HYPOTHESIS AND THEORY ARTICLE
published: 01 November 2011
doi: 10.3389/fpsyg.2011.00275

A cognitive architecture for the coordination of utterances

Chiara Gambi * and Martin J. Pickering

Department of Psychology, The University of Edinburgh, Edinburgh, UK

Edited by:
Christoph Scheepers, University of Glasgow, UK

Reviewed by:
Giovanni Pezzulo, National Research Council of Italy, Italy
Simon Garrod, University of Glasgow, UK

***Correspondence:**
Chiara Gambi, Department of Psychology, The University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK.
e-mail: c.gambi@sms.ed.ac.uk

Dialog partners coordinate with each other to reach a common goal. The analogy with other joint activities has sparked interesting observations (e.g., about the norms governing turn-taking) and has informed studies of linguistic alignment in dialog. However, the parallels between language and action have not been fully explored, especially with regard to the mechanisms that support moment-by-moment coordination during language use in conversation. We review the literature on joint actions to show (i) what sorts of mechanisms allow coordination and (ii) which types of experimental paradigms can be informative of the nature of such mechanisms. Regarding (i), there is converging evidence that the actions of others can be represented in the same format as one's own actions. Furthermore, the predicted actions of others are taken into account in the planning of one's own actions. Similarly, we propose that interlocutors are able to coordinate their acts of production because they can represent their partner's utterances. They can then use these representations to build predictions, which they take into account when planning self-generated utterances. Regarding (ii), we propose a new methodology to study interactive language. Psycholinguistic tasks that have traditionally been used to study individual language production are distributed across two participants, who either produce two utterances simultaneously or complete each other's utterances.

Keywords: coordination, joint action, prediction, shared representations

INTRODUCTION

The interactive use of language in conversation is a form of joint activity, in which individuals act together to achieve the common goal of communicative success. Clark (1996, 2002) proposed that conversation shares fundamental features with other joint activities, for example waltzing, playing a duet, or shaking hands. The most central, defining feature of all joint activities is coordination: the mutual process by which actors take into account the intentions and the (performed or to-be-performed) actions of their partners in the planning and performance of their own actions (Clark, 1996, pp. 61–62). Clark regards the process by which individual actors manage to coordinate to be a form of problem solving, and his focus is on "strategies" that they use to attain coordination. Despite the recognition that co-actors need to coordinate both on content (the common intended goal) and on processes ("the physical and mental systems they recruit in carrying out those intentions"; Clark, 1996, p. 59) to succeed in a joint action, very little is in fact said about such processes. To illustrate this point, we look at two aspects of coordination in language use: the synchronization of the processes of production and comprehension, and turn-taking.

First, production and comprehension never occur in isolation, but the speaker's act of production unfolds while the listener comprehends it. In order to reach mutual understanding, they need to process linguistic (and non-linguistic) signals as they occur, while monitoring for errors and misunderstandings, and usually compensating for a fair amount of noise present in the environment. Clark (1996, 2002) argued that speaker and listener synchronize their acts of production and comprehension by striving to comply with principles such as the continuity principle, which states that constituents should be produced fluently whenever possible (Clark and Wasow, 1998). When they have to deviate from these principles, they follow conventional strategies to help their listeners by signaling that one of the principles is being violated. For example, Clark (2002) assumes that speakers produce certain types of disfluencies to inform listeners that they are violating the continuity principle. But he is silent on the mechanisms that normally allow synchronization, merely pointing out that the listener needs to attend to a speaker's productions.

Second, speakers and listeners take turns by repeatedly switching roles in the conversation. This alternation is managed "on the fly" by the participants themselves, at least in informal conversations (Sacks et al., 1974; Clark, 1996). Transitions are so smooth that the average gap between turns ranges from approximately 0 ms to around 500 ms (De Ruiter et al., 2006; Stivers et al., 2009), depending on language and culture. This tight temporal coordination is coupled with coordination at the pragmatic level, since each contribution normally constitutes an appropriate response to a previous contribution by the other speaker. Coordination is thought to result from the application of a set of norms, which govern turn transitions and state who can claim the ground and when (Sacks et al., 1974). It is also recognized that the listener anticipates the end of the speaker's turn (Sacks et al., 1974; Clark, 1996, 2002); additionally, the listener starts planning her utterance in advance, while the previous speaker's turn is still unfolding.

A widespread claim in the literature on turn-taking is that speakers help their addressees by signaling whether they want to keep the floor or are about to end their turn (Clark, 2002). Many linguistic (e.g., pitch contour) and non-linguistic (e.g., breathing) cues are reliably associated with turn-holding or turn-yielding points in a conversation. However, very few studies have systematically investigated which features of the speech signal are actually exploited by listeners to discriminate between end-of-turn and turn-holding points (see Gravano and Hirschberg, 2011; Hjalmarsson, 2011) and even fewer studies have looked at listeners' ability to use such cues on-line to anticipate turn endings (Grosjean and Hirt, 1996; De Ruiter et al., 2006; Magyari and De Ruiter, 2008). Moreover, no mechanisms have been proposed to explain how listeners can simultaneously comprehend what the speaker is saying, use the available cues to predict when the speaker's turn is going to end, and prepare their own contribution.

Another important approach to conversation as a joint activity has developed the study of coordination from a quite different perspective. Two conversational partners tend to unconsciously coordinate their body postures (Shockley et al., 2003) and gaze patterns (e.g., Richardson and Dale, 2005; see Shockley et al., 2009). One way of explaining such findings is based on the properties of oscillators, systems characterized by a periodic cycle. Mechanical oscillators (e.g., pendulums) tend to spontaneously attune their cycles, so that they become entrained: their cycles come into phase (or anti-phase). Neural populations firing at certain frequencies might act as oscillators, and sensory information regarding the phase of another oscillator (e.g., in another human body) could serve to fine-tune them. The entrainment of oscillators is therefore an automatic coordinative mechanism. According to this account, coordination, in the form of synchronization, emerges from the interaction of two dynamic systems, without any need for intentions. This view therefore suggests that coordination need not be goal-directed (Richardson et al., 2005; Shockley et al., 2009; Riley et al., 2011).

The entrainment of oscillators might explain the remarkable timing skills shown by language users. Wilson and Wilson (2005) proposed that such entrainment accounts for speakers' ability to avoid gaps or overlaps in conversation. In their account, the production system of a speaker oscillates with a syllabic phase: the readiness to initiate a new syllable is at a minimum in the middle of a syllable and peaks half a cycle after syllable offset. They argued the interlocutors converge on the same syllable rate, but their production systems are in anti-phase, so that the speaker's readiness to speak is at minimum when the listener's is at a maximum, and *vice versa.* Cummins (2003, 2009) found that two people can read the same text aloud with almost perfect synchrony; his participants only reviewed the text once and, even without any practice, could easily maintain average lags as short as 40–60 ms (Cummins, 2003). This timing is impressive, considering the huge amount of variability in speech, even within one speaker. Cummins (2009) tentatively suggested that the production systems of synchronous readers become entrained.

However, the oscillator model cannot fully explain turn-taking. First, regularities in speech appear to take place over very short time-scales, with the cyclic pattern of syllables that Wilson and Wilson (2005) propose as the basis for entrainment occurring at 100–150 ms. If predictions were made on the basis of syllable-level information alone, there would simply be not enough time to prepare the next contribution and leave a 0-ms gap. Anticipation of the end of a turn, instead, must draw on information that spans units larger than the syllable. Thus there must be additional mechanisms underlying coordination between interlocutors. In addition, Wilson and Wilson's account cannot explain how entrainment of oscillators might lead to mutual understanding.

More generally, accounts within this framework can only explain instances of rhythmic, highly repetitive activities. As such, they have no explanation for the pragmatic link between two complementary actions, be they turns in a conversation or the acts of handing over a mug and pouring coffee in it. Consider, for example, how answers complement questions. For an addressee to produce an appropriate answer, it is not enough to talk in anti-phase with the speaker. She must be able to plan in advance not only *when* to start speaking, but also *what* to say (Sebanz and Knoblich, 2009; Vesper et al., 2010).

Clark's (1996, 2002) approach and the entrainment of oscillators clearly deal with separate levels of analysis. Clark describes the dynamics of coordination at what we might call the "intentional" level. Interlocutors coordinate by making inferences about the intentions underlying their partners' behavior. Ultimately, coordination is successful if they develop mutual beliefs about their intentions. In this, they are helped by the existence of conventions (e.g., turn-allocation norms) that map intentions onto behavior. On the other hand, the entrainment-of-oscillators approach focuses on the behavioral patterns exhibited by two coordinating systems. It maintains that very general physical principles can explain the emergence of such patterns. Importantly, recent reviews (Knoblich et al., 2011) and computational accounts (Pezzulo and Dindo, 2011) have emphasized that successful joint action is likely to require coordination at both a higher level (intentions) and a lower level (bodily movements). We argue that one needs an intermediate level of analysis. In essence, it is at this level that one can define a cognitive architecture for coordination. This should comprise a set of mechanisms (representations and processes acting on those representations) that underlie coordination and, ultimately, mutual understanding between interlocutors.

In this paper, we propose that the most promising way of identifying these mechanisms stems from a mechanistic account of language processing. This is of course what psycholinguistic theories have traditionally tried to develop. However, most of these theories are concerned with monolog, in which speakers and listeners act in isolation. Pickering and Garrod (2004) pointed out the need for a theory of dialog that can explain the seemingly effortless, automatic nature of conversation. They proposed that interlocutors come to a mutual understanding via a process of alignment, whereby their representational states tend to converge during the course of a conversation. Alignment occurs at many different levels, including words and semantics (Garrod and Anderson, 1987), syntax (Branigan et al., 2000), and ultimately the situation model. Importantly, they argued that the simple mechanism of priming (i.e., facilitation in processing of an item due to having just processed the same or a related item) underlies such alignment. Alignment facilitates coordination (i.e., similar representational states facilitate successful interaction). In their model, therefore, coordination among interlocutors results from a mechanism of priming that is known to operate within the individual speaker's production system and the individual listener's comprehension system.

To account for alignment between speaker and listener, Pickering and Garrod (2004) assumed representational parity between production and comprehension. Menenti et al. (2011) recently provided evidence for this assumption in an fMRI study, showing that brain areas that support semantic, lexical, and syntactic processing are largely shared between language production and language comprehension. In another fMRI study, Stephens et al. (2010) compared activation in a speaker with activation in listeners attending to the speech produced by that speaker. The speaker's and the listeners' neural activity were not only spatially overlapping, but also temporally coupled. As might be expected, areas of the listeners' brains were typically activated with some delay relative to the corresponding areas of the speaker's brain. However, some areas showed the opposite pattern: they were activated in the listener's brain before they were in the speaker's. These areas might be responsible for anticipatory processing of the sort that seems to be necessary for coordination. The size of areas showing anticipatory activity was positively correlated with listeners' comprehension performance. Interestingly, Noordzij et al. (2009) also found extensive overlap when comparing the planning and recognition of non-conventional communicative actions (e.g., moving a token to communicate its goal position on a game board). If the production and comprehension systems make use of the same representations, those representations that have just been built in comprehension can be used again in production and *vice versa.* Because interlocutors alternate between production and comprehension, their production and comprehension systems become increasingly attuned.

However, it is not certain that representational parity can by itself account for coordination in dialog. In addition to a common format for the representation of self-generated and other-generated actions (Sebanz et al., 2006a), addressees need to predict speakers' utterances (Pickering and Garrod, 2007) and make use of these predictions when producing their own utterances (Garrod and Pickering, 2009). To show this, the next section first reviews evidence that representational parity holds between perception and action. We show how perception–action links can serve as a basis for prediction of others' actions and explain how these predictions can in turn affect the planning of one's own actions. Then we apply these ideas specifically to the coordination of utterances.

As well as outlining a theoretical framework, we describe some experimental paradigms that can help answer the questions raised by this new approach. In fact, we believe that the inadequacy of the current accounts is partly due to the limitations associated with current experimental studies of dialog. These studies have traditionally looked at how coordination is achieved off-line, over quite long stretches of conversation, using measures such as changes in turn length or choice of referring expressions. Under these circumstances, time constraints are loose enough to allow for relatively slow and intentional cognitive processes to be the basis of coordination (e.g., Clark and Wilkes-Gibbs, 1986; Wilkes-Gibbs and Clark, 1992). Studies that focus on alignment have reduced the time-scale to consecutive utterances. Garrod and Anderson (1987), for example, analyzed the spatial descriptions produced during a co-operative maze game. They showed that interlocutors align locally on the method of description that they use to refer to locations in the maze. Studies of priming in dialog have systematically investigated this utterance-to-utterance alignment. Thus, Branigan et al. (2000) had participants alternate in the description of pictures and found that the addressee tends to re-use the syntactic structure of the description produced by the current speaker, in the following turn. However, this is still a relatively long time-scale.

In contrast, no study has looked at that moment-by-moment coordination that might explain how listeners and speakers synchronize and take turns with virtually no gap or overlap. We argue that the obvious way to do this would be to conduct experiments with more than one participant in which the relative timing of their contributions is carefully controlled and the relationship between their utterances is systematically varied. We would then be able to test whether aspects of others' utterances are indeed predicted and to what extent such predictions are taken into account when planning one's own utterances. Importantly, these experiments should focus on the study of mechanistic processes (rather than intentional behavior), and should in this respect be similar to the psycholinguistics of monolog.

REPRESENTING ANOTHER'S ACTIONS

The behavioral and neuroscientific literature on joint actions has investigated how actions performed by a co-actor are taken into account in the planning and performance of one's own actions (Sebanz et al., 2006a; Sebanz and Knoblich, 2009). Sebanz and colleagues have argued that acting together requires shared representations. This means that people should represent other people's actions alongside their own. In a series of experiments, they demonstrated that such representations are indeed formed and activated automatically, even when they are not relevant for one's own actions because the two participants are merely acting next to each other on alternating trials (as opposed to acting together to reach a common goal; Sebanz et al., 2003, 2005; see also Atmaca et al., 2008; Vlainic et al., 2010).

For example, when one participant is instructed to respond to red stimuli with right button presses and the other responds to green stimuli with left button presses (joint condition), reaction times are slower when the stimulus and the response are spatially incongruent (e.g., the red stimulus points to the left) than when they are congruent. A similar interference effect arises when a single participant is in charge of both responses (individual condition; Sebanz et al., 2003, 2005). In the individual condition, the irrelevant spatial feature of the stimulus automatically activates the spatially congruent response, which is part of the participant's response set. In the joint condition, there is only one response in each participant's response set. However, the partner's task is represented as well; the presentation of a leftward-pointing stimulus automatically evokes the partner's response (left button press) as well as one's own (right button press), yielding interference. Additionally, electrophysiological evidence suggests that the action associated with the partner's task is inhibited on no-go trials (Sebanz et al., 2006b). In these experiments, knowledge about the partner's task is available from the start (i.e., both participants listen while task instructions for each co-actor are given) and can be used to predict the partner's action response even when there is no sensory feedback from the other's actions (Atmaca et al., 2008; Vlainic et al., 2010); seeing the associated stimulus is enough to activate the appropriate response (Sebanz et al., 2006a).

When knowledge about others' actions is not available as part of a task specification, the mere observation of actions performed by others can still lead to the formation of shared representations (Sebanz et al., 2006a). More precisely, the action system might be involved in action observation. At least two lines of evidence support this claim. First, observing an action that is incompatible with a planned action affects execution of that action (e.g., Brass et al., 2000; see Wilson and Knoblich, 2005); second, areas of the motor system involved in action planning are activated during passive observation of the same actions (e.g., Iacoboni et al., 1999; see Rizzolatti and Craighero, 2004 for a review). This suggests that observed actions are coded in the same format as one's own actions (Prinz, 1997; Sebanz et al., 2006a).

Many researchers agree that motor involvement in action perception can aid action understanding (e.g., Blakemore and Decety, 2001; Buccino et al., 2004). Wilson and Knoblich (2005) proposed that action perception involves *covert imitation* of others' actions, as the perceiver internally simulates the observed action in her own motor system. The simulation is quicker than the actual performance of an action. Therefore, it can also be used to formulate perceptual predictions about what the observed actor is going to do next. Such predictions allow rapid and effective interpretation of the observed movement, even in cases where the movement needs to be partially reconstructed, because perceptual information is missing (predictions would serve to "fill in the gaps"). In addition, covert imitation of the partner in a joint activity could underlie quick and appropriate reactions to his or her actions (Wilson and Knoblich, 2005, p. 468).

More specifically, Wilson and Knoblich (2005) proposed that covert imitation of others is based on a model of one's own body (cf. Grush, 2004). Though this model can be adjusted to accommodate differences between the observer's and the actor's bodies, it follows that simulation (and hence prediction) of one's own actions should be more accurate than simulation of actions performed by others. In support of this claim, people are better at predicting a movement trajectory (e.g., in dart-throwing or handwriting) when watching a video of themselves vs. others (Knoblich and Flach, 2001; Knoblich et al., 2002) and pianists find it easier to synchronize with a recording of themselves than with a recording of somebody else (Keller et al., 2007).

The model that computes predictions is specifically a forward model (Wilson and Knoblich, 2005). It takes a copy of the motor command sent to the body as input and produces the expected sensory feedback as output. Expected sensory consequences of executing a motor command (e.g., expected limb position) can then be compared with actual feedback coming from the sensory system. This mechanism allows for fast, on-line control of movements (Wolpert and Flanagan, 2001). If the actual position of a limb, for example, does not match the predicted position, adjustments can be made to the motor command to minimize the difference. When the forward model is run, activation of the motor system normally ensues. However, when the forward model is used to covertly imitate another actor, covert imitation does not always result in overt imitation of another's movements. It is likely that the overt motor response is suppressed in such cases (Grush, 2004; Sebanz et al., 2006b).

Finally, and again following Sebanz et al. (2006a), we note that representing the actions performed by others and predicting what they are going to do are necessary but not sufficient for on-line coordination. What is also required is a mechanism for integrating self-generated and other-generated actions in real time. If individual actions are coordinated to the partner's actions on a moment-by-moment basis, then other-generated actions must be considered during planning of one's own actions. In support of this, Knoblich and Jordan (2003) had participants coordinate button presses that caused a circular stimulus to accelerate either to the right or to the left (with each participant being in charge of one direction) so that the stimulus remained aligned with a moving dot. Provided that feedback about the other's actions was available, participants mastered the task as successfully as participants acting alone. In particular, they learned to jointly anticipate sudden changes in the dot's movement direction.

The authors concluded that the participants were predicting the consequences of integrating their own and their partner's actions and suggested two mechanisms that could underlie this ability. Participants might run multiple simulations corresponding to the combination of the various action alternatives available to themselves and their partners (cf. Wilson and Knoblich, 2005). The other alternative, which they favored (Knoblich and Jordan, 2003; Sebanz and Knoblich, 2009), is based on the distal coding theory (Prinz, 1997; Hommel et al., 2001), which states that actions are coded in terms of the events resulting from them. Integration of self- and other-generated actions could occur at the level of these distal events. Rather than building and constantly updating a simulation of other-generated actions, then, people would simply take into account the perceptual consequences of others' actions (the events potentially resulting from them), in the same way as they would take into account other aspects of the environment (e.g., the presence of obstacles; cf. Sebanz and Knoblich, 2009, p. 361). One would then adjust one's own action plan accordingly, so that the intended event (corresponding to the joint action goal) is realized.

To summarize, the shared representational approach maintains that (i) other-generated and self-generated actions are represented in the same format, (ii) representations of other-generated actions can be used to drive predictions, and (iii) self-generated and other-generated actions are integrated in real time to achieve coordination (Sebanz et al., 2006a). By referring to representations and processes that make use of those representations, the account provides explanations at a level that bridges purely intentional and purely mechanistic accounts of coordination. Despite the above-mentioned limitations (see Introduction), entrainment of oscillators could still play an important role in coordination. In particular, it could serve as a basis to optimize other mechanisms (Vesper et al., 2010). Recall that covert imitation of other-generated actions is assumed to exploit a model of one's own body. If some basic properties of this system, such as the frequency of rhythmic unintentional movements, become attuned *via* entrainment, then simulations of another's actions would likely become more accurate, because the simulated system will end up sharing features of the system on which simulations are based. In accord with this view, co-actors that rocked chairs in synchrony were faster at jointly moving a ball through a labyrinth (Valdesolo et al., 2010). Therefore, entrainment with another actor can enhance performance on a subsequent, unrelated joint task. Entrained actors did feel more similar to each other and more connected, but these feelings did not predict performance. Instead, enhancement appeared to be mediated by increased perceptual sensitivity to each other's actions (Valdesolo et al., 2010).

REPRESENTING ANOTHER'S UTTERANCES

In this section, we propose that interlocutors also coordinate via three mechanisms: (i) they represent others' utterances in a similar format as their own utterances; (ii) they use these representations as a basis for prediction; and (iii) they integrate self- and other-representations on-line. Interestingly, there is plenty of evidence for a direct link between speech perception and speech production (Scott et al., 2009). Fowler et al. (2003) showed that people are faster at producing a syllable in response to hearing the same syllable than in response to a tone; in fact, shadowing a syllable yielded response latencies that were nearly as fast as those found when the to-be-produced syllable was fixed and known in advance. Moreover, Kerzel and Bekkering (2000) demonstrated an action perception compatibility effect for speech (due to a task-irrelevant stimulus). They found that participants pronounced a printed syllable while watching a video of a mouth producing the same syllable more quickly than when the mouth produced a different syllable. While the first study involves intentional imitation, the second one provides more compelling evidence for automaticity. However, they both deal with cases of *overt* imitation, where there is an overt motor response. Evidence that bears more on the issue of *covert* imitation comes from neuropsychological studies of speech perception. These studies found activation of motor areas during passive listening to speech (e.g., Wilson et al., 2004), showed that this activation is articulator-specific (Pulvermüller et al., 2006), and found that stimulation of motor areas with TMS can influence speech perception (Meister et al., 2007; D'Ausilio et al., 2009; see Pulvermüller and Fadiga, 2010).

In addition, some researchers have proposed that activation of motor areas during speech perception might reflect the dynamics of forward models. In Guenther and colleagues' model of speech production, a forward model is used to compute the auditory representation corresponding to the current shape of the vocal tract, which in turn is derived from combined proprioceptive feedback and a copy of the motor command sent to the articulators (e.g., Guenther et al., 2006). In an MEG study, Tian and Poeppel (2010) demonstrated that auditory cortex is activated very quickly (around 170 ms) when participants are asked to imagine themselves articulating a syllable. They therefore proposed that forward models involved in speech production can be decoupled from the movement of the articulators. Their findings open up the possibility that a forward model of the articulation system could be used in covert imitation of perceived speech.

Activation of motor areas during speech perception could serve a variety of purposes. First, it could help understanding, just as it may for other actions (see Representing Another's Actions). In support of this, overt imitation of an unfamiliar accent (which must of course involve activation of such areas) improves accent comprehension more than mere listening (Adank et al., 2010). Alternatively, it could reflect articulatory rehearsal in the verbal working memory system (Wilson, 2001). Scott et al. (2009) suggested that motoric activation during speech perception might also facilitate coordination between language users in dialog. In particular, they proposed that the activation of the motor system underlies synchronization of the rhythmic properties of speech (entrainment). Our proposal differs in that we claim that it could also be responsible for the covert imitation, and prediction, of others' utterances (Pickering and Garrod, 2007).

WHAT KIND OF INFORMATION IS REPRESENTED?

Consider two speakers, *A* (female) and *B* (male), producing two utterances roughly at the same time, in response to a shared stimulus, such as a to-be-named picture of a kite. **Figure 1** illustrates the range of information that *A* could represent about her own utterance (upper box) and about *B*'s utterance (lower box). Before we discuss the nature of these representations, we will briefly illustrate the time course of word production, taking *A*'s production of "kite" as an example (see the timeline at the top of **Figure 1**). Models of single word production (e.g., Levelt et al., 1999) involve at least (i) a semantic representation (sem_A) corresponding to the target concept (KITE); (ii) a syntactic representation (syn_A) – sometimes called a lemma – that incorporates syntactic information about the lexical item, such as that it is a noun ($kite_{(N)}$); (iii) a phonological representation (phon_A) that specifies a sequence of phonemes and its syllable structure (/kaIt/). Finally, the appropriate articulatory gestures are retrieved and executed (art_A).

Note that each processing level is characterized not only by the content of the associated representation, but also by its timing [$\text{t}(\text{sem}_A)$, $\text{t}(\text{syn}_A)$, etc.]. Some representations are typically ready before others and the processing stages take different amounts of time. Indefrey and Levelt (2004) derived indicative time windows from a meta-analysis of several word production experiments. Their estimates are also reported at the top of **Figure 1**, though the exact times might depend on the words used or the experimental conditions (cf. Sahin et al., 2009, for estimates based on intracranial electrophysiological recordings).

Now, consider the upper box of **Figure 1**. We assume that *A* can generate predictive estimates of the duration of each processing stage (indicated by $\hat{\text{t}}$ in **Figure 1**). For example, she might generate the estimate $\hat{\text{t}}\ (\text{syn}_A) \approx 250$ ms, meaning that she predicts retrieving the syntactic representation will take approximately 250 ms (from picture onset). These estimates can in turn be exploited by *A* to guide planning of her own utterance. Interestingly, some studies have shown that individual speakers can coordinate the production of two successive utterances so as to minimize disfluencies (Griffin, 2003; cf. Meyer et al., 2007). Similarly, Meyer et al. (2003) demonstrated that the amount of planning speakers perform before articulation onset can depend on the response time deadline they implicitly set for their performance at a naming task. This suggests that timing estimates are computed for one's own utterances and can be used to guide planning.

Clearly, for a speaker to be able to use the information provided by timing estimates effectively, the estimates must be ready before processing at the corresponding stages is completed. So, for instance, the estimate $\hat{\text{t}}\ (\text{syn}_A) \approx 250$ ms is useful only if it is available before syntactic processing is complete. This means that

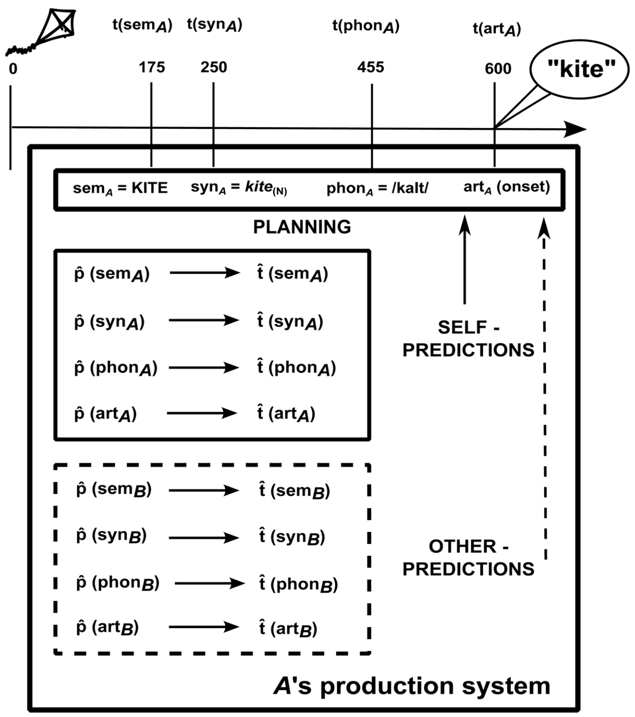

FIGURE 1 | Simultaneous production. A produces the word *kite* in response to the picture of a kite. sem_A, syn_A, phon_A are semantic, syntactic, and phonological representations for *A*'s utterance. $\text{t}(\text{sem}_A)$, $\text{t}(\text{syn}_A)$, $\text{t}(\text{phon}_A)$ indicate the actual time elapsed from picture onset (in ms) when processing is completed at each stage and the corresponding representation has been built (based on Indefrey and Levelt, 2004); $\text{t}(\text{art}_A)$ marks the onset of *A*'s utterance. $\hat{\text{t}}\ (\text{sem}_A)$, $\hat{\text{t}}\ (\text{syn}_A)$, $\hat{\text{t}}\ (\text{phon}_A)$, and $\hat{\text{t}}\ (\text{art}_A)$ are timing estimates computed by *A* for her own utterance. $\hat{\text{p}}\ (\text{sem}_A)$, $\hat{\text{p}}\ (\text{syn}_A)$, $\hat{\text{p}}\ (\text{phon}_A)$, $\hat{\text{p}}\ (\text{art}_A)$ are the content predictions for *A*'s own utterance, on which the timing estimates are based. *A* believes that *B* is speaking in response to the same picture. Dotted lines refer to representations of the other. $\hat{\text{t}}\ (\text{sem}_B)$, $\hat{\text{t}}\ (\text{syn}_B)$, $\hat{\text{t}}\ (\text{phon}_B)$, and $\hat{\text{t}}\ (\text{art}_B)$ are timing estimates computed by *A* for *B*'s utterance. $\hat{\text{p}}\ (\text{sem}_B)$, $\hat{\text{p}}\ (\text{syn}_B)$, $\hat{\text{p}}\ (\text{phon}_B)$, $\hat{\text{p}}\ (\text{art}_B)$ are *A*'s content predictions, at the various processing stages, for *B*'s utterance. Horizontal arrows [from $\hat{\text{p}}\ (\text{sem}_A)$ to $\hat{\text{t}}\ (\text{sem}_A)$, from $\hat{\text{p}}\ (\text{syn}_A)$ to $\hat{\text{t}}\ (\text{syn}_A)$, etc.] indicate that estimates of the timing at each level are based on content predictions at the same level. Timing estimates at one level could also be directly based on content estimates at other levels, but we ignore this here for simplicity. Vertical arrows from self- and other-predictions to planning represent the integration stage.

the estimates are predictions. What are such predictions based on? Importantly, in language production, timing aspects are known to be closely related to the content of the computed representations. For example, word frequency affects $\text{t}(\text{phon}_A)$, with phonological retrieval being slower for less frequent word forms (e.g., Caramazza et al., 2001). We therefore assume that *A* predicts aspects of the content of sem_A, syn_A, and phon_A. In other words, the speaker anticipates aspects of the semantics, syntax, and phonology of the utterance she is about to produce, before the representations corresponding to each level are built in the course of the production process itself. To distinguish these predictions that relate to content from predictions that relate to timing (i.e., the timing estimates), we label them $\hat{\text{p}}\ (\text{sem}_A)$, $\hat{\text{p}}\ (\text{syn}_A)$, $\hat{\text{p}}\ (\text{phon}_A)$, and $\hat{\text{p}}\ (\text{art}_A)$.

There is much evidence that content predictions of the sort we are assuming for production are indeed formulated by readers and listeners during comprehension. For example a series of sentence comprehension studies showed that predictions are made at the syntactic (lemma) level, in relation to syntactic category (e.g., Staub and Clifton, 2006) and gender (e.g., Van Berkum et al., 2005), and at the phonological level (e.g., DeLong et al., 2005; Vissers et al., 2006). For a review of some of this evidence, see Pickering and Garrod (2007), who also argued that such predictions rely on production processes. Note, however, that timing and content predictions for self-generated utterances need not always be as detailed as these studies may suggest. The specificity of predictions might depend on task demands (e.g., whether fine-grained control over the production process is needed) and be highly variable.

Having posited that predictions of timing and content can be generated for one's own utterances, we now propose that representing others' utterances can also involve the computation of predictions, and that those predictions are in a similar format to the timing and content predictions for self-generated utterances. The lower (dashed) box in **Figure 1** shows the range of information that *A* could represent about *B*'s utterance. Importantly, *A* may well not represent all of this information under all circumstances. Later, we describe experimental paradigms that can investigate the conditions under which aspects of *B*'s utterance are represented and how. Here, our aim is to provide a comprehensive framework in which such questions can be addressed.

First of all, *A* could estimate the time course of *B*'s production. Minimally, *A* could compute $\hat{\text{t}}\ (\text{art}_B)$, an estimate of *B*'s speech onset latency. In addition, *A* might compute timing estimates for the different processing stages, from semantics to phonology [$\hat{\text{t}}\ (\text{sem}_B)$, $\hat{\text{t}}\ (\text{syn}_B)$, $\hat{\text{t}}\ (\text{phon}_B)$, and $\hat{\text{t}}\ (\text{art}_B)$ in **Figure 1**], just as she does when anticipating the timing of her own productions. As timing estimates are likely to be based on information regarding the content of the computed representations, we suggest that *A* can also represent the content of *B*'s utterance. In particular, *A* builds predictive representations of the semantics, syntax, and phonology of the utterance produced by *B* [$\hat{\text{p}}\ (\text{sem}_B)$, $\hat{\text{p}}\ (\text{syn}_B)$, $\hat{\text{p}}\ (\text{phon}_B)$, and $\hat{\text{p}}\ (\text{art}_B)$ in **Figure 1**].

THE NATURE OF THE REPRESENTATION OF THE OTHER

We have just proposed that other-generated utterances can be represented in a format that is similar to that of content ($\hat{\text{p}}$) and timing ($\hat{\text{t}}$) predictions for self-generated utterances. How are such predictions computed? We propose that people can make content and timing predictions, for both self-generated and other-generated utterances, using forward models of their own production system. This, in essence, amounts to an extension of the covert imitation account (Wilson and Knoblich, 2005) to language. Pickering and Garrod (submitted) provide a detailed theory that incorporates these claims (see also Pickering and Garrod, 2007; Garrod and Pickering, 2009).

The model is primarily used in the planning and control of one's own acts (here, speech production acts), but it can be used to simulate the production system of another speaker. When this happens, the model is decoupled from the production system, so that covertly simulating another's utterances does not lead to the actual planning of that utterance or to its articulation. In other words, *A* does not build sem_B, syn_B, and phon_B (semantic, syntactic, and phonological representations for the utterance that *B* is going to produce) just as she does not initiate art_B (the articulation stage for *B*'s utterance).

Nevertheless, speakers can overtly imitate a speaker (e.g., in speech shadowing; see Marslen-Wilson, 1973) and they sometimes complete each other's utterances (see Pickering and Garrod, 2004). On occasion, therefore, covert simulation of *B*'s utterance, *via* the computation of a forward model, results in activation of *A*'s own production system. In this case, there will be activation of the semantic (sem_B), syntactic (syn_B), and phonological (phon_B) representations corresponding to *B*'s to-be-produced utterance, within *A*'s production system. Depending on the predictability of *B*'s utterance, and on the speed of the simulation, *A* might end up shadowing *B*'s speech, talking in unison with *B* or even anticipating a completion for *B*'s utterance. Note, however, that some activation of *A*'s production system does not necessarily entail that *A* overtly articulates *B*'s utterance.

Note that this account differs slightly from the dominant view in the action and perception literature (e.g., Grush, 2004). According to this view, the motor system is in fact always activated following the activation of the forward model, but this activation is inhibited and therefore does not result in an overt motor response (though residual muscle activation can be detected in the periphery; e.g., Fadiga et al., 2002). The system responsible for language prediction might function in the same way as the system responsible for motor predictions. However, it is also possible that predicting *B*'s utterances does not involve any (detectable) activation flow in *A*'s language production system. At present, determining exactly under which conditions *A*'s production system is activated, and to what extent, is still a matter for empirical investigation. In the section on "Simultaneous Productions" we indicate which experimental outcomes are to be expected under the alternative hypotheses.

Another important issue relates to the accuracy of both the timing and content representations of another's utterances. For example, how similar is $\hat{p}$ (sem_B) to *B*'s concept KITE, or how accurate an estimate of *B*'s speech onset latency is $\hat{t}$ (art_B)? We expect representations of another's utterances to be generally somewhat inaccurate. First, although context and task instructions might highly constrain the productions of both speakers in experimental settings, normally *A* would have only limited information regarding what *B* intends to say. Second, *A* has limited experience of other speakers' production systems. The forward model she uses to compute predictive estimates is fine-tuned to her own production system rather than to *B*'s production system (Wolpert et al., 2003). As a consequence, timing estimates based on a model of *A*'s production system are likely to diverge from the actual time course of *B*'s production. The degree of error will also depend on how much *B* differs from *A* in speed of information processing. Conversely, we expect accuracy to increase the more *A*'s and *B*'s systems are or become similar (Wolpert et al., 2003). In conversations, the two systems might become increasingly attuned *via* alignment (Pickering and Garrod, 2004), thanks to priming channels between the production and comprehension systems of the two interlocutors. Furthermore, interlocutors' breathing patterns and speech rates can converge *via* entrainment (see Wilson and Wilson, 2005 and references therein).

Finally, we might ask whether predictions about other-generated utterances can influence the planning of one's own utterances to the same extent as predictions about self-generated utterances. For example, say that $\hat{t}$ (art_A) is a prediction of when *A* will finish articulating her current utterance. *A* should take this prediction into account as she plans when to start her next utterance. Similarly, if *B* is the current speaker and *A* wants to take the next turn, *A* could compute $\hat{t}$ (art_B), an estimate of when *B* will stop speaking. Then the question is, will *A* pay as much attention to $\hat{t}$ (art_B) as she would to $\hat{t}$ (art_A) in the first case? This is likely to depend on the circumstances. For example, $\hat{t}$ (art_B) might be weighted as less important if its degree of accuracy is low (i.e., previous predictions have proved to be wrong). Alternatively, *A* might not take $\hat{t}$ (art_B) into account, simply because she does not share a goal with *B*; for example, she might be trying hard to be rude and interrupt *B* as much as possible.

THE TIME COURSE OF PLANNING, PREDICTION, AND THEIR INTEGRATION

What is the time course of predictions, both with respect to one another and to the time course of word production? Firstly, predictions should be ready before the corresponding production representations are retrieved in the process of planning an utterance. Secondly, since we assumed that timing estimates are computed on the basis of content predictions, $\hat{p}$ (sem_A) should be ready before $\hat{t}$ (sem_A), $\hat{p}$ (syn_A) before $\hat{t}$ (syn_A), etc. Similarly for other-predictions, $\hat{p}$ (sem_B) should be ready before $\hat{t}$ (sem_B), $\hat{p}$ (syn_B) before $\hat{t}$ (syn_B), etc. (see horizontal arrows in **Figure 1**).

However, we intend not make any specific claim about the order in which predictions at the different levels (semantics, syntax, and phonology) are computed. It might be tempting to stipulate that the prediction system closely mimics the production system in this respect. In fact, however, the prediction system is a (forward) model of the production system and such a model need not implement all aspects of the internal dynamics of the modeled system. In particular, the prediction system for language could involve the same representational levels as the language production system, but the time course with which predictions are computed could differ from the time course of language production. Predictions at the levels of semantics, syntax, and phonology might even be computed separately and (roughly) simultaneously (Pickering and Garrod, 2007). In other words there could be separate mappings from the intention to communicate to semantics, syntax, and phonology. For this reason, in **Figure 1** we simply list the different predictions. Nevertheless, it is certainly the case that predictions at different levels are related to each other. For example, a prediction that the upcoming word refers to an object (a semantic prediction) and that it is a noun (a syntactic prediction) are related (because nouns tend to refer to objects). It is likely that the prediction system for language exploits such systematic relations between levels.

Once predictions are computed, how are they integrated in the process of planning an utterance (cf. vertical arrows in **Figure 1**)? To illustrate, take the following situation. The speaker needs to initiate articulation (art_A) rapidly, perhaps because of task instructions (in an experiment) or because of an impatient listener trying to get the floor. But she also knows that her chosen word is long (e.g., helicopter). The speaker computes $\hat{p}$ (phon$_A$), a prediction of the phonology of the word. On the basis of this, the speaker estimates, $\hat{t}$ (phon$_A$), that the complete phonological representation for that word will take a long time to construct, and that she will not be able to get it ready before the timeout. The predicted failure to meet the goal either (i) causes more resources to be invested in planning to speed things up, or, if processing speed is already at limit (ii) leads to early articulation of the first syllable of the word, even if the remaining syllables have not been prepared yet (Meyer et al., 2003). In other words, predicted outcomes (i.e., the output of the forward model) can trigger corrections to the ongoing planning process, in case such outcomes do not correspond to the intended goal.

METHODOLOGICAL RATIONALE: COMPARING SELF's AND OTHER's REPRESENTATIONS

How can we test whether the proposed account is correct? First, we should identify the conditions under which other-representations are formed. Second, we should investigate the nature of such representations. To do so, we need to compare individual production and joint production (in analogy with the joint action literature; e.g., Sebanz et al., 2003). In particular, we consider two instances of joint production: simultaneous productions (see Simultaneous Productions) and consecutive productions (see Consecutive Productions). In both sections, we first introduce the rationale behind joint production tasks and present the model's general predictions. Then, we describe a few specific methods in more detail. These make use of psycholinguistic tasks that (i) have been successfully employed in the study of isolated individual production, and (ii) can be distributed between two participants to study joint production. After a brief overview of the results typically found in individual production experiments, we list the specific predictions that our account makes with regard to the comparison of the individual and the joint task in each case.

SIMULTANEOUS PRODUCTIONS

Consider two speakers planning two different or similar utterances at the same time (see **Figure 1**). If *A* automatically represents *B*'s utterance as well as her own, then her act of production will be affected by the nature of his utterance, even if there is no need for coordination; the same holds for *B*'s representation of *A*'s utterance. We therefore expect joint simultaneous production to differ from individual production. By manipulating the relationship between the two speakers' utterances (e.g., whether they produce the same or different utterances), we can further investigate the nature of *A*'s representations of *B*'s utterances.

In particular, if predictions regarding other-generated utterances are computed *via* a model of one's production system, it should be possible to simulate another's utterances *without* the corresponding representations being activated in one's own production system. Additionally, it might be possible to maintain two models active in parallel (Wolpert et al., 2003; Wilson and Knoblich, 2005), for one's own and one's partner's utterances. However, using the same format simultaneously for simulating oneself and another may well lead to competition (Hamilton et al., 2004; Wilson and Knoblich, 2005). If so, we expect greater interference from *B*'s utterance on *A*'s production when *A* and *B* perform the same act of production than when they perform different acts.

Nevertheless, if (at least partial) activation of *A*'s own production system follows her simulation of *B via* the forward model, then we expect representations of *B*'s utterances to interact with representations of *A*'s own utterances in the way that representations for different self-generated utterances should interact. What would be the effect of such interaction *within A*'s production system? There might be facilitation or interference, depending on a variety of factors (e.g., whether *B* is producing the same word or a different word; in the latter case, whether the two words are related in form or meaning; cf. Schriefers et al., 1990).

Besides, since some representations are harder to process than others, variables that affect processing difficulty of self-generated utterances should also exert an effect in relation to other-generated utterances. Consider, for instance, the following situation. *A* and *B* name different pictures. The frequency of picture names is varied, so that on some trials *B* produces low-frequency words, whereas on others he produces high-frequency words. Given that it is harder to access the phonological representation of a low-frequency word than a high-frequency word (cf. Miozzo and Caramazza, 2003), we predict that representing *B*'s utterance will interfere more with *A*'s naming in the low-frequency condition than the high-frequency condition. In general, the difficulty of *B*'s task will affect the degree to which the representation of *B*'s utterances affects *A*'s production of her own utterances.

To sum up, paradigms that involve two speakers' simultaneously or near-simultaneously producing utterances serve two purposes: they test whether self- and other-generated utterances are represented in the same way, and they can elucidate the nature of other-representations, and in particular whether they involve the activation of one's own production system. Below we describe two such paradigms in more detail: joint picture–word interference and joint picture–picture naming.

Joint picture–word interference

In the classical picture–word interference paradigm (individual task), naming latencies are affected by the relationship between the pictures that the participant is required to name and words superimposed on those pictures. For example, semantically related distractor words lead to longer latencies than unrelated distractor words (Schriefers et al., 1990). The task-irrelevant stimulus (word) is thought to be automatically processed and interfere with the response to the task-relevant stimulus (picture).

In a joint version of this task, participants take turns to name the picture and to perform a secondary task, which is either congruent or incongruent with the primary task of picture naming. One possibility is for the participants to be in the same room, with the congruent task being tacit naming of the picture and the incongruent task being tacit naming of the word. Alternatively, the participants could be in separate and soundproofed rooms, in which case the secondary task could be overt picture or word naming. In any case, we would have a SAME condition (congruent secondary task), in which both participants produce the same utterance (i.e., the picture's name) and a DIFFERENT condition (incongruent secondary task), in which they produce different utterances (i.e., the picture's name and the distractor word). If speakers represent the processes underlying their partners' acts of speaking, we expect both the SAME and DIFFERENT conditions to differ from the individual task. If speakers represent the processes underlying their partners' response *via* a forward model, we expect longer latencies in the SAME than the DIFFERENT condition. If representing the other involves activation of one's own production system, on the contrary, we expect faster latencies in the SAME than in the DIFFERENT condition. In addition, we may find enhanced effects of distractor words on the processing of the pictures (e.g., greater semantic interference) in the DIFFERENT condition.

Joint picture–picture naming

In picture–picture naming tasks, participants name a target picture which is presented in the context of another (distractor) picture. The distractor picture is either related or unrelated to the target picture. Unlike picture–word interference experiments, picture–picture naming experiments typically show no clear effect of semantically related distractors on target naming latencies (e.g., Navarrete and Costa, 2005). In a joint version of the picture–picture naming task, participants either name one picture or remain silent. For trials on which the participant is naming a picture, we vary whether the partner remains silent (NO condition) or names the same (SAME condition) or a different picture (DIFFERENT condition). Assuming that the task-irrelevant picture's name is not automatically activated when performing the individual task, the NO condition should act as a control. If the participant represents the fact that her partner is naming a picture, then this may similarly affect both the SAME and the DIFFERENT condition; if she represents that her partner is naming a specific picture, we predict the SAME and the DIFFERENT condition will differ from each other. Again, the direction of these effects will depend on whether or not the production system is implicated in the representation of the other (see The Nature of the Representation of the Other).

CONSECUTIVE PRODUCTIONS

One concern with the study of simultaneous production is that it is comparatively rare in real conversations. Of course, speakers do occasionally contribute at the same time, for example when two listeners both claim the ground (e.g., in response to a question; Wilson and Wilson, 2005) or in intended choral co-production (e.g., mutual greetings; Schegloff, 2000). But it may be that speakers do not need a system that is specialized for representing their own utterance and a simultaneous utterance by their partner.

In contrast, consecutive production occurs all the time in conversation. First, the norm in dyadic conversations is the alternation of speaking turns. Second, conversational analysts have noted the occurrence of "collaborative turn completion" (Lerner, 1991). As illustrated in Example 1 below, *B*'s act of production completes *A*'s act appropriately and with minimum delay (0.1 means 100 ms). Instances of "collaborative turn completion" are striking, because two people effectively coordinate to jointly deliver one well-formed utterance.

1. *A*: so if one person said he could not invest (0.1)
   B: then I'd have to wait

(Lerner, 1991, p. 445)

Thus, speakers have much more need of representing their own utterance and their partner's upcoming utterance. Consecutive production paradigms should then somewhat mimic the naturalistic situation exemplified in 1. For example, *A* and *B* could be shown two pictures (e.g., of a wig and of a carrot), one on the right and one on the left of a computer screen. *A* first names the left picture (*wig*); then *B* names the right picture (*carrot*; see **Figure 2A**). They are told to minimize delay between the two names (cf. Griffin, 2003). We therefore create a joint goal for them. This situation certainly differs from naturally occurring instances of "collaborative turn completion", but it allows clear experimental control, and is arguably comparable to using tasks such as picture naming to understand natural monolog. (In an alternative version of the task, participants might simply start speaking in response to cues, which might occur at different times (i.e., SOAs) depending on condition.)

Figure 2A presents a schematic description. Given the complexity of the situation, in order to ensure that the figure is readable, we illustrate what happens from the perspective of *A*, the speaker that names the first picture. The timeline at the top shows the time course of word production for *A*'s utterance (and the onset of *B*'s utterance). Just as for the simultaneous production paradigm, we assume that *A* generates timing estimates for her own utterance and that these estimates are based on content predictions (left box). In addition, we hypothesize that *A* represents *B*'s upcoming utterance in a similar format and computes timing estimates and content predictions for that utterance, as well (right box).

To test these hypotheses, we again compare joint tasks with solo tasks. In the solo task (see **Figure 2B**), which was first used by Griffin (2003), *A* produces both pictures' names, with the same instruction of avoiding pausing between the two. Clearly, *A* goes through all the processing levels for both words and builds representations at each level. The timeline at the top of panel B differs from the one in **Figure 1**: most notably, t(art$_A$) corresponds to 1200 ms, instead of the 600 ms posited by Indefrey and Levelt (2004). This reflects the finding that participants tend to delay the onset of the first word, presumably because they perform advance planning. They start planning the second word before initiating the articulation of the first one (Griffin, 2003). We also assume that *A* computes timing estimates and content predictions for the second word, as well as for the first word.

If content and timing predictions computed for *B*'s utterance in the JOINT condition are similar to those computed for *A*'s own second utterance in the SOLO condition, we expect the JOINT and the SOLO condition to show similar patterns of results. Of course, we might also expect any effects to be weaker in the JOINT than in the SOLO condition, if other-representations are weighted less than self-representations (see The Nature of the Representation of the Other). We know that the amount of planning that speakers perform before articulation onset (and, consequently, speech onset latency) depends on various properties of the planned material, such as its length (Meyer et al., 2003) or syntactic complexity (e.g., Ferreira, 1991). Therefore, we expect

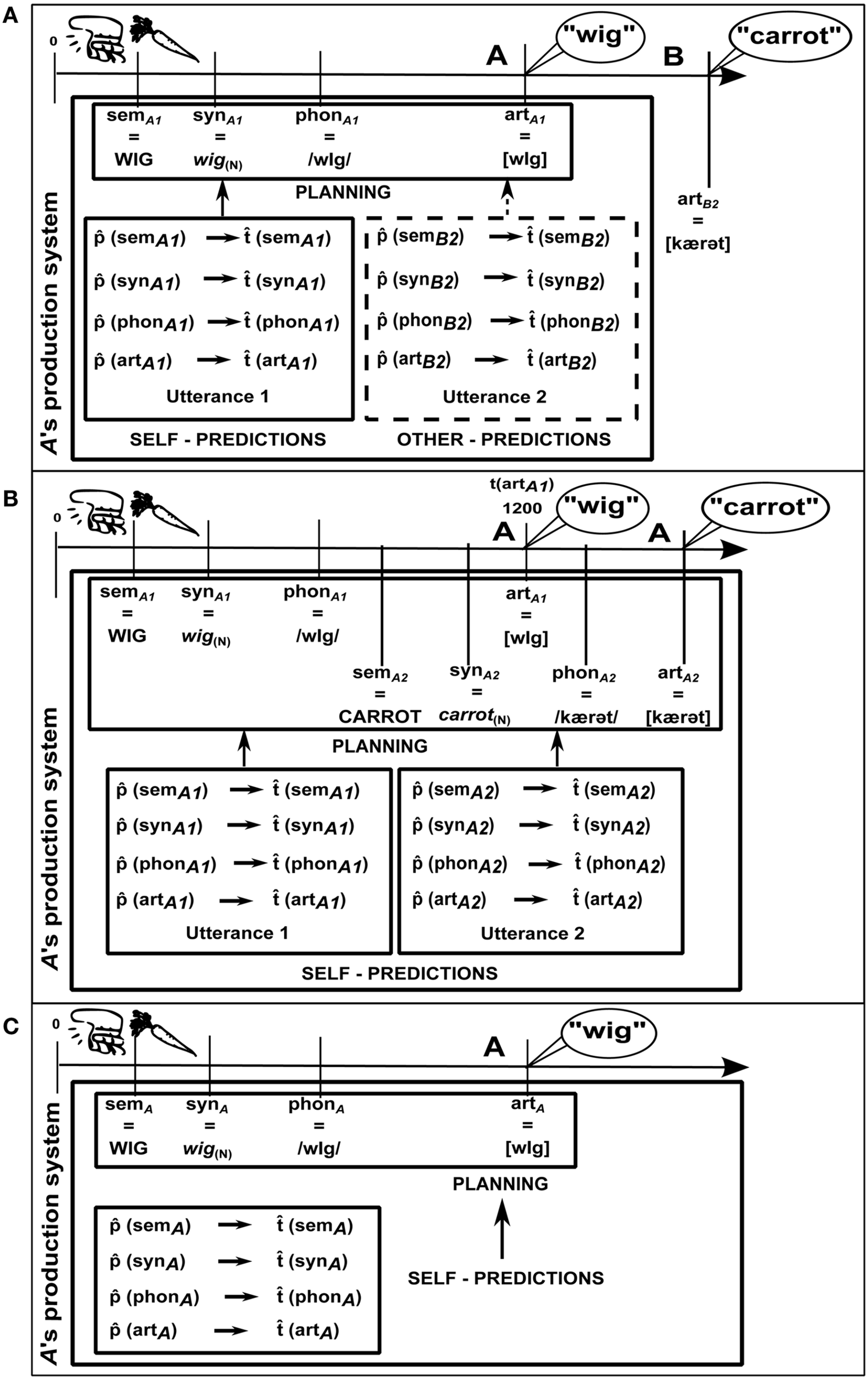

FIGURE 2 | Consecutive utterances: pictures of a wig and a carrot appear simultaneously. (A) JOINT: *A* names the left picture, then *B* names the right picture. **(B)** SOLO: *A* names the left picture, then *A* names the right picture. **(C)** NO: *A* names the left picture. Where two utterances are produced, we indicate the temporal relation between them by way of number subscripts (1 for the first utterance, 2 for the second utterance). In **(A)** art$_{B2}$ stands for the articulation stage of *B*'s utterance and p̂ (sem$_{B2}$) is the semantic content prediction that *A* generates in relation to *B*'s utterance. Time in ms. All other details as in **Figure 1**.

speech onset latencies for the first word to be affected by properties of the second word in the SOLO condition. This would reflect an influence of predictions of the second word's features on the planning of the first word. In the JOINT condition, we predict *A*'s speech onset will be similarly affected (though perhaps to a lesser degree), despite the fact that the second word is actually produced by *B*. This would show that predictions of the second word's features are computed and can affect planning of the first word also when the second word is generated by another speaker.

Additionally, the JOINT condition could be usefully contrasted to the NO condition, depicted in **Figure 2C**. The NO condition is equivalent to an instance of isolated production of a single word by *A*. Importantly, *A*'s task is the same in the NO and the JOINT conditions (i.e., producing Utterance 1), the only difference being that *B* does not produce Utterance 2 in the NO condition. The NO condition can therefore act as a control: no effect on onset latencies is expected.

Below we present various experiments that implement these ideas and discuss detailed predictions for each. Note that having the participants perform both roles is advisable, for two reasons. First, it allows data from both participants in a pair to be collected (therefore also comparisons between the behavior of the partners). Second, performing *B*'s task on half of the trials is likely to maximize the accuracy of *A*'s estimates of *B*'s timing.

Joint reversed length-effect

In Griffin's (2003) study, two pictures appeared simultaneously. The participant was told to name both pictures, avoiding pauses between the two names. She found a reversed length-effect: participants tended to initiate speech later when the first name was shorter than when it was longer; they also tended to look at the second picture more prior to speech onset and less after speech onset. Meyer et al. (2007) reported no effect on speech latencies, but they showed that the gaze–speech lag for the second picture was longer when the first name was shorter. Overall, these results seem to suggest that participants can estimate the amount of time that will be available for preparation of the second name during the articulation of the first name (Griffin, 2003).

We can therefore ask if they also estimate the time that their partner spends preparing the second name. In the SOLO condition, one participant names both pictures on a given trial; this condition is the same as Griffin (2003), except for the fact that two people are present and take turns in performing the task. In the NO condition, participants alternate in naming only the first picture, with both partners ignoring the second picture. In the critical JOINT condition, one participant names the first picture, then the other names the second picture; they alternate in performing either half of the task. We expect *B* (who has to name the second picture) to start looking at the second picture earlier (relative to when *A* starts speaking) when the first name is shorter. This would show that *B* is anticipating he will have less time to prepare his utterance when *A* is speaking. Besides, we expect *A* to initiate shorter words later than longer words. This would show that *A* is estimating *B*'s speech onset latencies and taking this estimate into account to successfully coordinate with *B* in producing a fluent utterance.

A related paradigm is based on Meyer (1996). She showed that when one participant is asked to name two pictures with a conjoined noun phrase, the auditory presentation of a distractor related in meaning to the second name delays onset latencies of the conjoined phrase. Again, if *A* contributes the first noun and *B* the second noun of the conjoined noun phrase and they have to coordinate to produce a fluent utterance (JOINT condition), we predict *A*'s speech will be affected by the relationship between the distractor and the second noun.

Joint syntactic encoding

The greater the syntactic complexity of the subject of a sentence, the longer it takes to start uttering the sentence. For example, a complex subject containing a prepositional phrase modifier or a relative clause slows down initiation times compared to a simple subject composed of two conjoined noun phrases, even when length is controlled for (Ferreira, 1991). The SOLO condition would be based on Ferreira's experiments (except for the presence of two participants): sentences could be first memorized and then produced upon presentation of a "go"-signal. In the JOINT condition, both participants would memorize the sentences. Then, depending on the cue presented at the beginning of the trial, either *A* or *B* would produce the subject (e.g., *The bike*), while their partner would contribute the rest of the sentence (e.g., *was damaged* vs. *that the cars ran over was damaged*). We expect a syntactic complexity effect on initiation times of the subject.

Active utterances are also initiated faster than the corresponding passives (Ferreira, 1994). Participants in the SOLO condition either produce sentences using a set of words provided by the experimenter or they describe pictures depicting a transitive event (e.g., of a girl hitting a boy). They are instructed to always start with the word or character presented in green (the so-called "traffic light" paradigm; Menenti et al., 2011). In this way, it is possible to control the voice of the sentence (e.g., if the boy is the first-named entity, a passive will be produced, otherwise an active). In the JOINT condition, participant *A* names only this first entity, while participant *B* produces the rest of the sentence. We expect *A*'s speech onset latencies to be slower when *B* produces a passive continuation than an active continuation; similar (or larger) results would occur in the SOLO condition, but not in the NO condition. A related paradigm could compare short vs. long continuations; it is known that more disfluencies are found at the start of longer constituents (Clark and Wasow, 1998) and it takes longer to start uttering a sentence when the subject is a conjoined noun phrase than when it is a simple noun phrase (Smith and Wheeldon, 1999).

Shared error-repair

In instances of spontaneous self-repair, people stop speaking because they detected an error in their speech and then resume with the intended output. In Hartsuiker et al. (2008), participants named pictures. On a small percentage of trials, an initial picture (the *error*) changed into a target picture (the *resumption*). Participants were told to stop speaking as fast as possible when they detected the change. In one experiment (Experiment 1), then, the same participant was asked to resume as fast as possible by naming the target picture, whereas in another experiment (Experiment 2) the task was simply to stop speaking (Hartsuiker et al., 2008).

Hartsuiker et al., 2008; see also Tydgat et al., 2011) showed that the process of stopping and the process of planning the resumption share resources: in Experiment 1, participants took longer to stop naming the error when the resumption was more difficult (through the target picture being degraded) than when it was less difficult (through the picture being intact). Moreover, there is evidence for strategic processing: when a resumption follows, people tend to withdraw resources from stopping, and instead invest them in planning the resumption while carrying on speaking. In other words, they prefer to complete the error rather than to interrupt it right away. A two-person version of Experiment 1 (stopping and resuming) would correspond to the SOLO condition, whereas a two-person version of Experiment 2 (stopping) would be our NO condition. In the critical JOINT condition, *A* stops, then *B* resumes. Therefore, *A* does not contribute the resumption. However, if she predicts that *B* will resume, we expect she will preferentially withdraw resources from stopping and complete the error, even if she does not need to invest these resources in planning the resumption.

CONCLUSION

After reviewing the literature on joint actions, we identified three mechanisms of action coordination: representational parity between self- and other-generated actions, prediction of observed actions, and integration of others' actions into the planning of one's own actions. We then claimed that similar mechanisms could underlie the coordination of utterances. We gave a comprehensive account of the type of information that could be represented about another's utterances. In considering the nature of these representations, we proposed that they are predictions generated by a forward model of one's own production system. Finally, we described two types of experimental paradigms (simultaneous productions and consecutive productions) that may prove informative as to the nature, extent, and accuracy of other-representations.

ACKNOWLEDGMENTS

Chiara Gambi is supported by a University of Edinburgh studentship. We thank Joris Van de Cavey and Uschi Cop for useful discussions.

REFERENCES

Adank, P., Hagoort, P., and Bekkering, H. (2010). Imitation improves language comprehension. *Psychol. Sci.* 21, 1903–1909.

Atmaca, S., Sebanz, N., Prinz, W., and Knoblich, G. (2008). Action corepresentation: the joint SNARC effect. *Soc. Neurosci.* 3, 410–420.

Blakemore, S.-J., and Decety, J. (2001). From the perception of action to the understanding of intention. *Nat. Rev. Neurosci.* 2, 561–567.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition* 75, B13–B25.

Brass, M., Bekkering, H., Wohlschläger, A., and Prinz, W. (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues. *Brain Cogn.* 44, 124–143.

Buccino, G., Binkofski, F., and Riggio, L. (2004). The mirron neuron system and action recognition. *Brain Lang.* 89, 370–376.

Caramazza, A., Costa, A., Miozzo, M., and Bi, Y. (2001). The specific-word frequency effect: implications for the representation of homophones in speech production. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1430–1450.

Clark, H. H. (1996). *Using Language.* Cambridge: Cambridge University Press.

Clark, H. H. (2002). Speaking in time. *Speech Commun.* 36, 5–13.

Clark, H. H., and Wasow, T. (1998). Repeating words in spontaneous speech. *Cogn. Psychol.* 37, 201–242.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *J. Phon.* 31, 139–148.

Cummins, F. (2009). Rhythm as entrainment: the case of synchronous speech. *J. Phon.* 37, 16–28.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385.

De Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language* 82, 515–535.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word preactivation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *J. Mem. Lang.* 30, 210–233.

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *J. Mem. Lang.* 33, 715–736.

Fowler, C. A., Brown, J. M., Sabadini, L., and Weihing, J. (2003). Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *J. Mem. Lang.* 49, 396–413.

Garrod, S., and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition* 27, 181–218.

Garrod, S., and Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Top. Cogn. Sci.* 1, 292–304.

Gravano, A., and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* 25, 601–634.

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychon. Bull. Rev.* 10, 603–609.

Grosjean, F., and Hirt, C. (1996). Using prosody to predict the end of sentences in English and French: normal and brain-damaged subjects. *Lang. Cogn. Process.* 11, 107–134.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–442.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301.

Hamilton, A., Wolpert, D. M., and Frith, U. (2004). Your own action influences how you perceive another person's action. *Curr. Biol.* 14, 493–498.

Hartsuiker, R. J., Catchpole, C. M., De Jong, N. H., and Pickering, M. J. (2008). Concurrent processing of words and their replacements during speech. *Cognition* 108, 601–607.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Commun.* 53, 23–35.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–937.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286, 2526–2528.

Indefrey, P., and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition* 92, 101–144.

Keller, P. E., Knoblich, G., and Repp, B. H. (2007). Pianists duet better when they play with themselves: on the possible role of action simulation in synchronization. *Conscious. Cogn.* 16, 102–111.

Kerzel, D., and Bekkering, H. (2000). Motor activaction from visible speech: evidence from stimulus response compatibility. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 634–647.

Knoblich, G., Butterfill, S., and Sebanz, N. (2011). "Psychological research on joint action: theory and data", in *The Psychology of Learning and Motivation*, ed. B. Ross (Burlington: Academic Press), 59–101.

Knoblich, G., and Flach, R. (2001). Predicting the effects of actions: interactions of perception and action. *Psychol. Sci.* 12, 467–472.

Knoblich, G., and Jordan, G. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 1006–1016.

Knoblich, G., Seigerschmidt, E., Flach, R., and Prinz, W. (2002). Authorship effects in the prediction of handwriting strokes: evidence for action simulation during action perception. *Q. J. Exp. Psychol.* 55A, 1027–1046.

Lerner, G. H. (1991). On the syntax of sentences-in-progress. *Lang. Soc.* 20, 441–458.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–75.

Magyari, L., and De Ruiter, J. P. (2008). "Timing in conversation: the anticipation of turn endings", in *12th Workshop on the Semantics and Pragmatics of Dialogue*, eds J. Ginzburg, P. Healey and Y. Sato (London: King's college), 139–146.

Marslen-Wilson, M. (1973). Linguistic structure ans speech shadowing at very short latencies. *Nature* 244, 522–523.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696.

Menenti, L., Gierhan, S., Segaert, K., and Hagoort, P. (2011). Shared language: overlap and segregation (of) the neuronal infrastructure for speaking and listening revealed by fMRI. *Psychol. Sci.* 22, 1173–1182.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: results from picture-word interference experiments. *J. Mem. Lang.* 35, 477–496.

Meyer, A. S., Belke, E., Häcker, C., and Mortensen, L. (2007). Use of word length information in utterance planning. *J. Mem. Lang.* 57, 210–231.

Meyer, A. S., Roelofs, A., and Levelt, W. J. M. (2003). Word length effects in object naming: the role of a response criterion. *J. Mem. Lang.* 48, 131–147.

Miozzo, M., and Caramazza, A. (2003). When more is less: a counterintuitive effect of distractor frequency in the picture-word interference paradigm. *J. Exp. Psychol. Gen.* 132, 228–252.

Navarrete, E., and Costa, A. (2005). Phonological activation of ignored pictures: further evidence for a cascade model of lexical access. *J. Mem. Lang.* 53, 359–377.

Noordzij, M. L., Newman-Norlund, S. E., De Ruiter, J. P., Hagoort, P., Levinson, S. C., and Toni, I. (2009). Brain mechanisms underlying human communication. *Front. Hum. Neurosci.* 3:14. doi:10.3389/neuro.3309.3014.2009

Pezzulo, G., and Dindo, H. (2011). What should I do next? Using shared representations to solve interaction problems. *Exp. Brain Res.* 211, 613–630.

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–226.

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci. (Regul. Ed.)* 11, 105–110.

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154.

Pulvermüller, F., and Fadiga, L. (2010). Action perception: sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* 11, 351–360.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso Del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870.

Richardson, D. C., and Dale, R. (2005). Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn. Sci.* 29, 1045–1060.

Richardson, M. J., Marsh, K. L., and Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 62–79.

Riley, M. A., Richardson, M. J., Shockley, K., and Ramenzoni, V. C. (2011). Interpersonal synergies. *Front. Psychol.* 2:38. doi:10.3389/fpsyg.2011.00038

Rizzolatti, G., and Craighero, L. (2004). The mirron-neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science* 326, 445–449.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Lang. Soc.* 29, 1–63.

Schriefers, H., Meyer, A. S., and Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: picture-word interference studies. *J. Mem. Lang.* 29, 86–102.

Scott, S. K., Mcgettigan, C., and Eisner, F. (2009). A little more conversation, a little less action: candidate roles for the motor cortex in speech perception. *Nat. Rev. Neurosci.* 10, 295–302.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006a). Joint action: bodies and minds moving together. *Trends Cogn. Sci. (Regul. Ed.)* 10, 70–76.

Sebanz, N., Knoblich, G., Prinz, W., and Wascher, E. (2006b). Twin peaks: an ERP study of action planning and control in coacting individuals. *J. Cogn. Neurosci.* 18, 859–870.

Sebanz, N., and Knoblich, G. (2009). Prediction in joint action: what, when, and where. *Top. Cogn. Sci.* 1, 353–367.

Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition* 88, B11–B21.

Sebanz, N., Knoblich, G., and Prinz, W. (2005). How two share a task: corepresenting stimulus-response mappings. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1234–1246.

Shockley, K., Richardson, D. C., and Dale, R. (2009). Conversation and coordinative structures. *Top. Cogn. Sci.* 1, 305–319.

Shockley, K., Santana, M.-V., and Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 326–332.

Smith, M., and Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition* 73, 205–246.

Staub, A., and Clifton, C. J. (2006). Syntactic prediction in language comprehension: evidence from either…or. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 425–436.

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14425–14430.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10587–10592.

Tian, X., and Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* 1:166. doi:10.3389/fpsyg.2010.00166

Tydgat, I., Stevens, M., Hartsuiker, R. J., and Pickering, M. J. (2011). Deciding where to stop speaking. *J. Mem. Lang.* 64, 359–380.

Valdesolo, P., Ouyang, J., and Desteno, D. (2010). The rythm of joint action: synchrony promotes cooperative ability. *J. Exp. Soc. Psychol.* 46, 693–695.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467.

Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010). A minimal architecture for joint action. *Neural Netw.* 23, 998–1003.

Vissers, C. T. W. M., Chwilla, D. J., and Kolk, H. H. J. (2006). Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. *Brain Res.* 1106, 150–163.

Vlainic, E., Liepelt, R., Colzato, L. S., Prinz, W., and Hommel, B. (2010). The virtual co-actor: the social Simon effect does not rely on online feedback from the other. *Front. Psychol.* 1:208. doi:10.3389/fpsyg.2010.00208

Wilkes-Gibbs, D., and Clark, H. H. (1992). Coordinating beliefs in conversation. *J. Mem. Lang.* 31, 183–194.

Wilson, M. (2001). The case for sensorimotor coding in working memory. *Psychon. Bull. Rev.* 8, 44–57.

Wilson, M., and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol. Bull.* 131, 460–473.

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* 12, 957–968.

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 593–602.

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Curr. Biol.* 11, R729–R732.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 July 2011; paper pending published: 26 August 2011; accepted: 03 October 2011; published online: 01 November 2011.

Citation: Gambi C and Pickering MJ (2011) A cognitive architecture for the coordination of utterances. Front. Psychology ***2****:275. doi: 10.3389/fpsyg.2011.00275*
This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.