke's distinction between semantic reference and speaker reference to extend this to the pragmatics of individual constituents within sentences.

[3] Space precludes a complete presentation of the formal details. In fact, there are substantial technical challenges involved—particularly the problem of representing the compositional structure and meaning of utterances in an action formalism, and the problem of reasoning about interlocutors' changing information as the conversation unfolds. Stone, Doran, Webber, Bleam & Palmer (2003) describe an implemented model of linguistic grammar, utterance meaning, and interpretation in context, and show how the formalism accounts for a wide range of efficient and natural utterances. Stone (2000) shows how to reason about the evolving context in this framework, focusing on the hypothetical reasoning that may be required to assess whether a provisional utterance contributes enough information to satisfy the system's communicative goals. Stone (2004) justifies the philosophical understanding of these formal structures as representing the communicative intentions of conversational agents. Koller & Stone (2007) formalize the close relationship between these models of sentence planning and other AI models of planning. Cassell,

Stone & Yan (2000) sketch a model that extends the approach to simple cases of embodied communicative actions. Developing formal details to handle cases like (5) and (6) remains a project for future work.