To be published in *Behavioral and Brain Sciences* (in press)

An Integrated Theory of Language Production and Comprehension

Martin J. Pickering
University of Edinburgh
Department of Psychology
7 George Square
Edinburgh EH8 9JZ
United Kingdom
Email: martin.pickering@ed.ac.uk
http://www.psy.ed.ac.uk/Staff/academics.html#PickeringMartin

Simon Garrod
University of Glasgow
Institute of Neuroscience and Psychology
58 Hillhead Street
Glasgow G12 8QT
United Kingdom
Email: simon@psy.gla.ac.uk
http://staff.psy.gla.ac.uk/~simon/

Abstract: Current accounts of language processing treat production and comprehension as quite distinct. This paper rejects the dichotomy. In its place, we propose that producing and understanding are tightly interwoven, and this interweaving underlies people's ability to predict themselves and each other. We start by noting that production and comprehension are forms of action and action perception. We then consider the evidence for interweaving in action, action perception, and joint action, and explain such evidence in terms of prediction. Specifically, we assume that actors construct forward models of their actions before they execute those actions, and that perceivers of others' actions covertly imitate those actions and then construct forward models of those actions. We then use these accounts of action, action perception, and joint action to develop accounts

of production, comprehension, and interactive language. Importantly, they incorporate well-defined levels of linguistic representation (such as semantics, syntax, and phonology). We show how speakers and comprehenders use covert imitation and forward modeling to make predictions at these levels of representation, how they interweave production and comprehension processes, and how they use these predictions to monitor the upcoming utterances. We show how this account explains a range of behavioral and neuroscientific data on language processing, and discuss some of the implications of the account.

Keywords: comprehension, covert imitation, dialogue, forward model, language, prediction, production

1. INTRODUCTION

Current accounts of language processing treat production and comprehension as quite distinct from each other. The split is clearly reflected in the structure of recent handbooks and textbooks concerned with the psychology of language (e.g., Gaskell, 2007; Harley, 2007). This does not merely reflect organizational convenience, but instead, comprehension and production are treated as two different questions to investigate. For example, researchers assume that the processes involved in comprehending a spoken or written sentence, such as resolving ambiguity, may be quite distinct from the processes involved in producing a description of a scene. In neurolinguistics, the "classic" Lichtheim-Broca-Wernicke model assumes distinct anatomical pathways associated with production and comprehension, primarily on the basis of deficit-lesion correlations in aphasia (see Ben Shalom & Poeppel, 2008). This paper rejects this dichotomy. In its place, we propose that producing and understanding are tightly interwoven, and this interweaving underlies people's ability to predict themselves and each other.

1.1. The traditional independence of production and comprehension

To see the effects of the split, we need to think about language use both within and between individuals, in terms of a model of communication (Fig. 1).

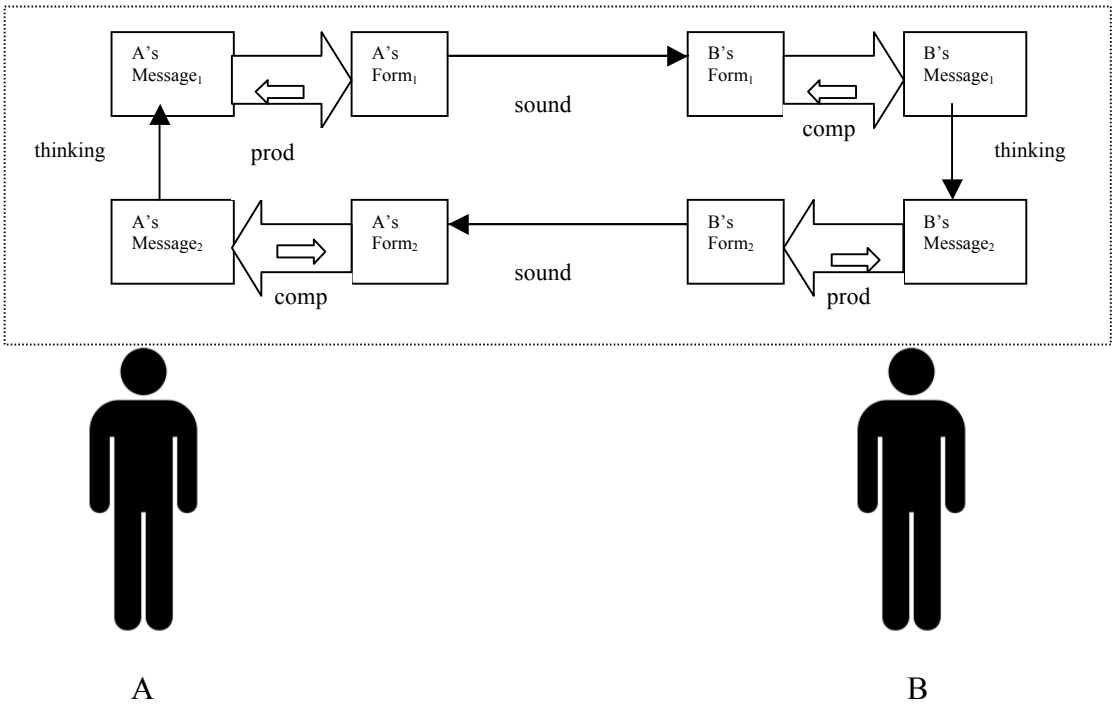

Figure 1. A traditional model of communication between *A* and *B*.

This account includes "thick" arrows between message and form, corresponding to production and comprehension. The production arrows represent the fact that production may involve converting one message into form (serial account) or the processor may convert multiple messages at once, then select one (parallel account). Within production, the "internal" arrows signify feedback (e.g., from phonology to syntax), which occurs in interactive accounts but not purely feedforward accounts. Note that these arrows are consistent with any type of information (linguistic or non-linguistic) being used during production. The arrows play an analogous role within comprehension (e.g., the internal arrows could signify feedback from semantics to syntax). In contrast, the arrows corresponding to sound are "thin" because a single sequence of sounds is sent forwards between the speakers. If communication is fully successful, then A's message$_1$ = B's message$_1$. Similarly, there is a "thin" arrow for

thinking because such models assume that each individual converts a single message (e.g., an understanding of a question, $message_1$) into another (e.g., an answer, $message_2$), and the answer does not affect the understanding of the question.

The model is split *vertically* between the processes in different individuals, who of course have independent minds. But it is also split *horizontally*, because the processes underlying production and comprehension within each individual are separated. The traditional model assumes discrete stages: one in which *A* is producing and *B* is comprehending an utterance, and one in which *B* is producing and *A* is comprehending an utterance. Each speaker constructs a message that is translated into sound before the addressee responds with a new message. Hence, dialogue is "serial monologue," in which interlocutors alternate between production and comprehension.

But in conversation, interlocutors' contributions often overlap, with the addressee providing verbal or non-verbal feedback to the speaker, and the speaker altering her contribution on the basis of this feedback. In fact, such feedback can dramatically affect both the quality of the speaker's contribution (e.g., Bavelas et al., 2000) and the addressee's understanding (Schober & Clark, 1989). This of course means that each interlocutor must simultaneously produce their own contribution and comprehend each other's contribution. Clearly an approach to language processing that assumes a temporal separation between production and comprehension cannot explain such behavior.

Interlocutors are not static, as the traditional model assumes, but are "moving targets" performing a joint activity (Garrod & Pickering, 2009). They do not simply transmit

messages to each other in turn, but rather negotiate the form and meaning of expressions that they use by interweaving their contributions (Clark, 1996), as illustrated in (1-3) below (from Gregoromichelaki et al., 2011). In (1b), *B* begins to ask a question, but *A*'s interruption (1c) completes the question and answers it. *B* therefore does not discretely encode a complete message into sound but rather *B* and *A* jointly encode the message across (1b-c).

1a-----A: I'm afraid I burnt the kitchen ceiling

1b-----B: But have you

1c-----A: burned myself? Fortunately not.

The horizontal split is also challenged by findings from isolated instances of comprehension or production. Take picture-word interference, in which participants are told to name a picture (e.g., of a dog) while ignoring a spoken or written distractor word (e.g., Schriefers et al., 1990). At certain timings, they are faster naming the picture if the word is phonologically related to it (*dot*) than if it is not. The effect cannot be caused by the speaker interpreting *dot* before producing *dog* – the meaning of *dot* is not the cause of the facilitation. Rather, the participant accesses phonology during the comprehension of *dot* and this affects the construction of phonology during the production of *dog*. So experiments such as these suggest that production and comprehension are tightly interwoven. Quite ironically, most psycholinguistic theories attempt to explain either production or comprehension, but a great many experiments appear to involve both. Single word naming is typically used to explain comprehension but involves production (see Bock, 1996). Sentence completion is often used to explain production but involves comprehension (e.g., Bock & Miller,

1991). Similarly, the finding that word identification can be affected by externally controlled cheek movement (Ito et al., 2009) suggests that production influences comprehension.

In addition, production and comprehension appear to recruit strongly overlapping neural circuits (Scott & Jonsrude, 2003; S.M. Wilson et al. 2004). For example, Paus et al. (1996) found activation (dependent on the rate of speech) of regions associated with speech perception when people whispered but could not hear their own speech. Listeners also activate appropriate muscles in the tongue and lips while listening to speech but not during non-speech (Fadiga et al., 2002; Watkins et al., 2003). Additionally, increased muscle activity in the lips is associated with increased activity (i.e., blood flow) in Broca's area, suggesting that this area mediates between the comprehension and production systems during speech perception (Watkins & Paus, 2004). There is also activation of brain areas associated with production during aspects of comprehension from phonology (Heim et al., 2003) to narrative structure (Mar, 2004); see Scott et al. (2009) and Pulvermüller and Fadiga (2010). Finally, Menenti et al. (2011) found massive overlap between speaking and listening for regions showing fMRI adaptation effects associated with repeating language at different linguistic levels (see also Segaert et al., in press). These results are inconsistent with separation of neural pathways for production and comprehension in the classical Lichtheim-Broca-Wernicke neurolinguistic model.

In conclusion, the evidence from dialogue, psycholinguistics, and cognitive neuroscience all casts doubt on the independence of production and comprehension, and therefore on the horizontal split assumed in Figure 1. Let us now address two

theoretical issues relating to the abandonment of this split, and then ask what kind of model is compatible with the interweaving of production and comprehension.

1.2. Modularity and the cognitive sandwich

Much of psycholinguistics has sought to test the claim that language processing is modular (Fodor, 1983). Such accounts investigate the way in which information travels between the boxes in a model such as in Figure 1. In particular, the arrows called *thinking* correspond to "central processes", and contain representations in some kind of language of thought. Researchers are particularly concerned with the extent to which *thinking* arrows are separated from the *production* and *comprehension* arrows. Modular theories assume that some aspects of production or comprehension do not make reference to "central processes" (e.g., Frazier, 1987; Levelt et al., 1999). In contrast, interactionist theories allow "central processes" to directly affect production or comprehension (e.g., Dell, 1986; MacDonald et al., 1994; Trueswell et al., 1994). But both types of theory maintain that production and comprehension are separated from each other. In this sense, both types of theory are modular, and are compatible with Figure 1.

In fact, Hurley (2008a) argued that traditional cognitive psychology assumes this type of modularity in order to keep action and perception separate. She referred to this assumption as the *cognitive sandwich*. Individuals perceive the world, reason about their perceptions using thinking (i.e., cognition), and act on the basis of those thoughts. Researchers assume that action and perception involve separate

representations and processes and study one or the other but not both (and they are kept separate in textbooks and the like). In Hurley's terms, the cognitive "meat" keeps the motor "bread" separate from the perceptual "bread".[1] She argues that perception and action are interwoven, and therefore rejects the cognitive sandwich.

Importantly, language production is a form of action and language comprehension is a form of perception. Therefore, traditional psycholinguistics also assumes the cognitive sandwich, with the thinking "meat" keeping apart the production and comprehension "bread." But if action and perception are interwoven, then production and comprehension are interwoven as well, and so accounts of language processing should also reject the cognitive sandwich.

1.3. Production and comprehension processes

How can production and comprehension both be involved in isolated speaking or listening? Within the individual, we mean that production and comprehension *processes* are interwoven. Production processes must of course be used when individuals produce language and comprehension processes must be used when they comprehend language. But production processes must also be used during, for example, silent naming, when no utterance is produced. Silent naming therefore involves some production processes (e.g., those associated with aspects of formulation such as name retrieval), but not others (e.g., those associated with articulation; see Levelt, 1989). Likewise, comprehension processes must occur when a participant retrieves the phonology of a masked prime word but not its semantics (e.g., Van den Bussche et al., 2009). And so it is also possible that production

processes are used during comprehension and comprehension processes used during production.

How can we distinguish production processes from comprehension processes? For this, we assume that (1) people represent linguistic information at different levels; (2) these levels are semantics, syntax, phonology[2]; (3) they are ordered "higher" to "lower", so that speaker's message is linked to semantics, semantics to syntax, syntax to phonology, and phonology to speech. We then assume that a producer goes from message to sound via each of these levels (message → semantics → syntax → phonology → sound), and a comprehender goes from sound to message in the opposite direction. Given this framework, we define a *production process* as a process that maps from a "higher" to a "lower" linguistic level (e.g., syntax to phonology) and a *comprehension process* as a process that maps from a "lower" to a "higher" level.[3] This means that producing utterances must involve production processes, but can also involve comprehension processes; and similarly, comprehending utterances must involve comprehension processes, but can also involve production processes.

One possibility is that people have separate production and comprehension systems. On this account, producing utterances may make use of feedback mechanisms that are similar in some respects to the mechanisms of comprehension; and comprehending utterances may make use of feedback mechanisms that are similar in some respects to the mechanisms of production. This is the position assumed by traditional interactive models of production (e.g., Dell, 1986) and comprehension (e.g., MacDonald et al., 1994). In such accounts, production and comprehension are internally non-modular,

but are modular with respect to each other. They do not take advantage of the comprehension system in production or the production system in comprehension (even though the other system is often lying dormant).

Very little work in comprehension makes reference to production processes, with classic theories of lexical processing (from e.g., Marslen-Wilson & Welsh, 1978 or Swinney, 1979 onwards) and sentence processing (e.g., Frazier, 1987; MacDonald et al., 1994) making no reference to production processes (see Bock, 1996 for discussion, and Federmeier, 2007 for an exception). In contrast, some theories of production do incorporate comprehension processes. Most notably, Levelt (1989) assumed that speakers monitor their own speech using comprehension processes. They can hear their own speech (external self-monitoring), in which case the speaker comprehends his own utterance just like another person's utterance; but they can also monitor a sound-based representation (internal self-monitoring), in which comprehension processes are used to convert sound to message (see Section 3.1).

In addition, some computationally sophisticated models can use production and comprehension processes together (e.g., Chang et al., 2006), use comprehension to assist in the process of learning to speak (Plaut & Kello, 1999), or assume that comprehension and production use the same network of nodes and connections so that feedback processes during production are the same as feedforward processes during comprehension (MacKay, 1982). In addition, Dell has proposed accounts in which feedback during production is a component of comprehension (e.g., Dell, 1988), though he has also queried this claim on the basis of neuropsychological evidence

(Dell, Schwartz, Martin, Saffran, & Gagnon, 1997, p. 830); see also the debate between Rapp and Goldrick (2000, 2004) and Roelofs (2004).

But none of these theories incorporate mechanisms of sentence comprehension (e.g., parsing or lexical ambiguity resolution) into theories of production. We believe that this is a consequence of the traditional separation of production and comprehension (as represented in Fig. 1). In contrast, we propose that comprehension processes are routinely accessed at different stages in production, and that production processes are routinely accessed at different stages in comprehension.

The rest of the paper develops an account of language processing in which processes of production and comprehension are integrated. We assume that instances of both production and comprehension involve extensive use of prediction – determining what you yourself or your interlocutor is likely to say next. Predicting your own utterance involves comprehension processes as well as production processes, and predicting another person's utterance involves production processes as well as comprehension processes.

As we have noted, production is a form of action and comprehension is a form of perception. More specifically, comprehension is a form of *action perception* – perception of other people performing actions. We first consider the evidence for interweaving in action and action perception, and explain such evidence in terms of prediction. We assume that actors construct forward models of their actions before they execute those actions, and that perceivers of others' actions construct forward

models of others' actions that are based on their own potential actions. Finally, we apply these accounts to joint action.

We then develop these accounts of action and action perception into accounts of production, comprehension, and dialogue. Unlike many other forms of action and perception, language has a clear structure, incorporating well-defined levels of linguistic representation such as semantics, syntax, and phonology. Thus our accounts also include such structure. We show how speakers and comprehenders predict the content of levels of representation by interweaving production and comprehension processes. We then explain a range of behavioral and neuroscientific data on language processing, and discuss some of the implications of the account.

2 INTERWEAVING IN ACTION AND ACTION PERCEPTION

For perception and action to be interwoven, there must be a direct link between them. If so, there should be much evidence for effects of perception on action, and there is. In one study, participants' arm movements showed more variance when they observed another person making a different versus the same arm movement (Kilner et al., 2003; see also Stanley et al., 2007). Conversely, there is good evidence for effects of action on perception. For example, producing hand movements can facilitate the concurrent visual discrimination of deviant hand postures (Miall et al., 2006), and turning a knob can affect the perceived motion of a perceptually bistable object (Wohlschläger, 2000). Such evidence immediately casts doubt on the "sandwich" architecture for perception and action.

What purpose might such a link serve? First, it could facilitate overt imitation, but overt imitation is not common in many species (see Prinz, 2006). Second, it could be used *postdictively*, with action representations helping perceivers develop a stable memory for a percept or a detailed understanding of it (e.g., via rehearsal), and perceptual representations doing the same for actors. But we propose a third alternative: people compute action representations during perception and perception representations during action to aid *prediction* of what they are about to perceive or to do, in a way that allows them to "get ahead of the game".[4] To explain this, we turn to the theory of forward modeling, which was first applied to action but has more recently been applied to perception. We interpret the theory in way that then allows us to extend it to account for language processing.

2.1. Forward modeling in action

To explain forward modeling, we draw on Wolpert's proposals from computational neuroscience (e.g., Davidson & Wolpert, 2005; Wolpert, 1997), but reframed using psychological terminology couched in the language of perception and action (see Fig. 2). We use the simple example of moving a hand to a target. The actor formulates the action (motor) command to move the hand. This command initiates two processes in parallel. First, it causes the action implementer to generate the act, which in turn leads the perceptual implementer to construct a percept of the experience of moving the hand. In Wolpert's terms, this percept is used as sensory feedback (*reafference*) and is partly proprioceptive, but may also be partly visual (if the agent watches her hand move).

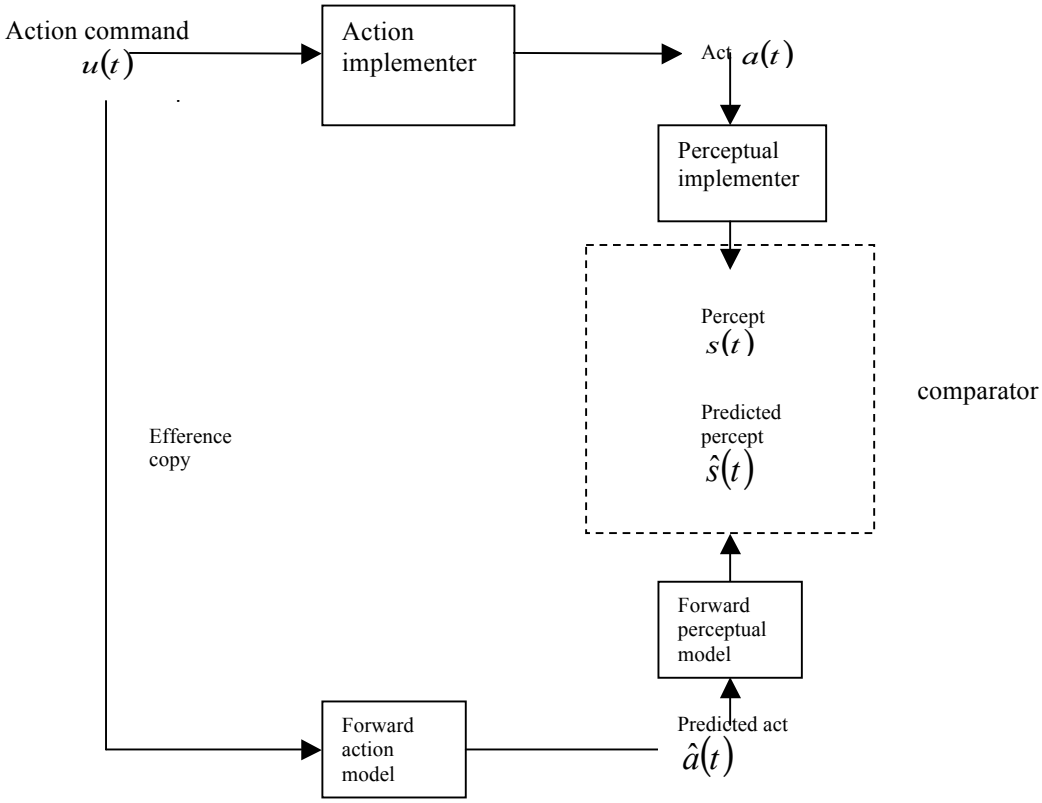

Figure 2. A model of the action system, using a snapshot of executing an act at time *t*. Boxes refer to processes, and terms not in boxes refer to representations. The action command $u(t)$ (e.g., to move the hand) initiates two processes. First, $u(t)$ inputs into the action (motor) implementer, which outputs an act $a(t)$ (the event of moving the hand). In turn this act inputs into the perceptual (sensory) implementer, which outputs a percept $s(t)$ (the perception of moving the hand). Second, an efference copy of $u(t)$ inputs into the forward action model, a computational device (distinct from the action implementer) which outputs a predicted act $\hat{a}(t)$ (the predicted event of moving the hand); the carat indicates an approximation. In turn $\hat{a}(t)$ inputs into the forward perceptual model, a computational device (distinct from the perceptual implementer) which outputs a predicted percept $\hat{s}(t)$ (the predicted perception of

moving the hand). The comparator can be used to compare the percept and the predicted percept.

Second, it sends an *efference copy* of the action command to cause the forward action model to generate the predicted act of moving the hand.[5] Just as the act depends on the application of the action command to the current state of the action implementer (e.g., where the hand is before the command), so the predicted act depends on the application of the efference copy of the action command to the current state of the forward action model (e.g., a model of where the hand is before the command). The predicted act then causes the forward perceptual model to construct a predicted percept of the experience of moving the hand. (This percept would not form part of a traditional action plan.) Note that this predicted percept is compatible with the Theory of Event Coding (Hommel et al., 2001), in which actions are represented in terms of their predicted perceptual consequences.

Importantly, the efference copy is (in general) processed more quickly than the action command itself (see Davidson & Wolpert, 2005). For example, the command to move the hand causes the action implementer to activate muscles, which is comparatively slow. In contrast, the forward action model and the forward perceptual model make use of representations of the position of the hand, state of the muscles, and so on (and may involve simplifications and approximations). These representations may be in terms of equations (e.g., hand coordinates), and such equations can (typically) be solved rapidly (e.g., using a network that represents relevant aspects of mathematics). So the predicted percept (the predicted sensations of the hand's movement and position) is usually "ready" before the actual percept.

The action then occurs and the predicted percept is compared to the actual percept (the sensations of the hand's actual movement and position).

Any discrepancy between these two sensations (as determined by the comparator) is fed back so that it can modify the next action command accordingly. If the hand is to the left of its predicted position, the next action command can move it more to the right. In this way, perceptual processes have an "on-line" effect on action, so that the act can be repeatedly affected by perceptual processes as well as action processes. (Alternatively, the actor can correct the forward model rather than the action command, depending on her confidence about the relative accuracy of the action command and the efference copy.) Such prediction is necessary because determining the discrepancy on the basis of reafferent feedback would be far too slow to allow corrective movements (see Grush, 2004, who refers to forward models as *emulators*).

We assume that the central role of forward modeling is perceptual prediction (i.e., predicting the perceptual outcomes of an action). But it has other functions. First, it can be used to help estimate the current state, given that perception is not entirely accurate. The best estimate of the current position of the hand combines the estimate that comes from the percept and the estimate that comes from the predicted percept. For example, a person can estimate the position of her hand in a dark room by remembering the action command that underlay her hand movement to its current location. Second, forward models can cancel the sensory effects of self-motion (*reafference cancellation*), when these sensory effects matched the predicted movement. This enables people to differentiate between perceptual effects of their

own actions and those reflecting changes in the world, for example explaining why self-applied tickling is not effective (Blakemore et al., 1999).

A helpful analogy is that of an old-fashioned sailor navigating across the ocean (cf. Grush, 2004). He starts at a known position which he marks on his chart (i.e., model of the ocean) and determines a compass course and speed. He lays out the corresponding course on the chart and traces out where he should be at noon (his predicted act, $\hat{a}(t)$), and determines what his sextant should read at this time and place (his predicted percept, $\hat{s}(t)$). He then sets off through the water until noon (his act, $a(t)$). At noon, he uses his sextant to estimate his position from the sun (his percept, $s(t)$), and can compare the predicted and observed sextant readings (using the comparator). He can then use this in various ways. If he is not confident of his course keeping, he pays more attention to the actual reading; if he is not confident of his sextant reading (e.g., it is misty), he pays more attention to the predicted reading. If the predicted and actual readings match, he assumes no other force (this is equivalent to reafference cancellation). But if they do not match and he is confident about both course keeping and sextant reading, he assumes the existence of another force, in this case the current.

Forward modeling also plays an important role in motor learning (Wolpert, 1997). To be able to pick up an object you need a model that maps the object's location onto a action (motor) command to move the hand to that location. This is called an inverse model because it represents the inverse of the forward model. Learning a motor skill requires learning both an appropriate forward model and an appropriate inverse model.

More sophisticated motor control theories use linked inverse-forward model pairs to explain how actors can adapt dynamically to changes in the context of an unfolding action. In their MOSAIC account, Haruno et al. (2001) proposed that actors run sets of model pairs in parallel, with each forward model making different predictions about how the action might unfold in different contexts. By matching actual movements against these different predictions, the system can shift responsibility for controlling the action toward the model pair whose forward model prediction best fits that movement. For example, you start to pick up a small (and apparently light) object using a weak grip but subsequently find the grip insufficient to lift the object. According to MOSAIC, you would then shift the responsibility for controlling the action to a new inverse-forward model pairing which produces a stronger grip.

The same principles apply to more complex structured activities such as the process of drinking a cup of tea. Here the forward model provides information ahead of time about the sequence and overlap between the different stages in the process (moving the hand to the cup, picking it up, moving it to the mouth, opening the mouth etc.) and represents the predicted sensory feedback at each stage (i.e., the predicted percept). Controlling such complex sequences of actions has been implemented by Haruno et al. (2003) in their Hierarchical MOSAIC (HMOSAIC) model. HMOSAIC extends MOSAIC by having hierarchically organized forward-inverse model pairings which link “high level” intentions to “low level” motor operations – in our terms, from high-level to low-level action commands.

In conclusion, forward modeling in action allows the actor to predict her upcoming action, in a way that allows her to modify the unfolding action if it fails to match the prediction. In addition, it can be used to facilitate estimation of the current state, to cancel reafference, and to support short- and long-term learning. In doing so, it closely interweaves representations associated with action and representations associated with perception, and can therefore explain effects of perception on action.

2.2. Covert imitation and forward modeling in action perception

When you perceive inanimate objects, you draw on your perceptual experience of objects. For example, if an object's movement is unclear, you can think about how similar objects have appeared to move in the past (e.g., obeying gravity). When you perceive other people (i.e., *action perception*), you can also draw on your perceptual experience of people. We refer to this as *the association route* in action perception. For example, you assume someone's ambiguous arm movement is compatible with your experience of perceiving other people's arm movements. People can clearly predict each other's actions using the association route, just as they can predict the movement of physical objects on the basis of past experience (e.g., Freyd & Finke, 1984).

But you can also draw on your experience of your own body – you assume that someone's arm movement is compatible with your experience of your own arm movements. We refer to this as *the simulation route* in action perception. The simplest possibility is that the perceiver determines "what they would do under the circumstances". In the case of hand movement, she would see the start of the actor's

hand movement and would then determine how she would move "if it were her hand", thereby determining the actor's intention. Informally, she would see the hand and the way it was moving, then think of it as her own hand, and use the mechanisms that she would use to move her own hand to predict her partner's hand movement. In other words, she would covertly imitate her partner's movements, treating his arm positions as though they were her own arm positions. However, the perceiver cannot simply use the same mechanisms the actor would use, but must "accommodate" to the differences in their bodies (the *context*, in motor control theory), for example applying a smaller force if her body is lighter than her partner's.[6] In any case, her reproduction is unlikely to be perfect – she is in the position of a character actor attempting to reproduce another person's mannerisms.

In theory, the perceiver could simulate by using her own action implementer (and inhibiting its output). However, this would be too slow – much of the time, she would determine her partner's action after he had performed that action. Instead, she can use her forward action model to derive a prediction of her partner's act (and the forward perceptual model to derive a prediction of her percept of that act). To do this, she would identify the actor's intention from her perception of the previous and current states of his arm (or from background information such as knowledge of his state of mind) and use this to generate an efference copy of the intended act. If she determined that the actor was about to punch her face, she would have time to move. She can also compare this predicted percept with her actual percept of his act when it happens. We illustrate this account in Figure 3.

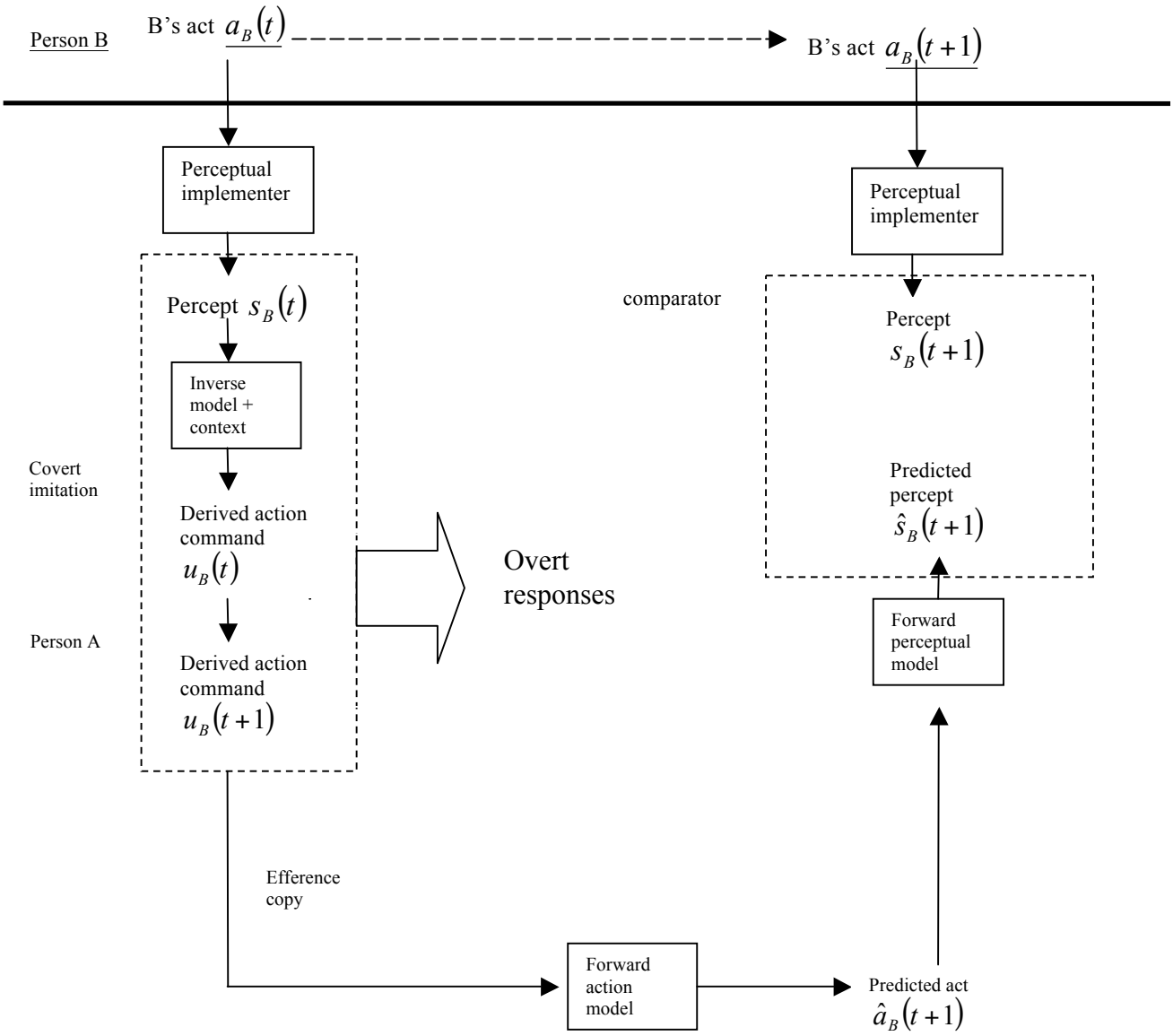

Figure 3. A model of the simulation route to prediction in action perception in Person *A*. Everything above the solid line refers to the unfolding action of Person *B* (who is being observed by *A*), and we underline *B*'s representations. For instance, $\underline{a_B(t)}$ can refer to B's initial hand movement (at time *t*) and $\underline{a_B(t+1)}$ to *B*'s final hand movement

(at time *t+1*). *A* predicts *B*'s act $\underline{a_B(t+1)}$ given *B*'s act $\underline{a_B(t)}$. To do this, *A* first covertly imitates *B*'s act. This involves perceiving B's act $\underline{a_B(t)}$ to derive the percept $s_B(t)$, and from this using the inverse model and context (e.g., information about differences between *A*'s body and *B*'s body) to derive the action command (i.e., the intention) $u_B(t)$ that *A* would use if *A* were to perform *B*'s act (without context, the inverse model would derive the command that *B* would use to perform *B*'s act – but this command is useless to *A*) and from this the action command that *A* would use if *A* were to perform the subsequent part of *B*'s act $u_B(t+1)$. *A* now uses the same forward modeling that she uses when producing an act (see Fig. 2) to produce her prediction of *B*'s act $\hat{a}_B(t+1)$, and her prediction of her perception of *B*'s act $\hat{s}_B(t+1)$. This prediction is generally ready before her perception of *B*'s act $s_B(t+1)$. She can then compare $\hat{s}_B(t+1)$ and $s_B(t+1)$ using the comparator. Notice that *A* can also use the derived action command $u_B(t)$ to overtly imitate *B*'s act and the derived action command $u_B(t+1)$ to overtly produce the subsequent part of *B*'s act (see Overt Responses).

This simulation account uses the mechanisms involved in the prediction of action (as illustrated in Figure 2), but adds a mechanism for covert imitation. This mechanism also allows for overt imitation of the action itself or a continuation of that action (overt responses). In fact, the strong link between actions and predictions of those actions means that perceivers tend to activate their action implementers as well as forward action models. Note that Figure 3 ignores the association route to action prediction, which uses the percept $s_B(t)$ and knowledge about percepts that tend to follow $s_B(t)$ to predict the percept of the act. (The perceiver may of course be able to

combine the action-based and perceptual predictions into a single prediction.) Additionally, we have glossed over the computationally complex part of this proposal – the mapping from the percept $s_B(t)$ to the action commands $u_B(t)$ and $u_B(t+1)$. How can the perceiver determine the actor's intention?

In fact, Wolpert et al. (2003) showed how to do this using HMOSAIC, which can make predictions about how different intentional acts unfold over time. In this account, the perceiver runs parallel linked forward-inverse model pairings at multiple levels from "low-level" movements to "high-level" intentions. By matching actual movements against these different predictions, HMOSAIC determines the likelihood of different possible intentions (and dynamically modifies the space of possible intentions). This in turn modifies the perceiver's predictions of the actor's likely behavior. For example, a first level might determine that a movement of the shoulder is likely to lead to a movement of the arm (and would draw on information about the actor's body shape); a second level might determine whether such an arm movement is the prelude to a proffered handshake or a punch (and would draw on information about the actor's state of mind). At the second level, the perceiver runs forward models based on those alternative intentions to determine what the actor's hand is likely to do next. If I predict you are more likely to initiate a handshake but then your fist starts clenching, I modify my interpretation of your intention and now predict that you will likely throw a punch. At this point, I have determined your intention and confidently predict the upcoming position of your hand, just as I would do if I were predicting my own hand movements.

Good evidence that covert imitation plays a role in prediction comes from studies showing that appropriate motor-related brain areas can be activated before a perceived event occurs (Haueisen & Knösche, 2001). Similarly, mirror neurons in monkeys can be activated by perceptual predictions as well as by perceived actions (Umiltá et al. 2001); note there is recent direct evidence for mirror neurons in people (Mukamel et al, 2010).[7] Additionally, people are better at predicting a movement trajectory (e.g., in dart-throwing or handwriting) when viewing a video of themselves versus others (Knoblich & Flach, 2001; Knoblich et al. 2002). Presumably, prediction-by-simulation is more accurate when the object of the prediction is one's own actions than when it is someone else's actions. This yoking of perceptual and action-based processes can therefore explain the experimental evidence for interweaving (e.g., Kilner et al., 2003).

Notice that such covert imitation can also drive overt imitation. However, the perceiver does not simply copy the movements of the actor, but rather bases her actions on her determination of the actor's intentions. This is apparent in infants' imitation of caregivers' actions (Gergely et al. 2002) and in the behavior of mirror neurons, which code for intentional actions (Umiltá et al. 2001). Importantly, mirror neurons do not exist merely to facilitate imitation (because imitation is largely or entirely absent in monkeys), and so one of their functions may be to drive action prediction via covert imitation (Prinz, 2006; Csibra & Gergly 2007). In conclusion, we propose that action perception interweaves action-based and perceptual processes in a way that supports prediction.

2.3. Joint action

People are highly adept at joint activities, such as ballroom dancing, playing a duet, or carrying a large object together (Sebanz et al., 2006a). Clearly, such activities require two (or more) agents to coordinate their actions, which in turn means that they are able to perceive each other's acts and perform their own acts together. In many of these activities, precise timing is crucial, with success occurring only if each partner applies the right force at the right time in relation to the other. Such success therefore requires tight interweaving of perception and action. Moreover, people must predict each other's actions, as responding after they perceive actions would simply be too slow. Clearly it may also be useful to predict one's own actions, and to integrate these predictions with predictions of others' actions.

We therefore propose that people perform joint actions by combining the models of prediction in action and action perception in Figures 2 and 3. Figure 4 shows how *A* and *B* can both predict *B*'s upcoming action (using prediction-by-simulation). *A* perceives *B*'s current act and then uses covert imitation and forward modeling; *B* formulates his forthcoming act and uses forward modeling based on that intention. If successful, they should make similar predictions about *B*'s upcoming act, and can use those predictions to coordinate. Note that they can both compare their predictions with *B*'s forthcoming act when it takes place.

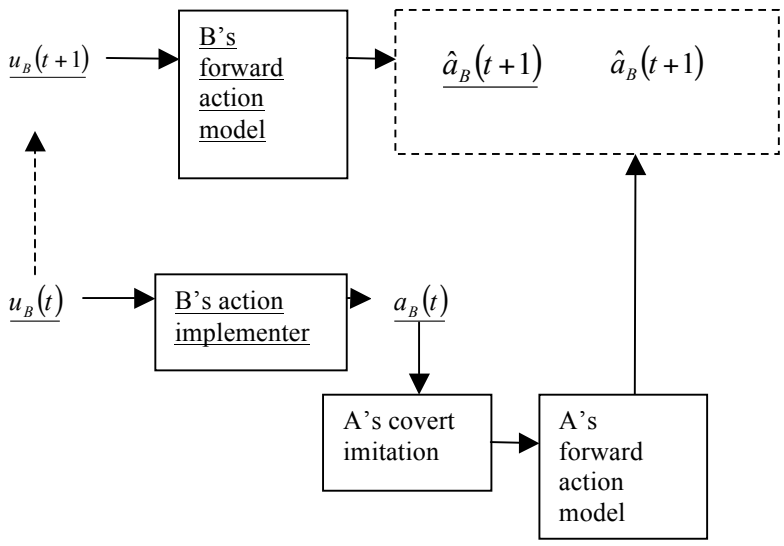

Figure 4. *A* and *B* predicting *B*'s forthcoming action (with *B*'s processes and representations underlined). *B*'s action command $\underline{u_B(t)}$ feeds into *B*'s action implementer and leads to B's act $\underline{a_B(t)}$. *A* covertly imitates *B*'s act and uses *A*'s forward action model to predict B's forthcoming act (at time *t*+1). *B* simultaneously generates the next action command (the dotted line indicates that this command is causally linked to the previous action command for *B* but not *A*) and uses *B*'s forward action model to predict *B*'s forthcoming act. If *A* and *B* are coordinated, then *A*'s prediction of *B*'s act and *B*'s prediction of *B*'s act (in the dotted box) should match. Moreover, they may both match *B*'s forthcoming act at time *t*+1 (not shown). *A* and *B* also predict *A*'s forthcoming action (see text).

Joint action can involve overt imitation, continuation of other's behavior, or complementary behavior. Overt imitation and continuation follow straightforwardly from Figure 3 (see the arrow from Covert imitation to Overt responses). There is

much evidence that people overtly imitate each other without intending to or being aware that they are doing so, from studies involving the imitation of specific movements (e.g., Chartrand & Bargh, 1999; Lakin & Chartrand, 2003) or the synchronization of body posture (e.g., Shockley et al., 2003). For example, pairs of participants tend to start rocking chairs at the same frequency, even though the chairs have different natural frequencies (Richardson et al., 2007), and crowds come to clap in unison (Neda et al., 2000). Such imitation appears to be on a perception-behavior expressway (Dijksterhuis & Bargh, 2001), not mediated by inference or intention. Many of these findings demonstrate close temporal coordination and appear to require prediction (see Sebanz & Knoblich, 2009). For instance, in a joint go-nogo task Sebanz et al. (2006b) found enhanced N170 ERPs (reflecting response inhibition) for the non-responding player when it was the partner's turn to respond. They interpreted this as suggesting that you suppress your own actions at the point when your partner is about to act. In addition, people continue each other's behavior by overtly imitating their predicted behavior (in contrast to overt imitation of actual behavior). For example, early studies showed that some mirror neurons fired both when the monkey observed an action and a different action that could follow the observed action (di Pelligrino et al., 1992).

Complementary behavior occurs when the co-actors use their same predictions to derive different (but coordinated) behaviors. For example, in ballroom dancing, both *A* and *B* predict that *B* will move his foot forwards; *B* will then move his foot, and *A* will plan her complementary action of moving her foot backwards. Graf et al. (2010) reviewed much evidence for complementary motor involvement in action perception (see Haberle et al., 2008; Newman-Norland et al., 2007; van Schie et al., 2008).

So far we have described how *A* and *B* predict *B*'s action. To explain joint activity, we first note that *A* and *B* predict *A*'s action as well (in the same way). They then integrate these predictions with their predictions of *B*'s action. To do this, they must simultaneously predict their own action and their partner's action. They can determine whether these acts are compatible (does my upcoming act fit with your upcoming act?). If not, they can modify their own upcoming actions accordingly (so that such modifications can occur on the basis of comparing predictions alone, without having to wait for the action). (If I find out that I am likely to collide with you, I can move out the way.) This account can therefore explain tight coupling of joint activity, as well as the experience of "shared reality" that occurs when *A* and *B* realize that they are experiencing the world in similar ways (Echterhoff et al. 2009).

Importantly, the participants in a joint action perform actions that are related to each other. It is of course easier for *A* to predict both *A* and *B* if *A* and *B*'s actions are closely related (as is the case in tightly coupled activities such as ballroom dancing). If *A*'s predictions of her own action ($\hat{a}_A(t+1)$) and her prediction of *B*'s action ($\hat{a}_B(t+1)$) were unrelated, she would find both predictions hard; but if the predictions are closely related, *A* is able to use many of the computations involved in one prediction to support the other prediction. In other words, it is easier to predict another person's actions when you are performing a related action than when you are performing an unrelated action. (Notice also that *A* and *B* are likely to overtly imitate each other and that such overt imitation will make their actions more similar, hence the predictions easier to integrate.) In conclusion, joint action can be successful

because the participants are able to integrate their own action with their perception of their partner's action.

3. A UNIFIED FRAMEWORK FOR LANGUAGE PRODUCTION AND COMPREHENSION

We noted that language production is a form of action and comprehension is a form of action perception, and therefore now apply the above framework to language. This is of course consistent with the evidence for interweaving that we briefly considered in Section 1: the tight coupling between interlocutors in dialogue, the evidence for effects of comprehension processes on acts of production and *vice versa* in behavioral experiments, and the overlap of brain circuits involved in acts of production and comprehension. We now argue that such interweaving occurs primarily to facilitate prediction, which in turn facilitates production and comprehension.

We first propose that speakers use forward production models of their utterances in the same way that actors use forward action models, by constructing efference copies of their predicted utterance and comparing those copies with the output of the actual production implementer. We then propose that listeners predict speakers' upcoming utterances by covertly imitating what they have uttered so far, deriving their underlying message, generating efference copies, and comparing those copies with the actual utterances when they occur, just as in our account of action perception. Dialogue involves the integration of the models of the speaker and the listener. These proposals are directly analogous to our proposals for action, action perception, and

joint action, except that we assume structured representations of language involving (at least) semantics, syntax, and phonology.

3.1. Forward modeling in language production

In acting, the action command drives the action implementer to produce an act, which the perceptual implementer uses to produce a percept of that act (see Fig. 2). But typically before this process is complete, the efference copy of the action command drives the forward action model to produce a predicted act, which the forward perceptual model uses to produce a predicted percept. The actor can then compare these outputs and adjust the action command (or the forward model) if they do not match.

In language production (see Fig. 5), the action command is specified as a production command. The action implementer is specified as the production implementer and the perceptual implementer is specified as the comprehension implementer. Similarly, the forward action model is specified as the forward production model and the forward perceptual model is specified as the forward comprehension model. The comparison of the utterance percept and the predicted utterance percept constitutes self-monitoring.

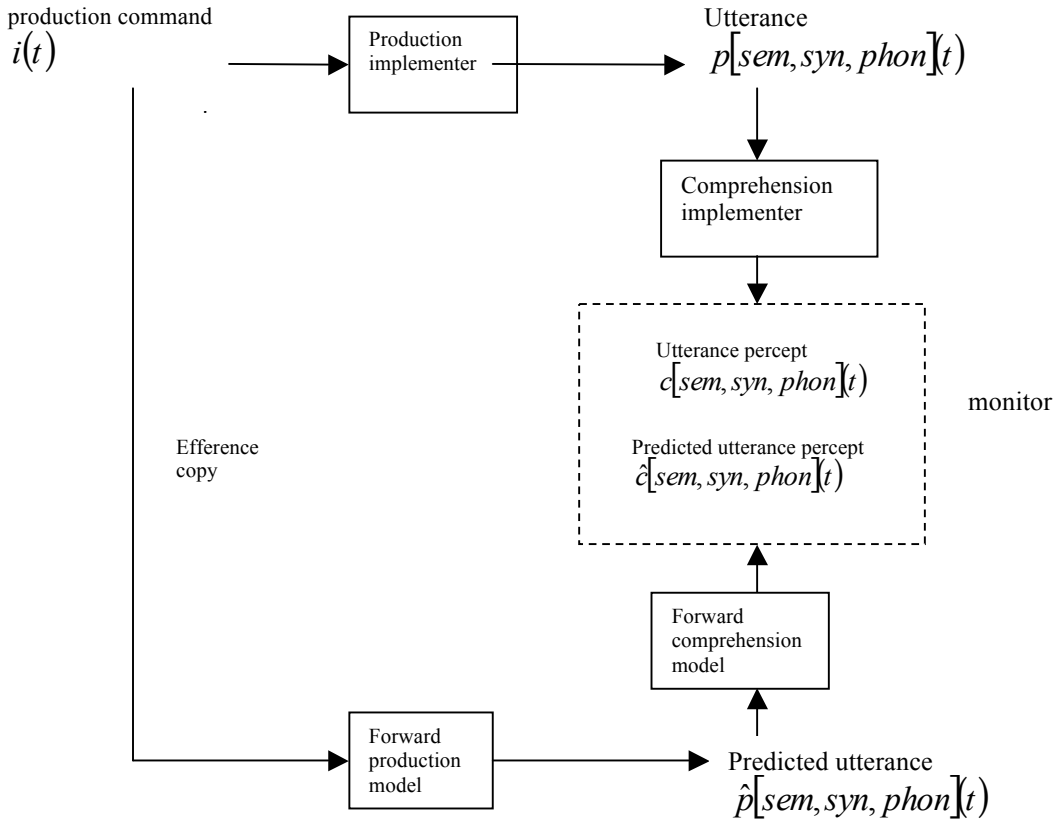

Figure 5. A model of production, using a snapshot of speaking at time *t*. The production command *i(t)* is used to initiate two processes. First, $i(t)$ inputs into the production implementer, which outputs an utterance p$[sem, syn, phon](t)$, a sequence of sounds that encodes semantics, syntax, and phonology. Notice that *t* refers to the time of the production command, not the time at which the representations are computed. In turn the speaker processes this utterance to create an utterance percept, the perception of a sequence of sounds that encodes semantics, syntax, and phonology. Second, an efference copy of $i(t)$ inputs into the forward production model, a computational device which outputs a predicted utterance. This inputs into the forward comprehension model, which outputs a predicted utterance percept (i.e., of the predicted semantics, syntax, and phonology). The monitor can then compare

the utterance percept and the predicted utterance percept at one or more linguistic levels (and therefore performs self-monitoring).

In Figure 5, the production command constitutes the message that the speaker wishes to convey (see Levelt, 1989) and includes information about communicative force (e.g., interrogative), pragmatic context, and a non-linguistic situation model (e.g., Sanford & Garrod, 1981). In addition, Figure 5 does not merely differ from Figure 2 in terminology, but also assumes structured linguistic representations, such as $p[sem, syn, phon](t)$ rather than $a(t)$. As we have noted, language processing appears to involve a series of intermediate representations between message and articulation. So Figure 5 is a simplification: We assume that speakers construct representations associated with the semantics, syntax, and phonology of the actual utterance, with the semantics being constructed before the syntax, and the syntax before the phonology (in accord with all theories of language production, even if they assume some feedback between representations).[8] We can therefore refer to these individual representations as $p[sem](t)$, $p[syn](t)$, and $p[phon](t)$. Note that the mappings from $p[sem](t)$ to $p[syn](t)$ and $p[syn](t)$ to $p[phon](t)$ involve aspects of the production implementer, but Figure 5 places the production implementer before a single representation $p[sem, syn, phon](t)$ for ease of presentation. Assuming Indefrey and Levelt's (2004) estimates (based on single-word production), semantics (including message preparation) takes about 175ms, syntax (lemma access) takes about 75ms, and phonology (including syllabification) takes around 205ms. Phonetic encoding and articulation takes an additional 145ms (see Sahin et al., 2009, for slightly longer estimates of syntactic and phonological processing).

Finally, speakers use the comprehension implementer to construct the utterance percept. Again, we assume that this system acts on each production representation individually, so that $p[sem](t)$ is mapped to $c[sem](t)$, $p[syn](t)$ to $c[syn](t)$, and $p[phon](t)$ to $c[phon](t)$; therefore Fig. 5 is a simplification in this respect as well. Importantly, the speaker constructs her utterance percept for semantics before syntax before phonology. Unlike Levelt (1989), we therefore assume that the speaker maps between representations associated with production and comprehension at all linguistic levels (see below).

The forward production model constructs $\hat{p}[sem](t)$, $\hat{p}[syn](t)$, and $\hat{p}[phon](t)$, and the forward comprehension model constructs $\hat{c}[sem](t)$, $\hat{c}[syn](t)$, and $\hat{c}[phon](t)$. Most importantly, these representations are typically ready before the representations constructed by the production implementer and the comprehension implementer. The speaker can then use the monitor to compare the predicted utterance percept with the (actual) utterance percept at each level (see Fig. 5) when those actual percepts are ready. Thus the monitor can compare predicted with actual semantics first, then predicted with actual syntax, then predicted with actual phonology. The production implementer makes occasional errors, and the monitor detects such errors by noting mismatches between outputs of the production implementer and outputs of the forward model. It may then trigger a correction (but does not need to do so). To do this, the monitor must of course be fairly accurate and use predictions made independently of the production implementer itself.

Let us now consider the content of these predictions and the organization of the forward models in more detail using examples. In doing so, we address the obvious

criticism that if the speaker is computing a forward model, why not just use that model in production itself? The answer is that the predictions are not the same as the implemented production representations, but are easier-to-compute "impoverished" representations. They leave out (or simplify) many components of the implemented representations, just as a forward model of predicted hand movements might encode coordinates but not distance between index finger and thumb, or a forward model for navigation might include information about the ship's position and perhaps fuel level but not its response to the heavy swell.

Similarly, the forward model does not form part of the production command. The production command incorporates a conceptual representation that describes a situation model and communicative force. It cannot represent information such as the first phoneme of the word the speaker is to use, because such information is phonological, not conceptual. In addition, the production command does not involve perceptual representations (what it "feels like" to perform an act), unlike the forward comprehension model.

Additionally, the forward model represents rather than instantiates time. For example, a speaker utters *The boy went outside to fly* …, and has decided to produce a word corresponding to a conceptual representation of a kite. At this point, she has predicted that the next word will be a definite determiner with phonology /*ðe/*, and that this should start in 100ms. (She does not wait 100ms to make this prediction.) She may also have predicted some aspects of the following word (*kite*) and that it should start in 300ms.

But apart from the timing, in what sense is this forward model impoverished? The phonological prediction ($\hat{p}[phon](t)$) might indicate (for example) the identities of the phonemes (/k/, /a/, /I/, /t/) and their order, but not how they are produced. So when the speaker decides to utter *kite*, she might simply look up the phonemes in a table and associate them with the numbers 1, 2, 3, and 4. Importantly, she does not necessarily have the prediction of /k/ ready before the prediction of /t/. Alternatively, she might look up the first phoneme, in which case the forward model would include information about /k/ only.

Similarly, the syntactic prediction ($\hat{p}[syn](t)$) might include the grammatical category of noun, but not whether the noun is singular or plural (or its gender, in a gender-marking language). The speaker might simply look up that a flyable object is likely to be a noun. This information then suggests that the word should occur at particular positions, for instance following a determiner. In addition, it is not necessary that the predicted representations are computed sequentially. Although the implemented syntax ($p[syn](t)$) must be ready before the implemented phonology ($p[phon](t)$), the syntactic prediction need not be ready before the phonological prediction. For example, the speaker might predict that the kite concept should have the first phoneme /k/ and predict that it should be a noun at the same time, or indeed predict the first phoneme without making any syntactic prediction at all. In summary, we assume that the production system "intervenes" between the implemented semantics and the implemented syntax, and between the implemented syntax and the implemented phonology, but do *not* assume intervention in the forward production model.

For example, a speaker might decide to describe a transitive event. At this point, she constructs a forward model of syntax, say $[NP\ [V\ NP]_{VP}]_S$. This appears appropriate if the speaker knows that transitive events are usually described by transitive constructions, a piece of information assumed in construction grammar (Goldberg, 1995), which associates constructions with "general" meanings. The speaker can therefore make this prediction before having decided on other aspects of the semantics of the utterance, thus allowing the syntactic prediction to be ready before the implemented semantics.

At a more abstract level, consider when the speaker wishes to refer to something in common ground (but not highly focused). On the basis of extensive experience, she can predict that the utterance will have the semantics definite nominal, the syntax $[Det\ N]_{NP}$, and the phonology starting with */ðe/*; she may also predict that she will start uttering the noun in 200 ms.

This approach might underlie choice of syntactic structure during production. For example, speakers of English favor producing short constituents before long ones (e.g., Hawkins, 1994). To do this, they might start constructing short and long constituents at the same time but tend to produce short ones first because they are ready first (see V.S. Ferreira, 1996). However, this appears inefficient because it would lead to sharp increases in processing difficulty at specific points (here, when producing the short phrase), and would therefore work against a preference for uniform information density during production (Jaeger, 2010, p. 25). It would mean that the long phrase would often be ready much too early, and would incorrectly predict that blend errors should be very common.

But alternatively, the speaker decides to describe a complex event and a simple event. She uses forward modeling to predict that the complex event will require a heavy phrase and the simple event a light phrase. She then evokes the "short before long" principle, and uses it to convert the simple event into a light phrase (using the production implementer). She can then wait till quite near the end of the phrase before beginning to produce the heavy phrase (again, using the implementer). In this way, she keeps information density fairly constant, prevents blending errors, and reduces memory load.

Just as in action, the speaker "tunes" the forward model based on experience speaking. If she has repeatedly formulated the intention to refer to a kite concept and then uttered the phoneme /k/, she will start to construct an accurate forward model ($\hat{p}[phon](t)$ = /k/) when she next decides to refer to such a concept. If she then constructs an incorrect phonological representation (e.g., $p[phon](t)$ = /g/), the monitor will likely immediately notice the mismatch between these two representations. If she believes the forward model is accurate, she will detect a speech error, perhaps reformulate, and modify her production implementer for subsequent utterances; if she believes that it may not be accurate, she will not reformulate but will alter her forward model accordingly (cf. Wolpert et al., 2001).

Evidence from speech production. There is good evidence for use of forward perceptual models during speech production. In an MEG study, Heinks-Maldonado et al. (2006) found that the M100 was reduced when people spoke and concurrently listened to their own unaltered speech versus a pitch-shifted distortion of the speech.

We assume that they construct a predicted phonological percept, $\hat{c}[phon](t)$. This typically matches their phonological percept ($c[phon](t)$) and thus suppresses the M100 (i.e., via reafference cancellation). But when the actual speech is distorted, the percept and the predicted percept do not match, and thus the M100 is enhanced. (The M100 could not reflect distorted speech itself as it was not enhanced when distorted speech was replayed to the speakers.) The rapidity of the effect suggests that speakers could not be comprehending what they heard and comparing this to their memory of their planned utterance. Additionally, Tian and Poeppel (2011) had participants produce or imagine producing a syllable, and found the same rapid MEG response in auditory cortex. This suggests that speakers construct a forward model incorporating phonological information under conditions when they do not speak (i.e., do not use the production implementer).

Tourville et al. (2008) had participants read aloud monosyllabic words while recording fMRI. On a small proportion of trials, participants' auditory feedback was distorted by shifting the first formant either up or down. Participants compensated by shifting their speech in the opposite direction within 100 ms. Such rapid compensation is a hallmark of feed-forward (predictive) monitoring (as correction following feedback would be too slow). Moreover, the fMRI results identified a network of neurons coding mismatches between expected and actual auditory signals. These three studies therefore provide clear evidence for forward models in speech production. In fact, Tourville and Guenther (2011) describe a specific implementation of such forward-model-based monitoring in the context of their DIVA and GODIVA models of speech production. However, these data and implementations do not relate to the full set of stages involved in language production.

Language production and self-monitoring. In psycholinguistics, well-established accounts of language production (e.g., Bock & Levelt, 1994; Dell, 1986; Garrett, 1980; Levelt, 1989; Levelt et al., 1999; Hartsuiker & Kolk, 2001) make no reference to forward modeling, and instead debate the operations of the production implementer (top line in Fig. 4). They tend to assume that self-monitoring uses the comprehension system. Levelt (1989) proposed that people can monitor what they utter (using an external loop) and thus repair errors. But he noted that they also make many repairs before completing the word, as in *to the ye- to the orange node*, where it is clear that they were going to utter *yellow* (Levelt, 1983), and show arousal when they are about to utter a taboo word but do not do so (Motley et al., 1975). He therefore proposed that they construct a sound-based representation (originally phonetic, but phonological in Wheeldon & Levelt, 1995) and input that representation directly into the comprehension system (using an internal loop). Note that other accounts assume more limited monitoring (e.g., suggesting that some evidence for monitoring is in fact due to feedback in the production system; Dell, 1986). But they do not deny the existence of a comprehension-based monitor.

However, alternative accounts assume that at least some monitoring can be "internal" to language production (e.g., Laver, 1980; Schlenck et al., 1987; Van Wijk & Kempen, 1987; see Postma, 2000). Such monitoring could involve the comparison of different aspects of implemented production, for example if the process is redundantly organized and a problem is noted if the outputs do not match (see Schlenck et al., 1987). Alternatively, it could register a problem if there is high conflict between potential words or phonemes (Nozari et al., 2011). Our account makes the rather

different claim that the monitor compares the output of implemented production (the utterance percept) with the output of the forward model (the predicted utterance percept).[9]

Of course, speakers clearly can perform comprehension-based monitoring using the external loop and indeed may be able to perform it using the internal loop as well. But a purely comprehension-based account cannot explain the data from Heinks-Maldonado et al. (2006) and Tourville et al. (2008). In addition, it has difficulty explaining the timing of error detection. To correct *to the ye- to the orange node*, the speaker prepares $p_{phon}(t)$ for *yellow*, converts it into $c_{phon}(t)$, uses comprehension to construct $c_{sem}(t)$, judges that $c_{sem}(t)$ is not appropriate (i.e., it is incompatible with $p_{sem}(t)$ or it does not make sense in the context), and manages to stop speaking, *before* she articulates more than *ye*. Given Indefrey and Levelt's (2004) estimates, the speaker has about 145ms plus the time to utter *ye-*, which is arguably less than the time it takes to comprehend a word (e.g., Levelt, 1989). Speakers might therefore "buffer" by delaying phonetic encoding and articulation (e.g., Blackmer & Mitton, 1991), but this is unlikely given that they speed up the process of monitoring and repair when speaking faster (see Postma, 2000). Such findings appear incompatible with a purely comprehension-based approach to monitoring.[10] In addition, Nozari et al. (2011) argue that non-speakers may be able to use the internal loop (as in Wheeldon & Levelt, 1995), but that speakers would face the extreme complexity of simultaneously comprehending different parts of an utterance with the internal and the external loops (see also Vigliocco & Hartsuiker, 2002). They also note that there is

much evidence for a dissociation between comprehension and self-monitoring in patients.

Huettig and Hartsuiker (2010) monitored speakers' eye movements as they referred to one of four objects in an array. The array contained an object whose name was phonologically related to the name of the target object. In comprehension experiments, people tend to look at such phonological competitors more than unrelated objects (Allopena et al., 1998). Huettig and Hartsuiker found that their speakers also tended to look at competitors after they had produced the target word. This suggests that they monitored their speech using the comprehension system. But they did not look at competitors while producing the target word. This suggests that they did not use a comprehension-based monitor of a phonological representation. Their findings therefore imply that speakers first monitor using a forward model (as we propose) but can later perform comprehension-based monitoring.

Accounts using an internal loop imply that phonological errors should be detected before semantic errors (assuming that both forms of detection are equally difficult). In contrast, our account claims that speakers construct the predicted semantic, syntactic, and phonological percepts early. They then construct the semantic percept, and compare it with the predicted semantic percept; then construct the syntactic percept, and compare it with the predicted syntactic percept; and finally construct the phonological percept, and compare it with the predicted phonological percept. Thus, they should detect semantic errors before syntactic errors, and syntactic errors before phonological errors.[11, 12]

3.2. Covert imitation and forward modeling in language comprehension

We now propose a model of prediction during language comprehension that incorporates the model of prediction during language production (see Fig. 5) in just the same way that the model of prediction during action perception (see Fig. 3) incorporates the model of prediction during action (see Fig. 2). It assumes that people make use of their ability to predict aspects of their own utterances to predict other people's utterances. Of course language comprehension involves structured linguistic representations (semantics, syntax, and phonology), and different predictions can be made at different levels. This makes prediction very powerful, because it is often the case that language is highly predictable at one linguistic level at least. An upcoming content word is sometimes predictable. Often, syntactic category can be predicted when the word itself cannot. On other occasions the upcoming phoneme is predictable. We propose that comprehenders make whatever linguistic predictions they can.

We assume that people can predict language using the association route and the simulation route. The association route is based on experience comprehending others' utterances. A comparable mechanism could be used to predict upcoming natural sounds (e.g., of a wave crashing against rocks). The simulation route is based on experience producing utterances. As in action perception, the simplest possibility is that the comprehender works out "what he would say under the circumstances" more quickly than the producer speaks, using a forward model. But just as with action perception, he needs to be able to represent what the speaker would say, not what he himself would say, and to do this, he needs to take into account the context. We

illustrate the model in Figure 6, in which the comprehender *A* covertly imitates *B*'s unfolding utterance (at time *t*) and uses forward modeling to derive the predicted utterance percept, which can then be compared with *A*'s percept of *B*'s actual utterance (at time *t+1*). Note that the account differs from Pickering and Garrod (2007), in which the comprehender simply predicts what he would say (and where these representations are not impoverished). Other-monitoring can take place at different linguistic levels, just like self-monitoring.

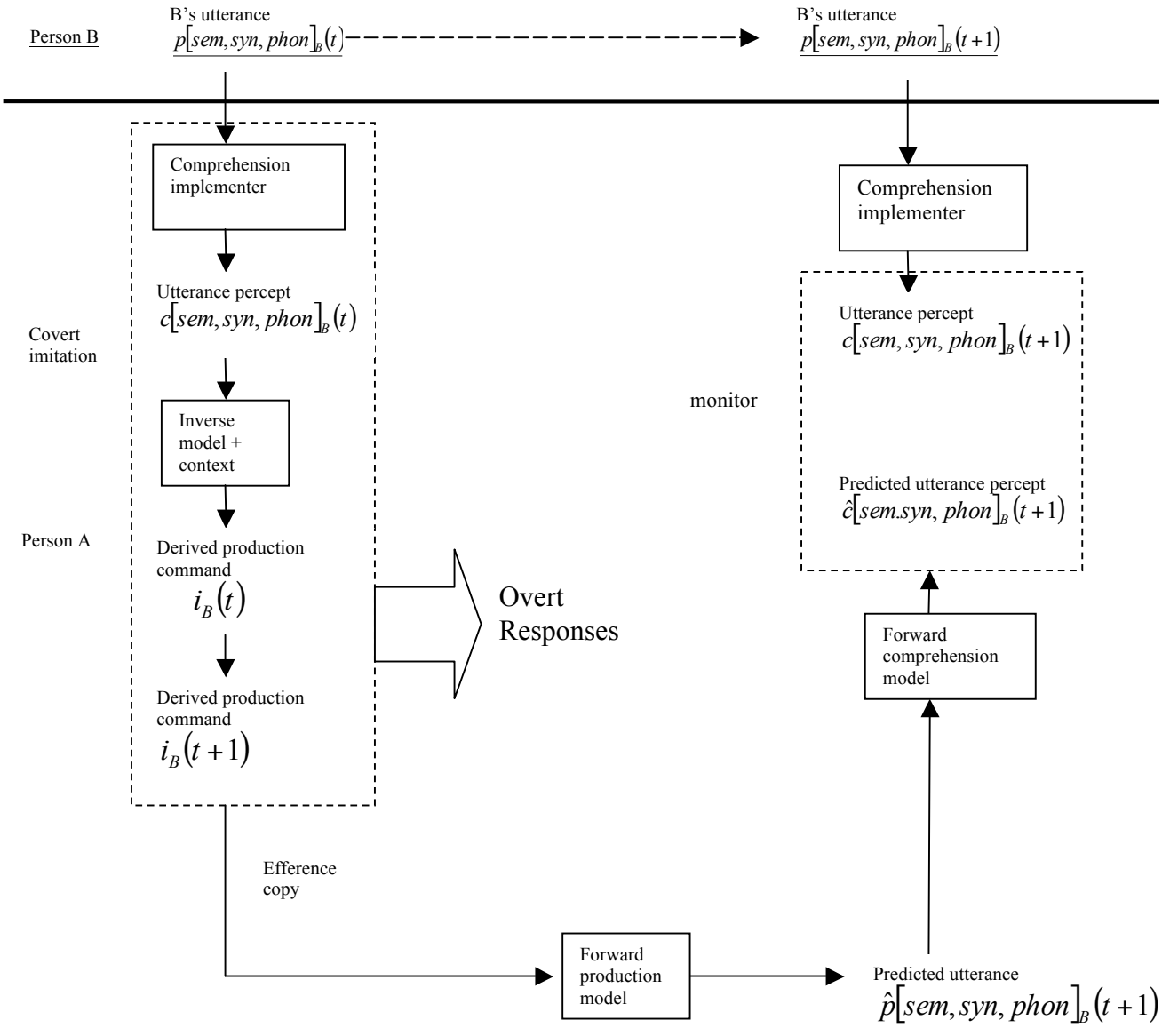

Figure 6. A model of the simulation route to prediction during comprehension in Person *A*. Everything above the solid line refers to *B*'s unfolding utterance (and is underlined). *A* predicts *B*'s utterance $\underline{p[sem, syn, phon]_B(t+1)}$ (i.e., its upcoming semantics, syntax, and phonology) given *B*'s utterance (i.e., at the present time *t*). To do this, *A* first covertly imitates *B*'s utterance. This involves deriving a representation of the utterance percept, and then using the inverse model and context (e.g., information about differences between *A*'s speech system and *B*'s speech system) to derive the production command $i_B(t)$ that *A* would use if *A* were to produce *B*'s utterance and from this the production command $i_B(t+1)$ associated with the next part of *B*'s utterance (e.g., phoneme or word). *A* now uses the same forward modeling as she does when producing an utterance (see Fig. 4) to produce her predictions of *B*'s utterance and of *B*'s utterance percept (at different linguistic levels). These predictions are typically ready before her comprehension of *B*'s utterance (the utterance percept). She can then compare the utterance percept and the predicted utterance percept at different linguistic levels (and therefore performs other-monitoring). Notice that *A* can also use the derived production command $i_B(t)$ to overtly imitate *B*'s utterance and the derived production command $i_B(t+1)$ to overtly produce the subsequent part of *B*'s utterance (see Overt Responses).

We now illustrate this account using a situation in which *A* (a boy) and *B* (a girl) have been given presents of an airplane and a kite respectively. *B* utters *I want to go out and fly the*. It is of course highly likely that *B* will say *kite*, which has $\underline{p[sem, syn, phon]_B(t+1)}$ = [KITE, noun, /kaIt/]. The utterance at time *t* is the semantics,

syntax, and phonology of *I want to go out and fly the*. To predict the situation at time *t+1*, *A* covertly imitates *B*'s production of *I want to go out and fly the*, and derives the production command that *A* would use to produce this utterance. *A* then derives the production command that *A* would use to produce the word that *B* would likely say (*kite*) and runs his forward models to derive his predicted utterance percept. If *A* feels sufficiently certain of what *B* is likely to say, *A* can act on this prediction, for example looking for a kite before *B* actually says *kite*. In addition, *A* can compare his prediction of what *B* will say with what *B* actually says using the monitor. In this case, *A* has no access to *B*'s representations during production, and therefore derives the utterance percept from *B*'s actual utterance. This means that *A* will access *B*'s phonology before *B*'s semantics. In this respect, other-monitoring is different from self-monitoring.

Importantly, *A* derives the production command of what *A* assumes *B* is likely to say (i.e., *kite*), rather than what *A* himself would be likely to say (i.e., *airplane*). This is the effect of using context together with the inverse model. It is consistent with the finding that comprehenders often pay attention to the speaker's state of knowledge (e.g., Hanna et al., 2003; Metzing & Brennan, 2003). However, comprehenders also show some "egocentric biases" (e.g., Keysar et al., 2000), a finding which is expected given that the comprehender's use of context cannot be perfect. Note also that predictions are driven by the forward production model, not by the production system itself. The production system would normally be too slow, given that the speaker should be at least as aware of what she is trying to say as the listener is. Use of the forward model also tends to cause some co-activation of the production system (as is typically the case when forward models are constructed). Such activation is not

central to prediction-by-simulation, but can lead to interference between production and comprehension, and serves as the basis for overt imitation (see Overt Responses in Fig. 6).

Note that Glenberg and Gallese (in press) recently proposed an Action Based Language (ABL) model of acquisition and comprehension that also uses paired inverse and forward models as in MOSAIC. The primary goal of ABL is to account for the content (rather than form) of language understanding, with language comprehension leading to the activation of action-based (embodied) representations. To do this, they specifically draw on evidence from mirror-neuron systems (see Section 4).

To assess our account, we discuss the evidence that comprehenders make predictions, that they covertly imitate what they hear, and that covert imitation leads to prediction that facilitates comprehension.

3.2.1. Evidence for prediction

A great deal of evidence shows that people predict other people's language (see Kutas et al., 2011; Pickering & Garrod, 2007, for reviews). In fact, expectation-based probabilistic models of language comprehension can explain much behavioral data (e.g., Hale, 2006; Levy, 2008), as can models of complexity that incorporate prediction (Gibson, 1998). Such findings are also compatible with accounts based on simple recurrent networks (Elman, 1990; see also Altmann & Mirkovic, 2009). But much of the evidence provides support for aspects of the account in Figure 6.

First, prediction occurs at different linguistic levels. Some research shows prediction of phonology (or associated visual or orthographic information). Delong et al. (2005) recorded ERPs while participants read sentences such as *The day was breezy so the boy went outside to fly…* They showed an N400 effect when the sentence ended with the less predictable *an airplane* than the more predictable *a kite*. The striking finding was that this effect occurred at *a* or *an*. It could not relate to ease of integration but must have involved prediction of the word and its phonological form (i.e., that it began with a consonant). Visser et al. (2006) found evidence of disruption when a highly predictable word was misspelt, presumably because it clashed with the predicted orthographic representation of the correct word.

Other experiments show prediction of syntax. Van Berkum et al. (2005) found disruption when Dutch readers and listeners encountered an adjective that did not agree in grammatical gender with an upcoming, highly predictable noun. Staub and Clifton (2006) found that people read *or the subway* faster after *The team took either the train* ...than after *The team took the train*: *Either* makes the sentence more predictable by ruling out an analysis in which *or* starts a new clause. Similarly, early syntactic anomaly effects in the ERP record are affected by whether the linguistic context predicts a particular syntactic category for the upcoming word or whether the linguistic context is compatible with different syntactic categories (Lau et al., 2006), and reading times are affected by predicted syntactic structure associated with ellipsis (Yoshida et al., in press).

Clear evidence for semantic prediction comes from eye-tracking studies in which participants listen to sentences while viewing arrays of objects or depictions of events.

They started looking at edible objects more than at inedible objects while hearing *the man ate the* (but not when *ate* was replaced with *moved*; Altmann & Kamide, 1999). These predictive eye movements do not just depend on the meaning (or lexical associates) of the verb, but are affected by properties of the prior context (Kaiser & Trueswell, 2004; Kamide et al., 2003) or other linguistic information such as prosody (Weber et al., 2006). People also predict the upcoming event as well as the upcoming referent (Knoeferle et al., 2005).

Some of these studies do not clearly demonstrate that the predictions are used more rapidly than would be possible with the production implementer. The eye-tracking studies reveal faster predictions, but they may show prediction of semantics (e.g., edible things) rather than a word (e.g., *cake*). However, recent MEG evidence shows sensitivity to syntactic manipulations in little over 100 ms, in visual cortex (Dikker et al., 2009, 2010). For example, the M100 was affected by predictability when the upcoming word looked like a typical noun (e.g., *soda*) but not when it did not (e.g., *infant*). Presumably these results cannot be due to integration, because activation of the grammatical category of this word (as part of the process of lexical access) could not occur so rapidly or in an area associated with visual form. Instead, the comprehender must predict both syntactic categories and the form most likely associated with those categories, then match those predictions against the upcoming word. Given that syntactic processing does not take place in the visual cortex (or indeed so quickly), these results reflect the visual correlates of syntactic predictions. They suggest that the comprehender constructs a forward model of visual properties (presumably closely linked to phonological properties) on the basis of sentence

context and can compare these predicted visual properties with the input within around 100ms.

Dikker and Pyllkänen (2011) found evidence for form prediction on the basis of semantics. Participants saw a picture followed by a noun phrase that matched (or mismatched) the specific item in the picture (e.g., an apple) or the semantic field (e.g., a collection of food). They found an M100 effect in visual cortex associated with matching the specific item but not the semantic field, suggesting that participants predicted the form of the specific word.

Kim and Lai (in press) conducted a similar study to Vissers et al. (2006) and found a P130 effect for contextually supported pseudowords (e.g., ... *bake a ceke*) but not for non-supported pseudowords (e.g., *bake a tont*). In contrast, an N170 effect occurred for non-supported pseudowords (and non-words). The N170 may relate to lexical access, but the P130 occurs before lexical access can have occurred and again appears to reflect a forward model, in which the comprehender predicts the form of the word (*cake*) and matches the input to that form.[13] In conclusion, these four studies support forward modeling, but they do not discriminate between prediction-by-simulation and prediction-by-association.

3.2.2. Evidence for covert imitation

Much evidence suggests that comprehenders activate mechanisms associated with aspects of language production. As we have noted, there appear to be integrated circuits associated with production and comprehension (Pulvermüller & Fadiga, 2010). For example, the lateral part of the precentral cortex is active when listening

to /p/ and producing /p/, whereas the inferior precentral area is active when listening to /t/ and producing /t/ (Pulvermüller et al., 2006; see also Vigneux et al., 2006; S.M. Wilson et al., 2004). We have also noted that tongue and lip muscles are activated during listening to speech but not other sounds (Fadiga et al., 2002; Watkins et al., 2003). More specifically, Yuen et al. (2010) found that listening to incongruent /t/ initial distracters leaves articulatory traces on simultaneous production of /k/ or /s/ target phonemes, in the form of increased alveolar contact. Furthermore, this effect only occurred with incongruent distracters and not with distinct but congruent distracters (e.g., /g/ for /k/ targets). These results suggest that perceiving speech results in selective, covert, and automatic activation of the speech articulators. Note that these findings show activation of the production implementer (not a forward model).

There is also much evidence for both overt imitation and overt completion. Speakers tend to imitate the speech of other people after they have comprehended it (see Pickering & Garrod, 2004), and to repeat each other's choice of words and semantics (Garrod & Anderson, 1987), syntax (Branigan et al., 2000), and sound (Pardo, 2006). Such imitation can be rapid and apparently automatic; for instance, speakers are almost as quick imitating a phoneme as they are making a simple response to it (Fowler et al., 2003). Speakers also tend to complete others' utterances. For example, Wright and Garrett (1984; see also Peterson et al., 2001) found that participants were faster at naming a word that was syntactically congruent with prior context than a word that was incongruent (even though neither word was semantically appropriate). Moreover, people regularly complete each other's utterances during dialogue (e.g., 1a-c); see for example Clark and Wilkes-Gibbs (1986). Rapid overt

imitation and overt completion are of course compatible with prior covert imitation (see Overt Responses in Fig. 6).

3.2.3. Evidence that covert imitation facilitates comprehension via prediction

The previous sections presented evidence that comprehenders make rapid predictions and that they covertly imitate what they hear. But are they causally linked in the way suggested in Figure 6? The evidence for prediction could involve the association route. In addition, covert imitation of language could be used postdictively, to facilitate memory (as a component of rehearsal) or to assist when comprehension leads to incomplete analyses or fails to resolve an ambiguity (see Garrett, 2000).

However, recent evidence suggests that covert imitation drives predictions that facilitate comprehension. Adank and Devlin (2010) used fMRI to show that during adaptation to time-compressed speech there was increased activation in the left ventral premotor cortex, an area concerned with planning articulation. This suggests that participants covertly imitated the compressed speech as part of the adaptation process that facilitates comprehension. Adank et al. (2010) found that overt imitation of sentences in an unfamiliar accent facilitated comprehension of subsequent sentences in that accent, in the context of noise. This suggests that overt imitation adapts the production system to unfamiliar accent and therefore that the production system plays a causal role in comprehension as it occurs.

Ito et al. (2009) manipulated listeners' cheeks as they heard words on a continuum between *had* and *head*. When the skin of the cheek was stretched upwards, listeners reported hearing *head* in preference to *had*; when the skin was stretched downwards

they reported hearing *had* in preference to *head*. Because production of *had* requires an upward stretch of cheek skin and production of *head* a downward stretch, the results suggest that proprioceptive feedback from the articulators causally affected comprehension (see also Sams et al., 2005). These results could conceivably be postdictive, perhaps relating to reconstruction occurring during self-report. Clearer evidence comes from Möttönen and Watkins (2009), who used repetitive TMS to temporarily disrupt specific articulator representations during speech perception. Disrupting lip representations in left primary motor cortex impaired categorical perception of speech sounds involving the lips (e.g., /*ba*/-/*da*/), but not of those involving other articulators (e.g., /*ka*/-/*ga*/). Furthermore, D'Ausilio et al. (2009) found that double-pulse TMS administered to motor cortex controlling lips speeded up and increased accuracy of responses to lip-articulated phonemes, whereas TMS administered to motor cortex controlling tongues speeded up and increased accuracy of responses to tongue-articulated phonemes. More recently, D'Ausilio et al. (2011) had participants repeatedly hear a pseudoword (e.g., *birro*) and used TMS to reveal immediate appropriate articulatory activation (associated with *rr*) if they heard the first part of the same word (*bi*, when coarticulated with *rro*) than if they heard the first part of a different word (*bi*, when coarticulated with *ffo*). Thus, covert imitation facilitates speech recognition as it occurs and before it occurs.

A different type of evidence comes from Stephens et al. (2010), who correlated cortical BOLD signal changes between speakers and listeners during the course of a narrative. There was aligned neural activation in many cortical areas at different lags. Sometimes the speaker's neural activity preceded that of the listener, but sometimes the listener's activity preceded that of the speaker. Importantly, listeners whose

activity preceded that of the speaker showed better comprehension, suggesting that covert imitation led to prediction and that this prediction facilitated comprehension.

Finally, speakers may use the production system to predict upcoming words (and events) in relation to scenes. In "visual world" experiments, participants activate the phonology associated with the names of the objects (see Huettig et al., 2011). For example, Huettig and McQueen (2007) had participants listen to a sentence and found that they look at a picture whose name was phonologically related to a target word (cf. Allopena et al., 1998) when they viewed the pictures for 2-3s before hearing the target word but not when they viewed the pictures for 200ms. In the former case, they presumably had enough time to access the phonological form of the name of the picture.

These studies therefore show that the results of covert imitation have immediate effects on comprehension as a result of prediction. Moreover, we have shown that covert imitation and prediction take place at many linguistic levels. Together, all of these findings provide support for the model of prediction-by-simulation in Figure 6. Of course, comprehenders may also perform prediction-by-association, just as they can for predicting non-linguistic events (see General Discussion).

3.3. Interactive language

Interactive conversation is a highly successful form of joint activity. It appears to be very complex, with interlocutors having to switch between production and comprehension, perform both acts at once, and develop their plans on the fly (Garrod

& Pickering, 2004). Just as we explained joint actions by combining the accounts of action and action perception (see Fig. 4), so we explain conversation by combining the accounts of language production and comprehension (as in Figs 5 and 6).

Figure 7 shows how *A* and *B* can both predict *B*'s upcoming utterance (using prediction-by-simulation). *A* comprehends *B*'s current utterance and then uses covert imitation and forward modeling; *B* formulates his forthcoming production command and uses forward modeling based on that command. If successful, they should make similar predictions about *B*'s upcoming utterance, and can use those predictions to coordinate (i.e., have a well-organized conversation). Note that they can both compare their predictions with *B*'s forthcoming utterance when produced, with *A* using other-monitoring and *B* using self-monitoring. In addition, *A* and *B* can also predict *A*'s forthcoming utterance (so both *A* and *B* predict both *A* and *B*). Of course these predictions will be related to *A* and *B*'s predictions of *B*'s utterance (e.g., they might both predict *A*'s upcoming word and *B*'s response following that word), in a way that will reduce the difficulty of making two predictions.

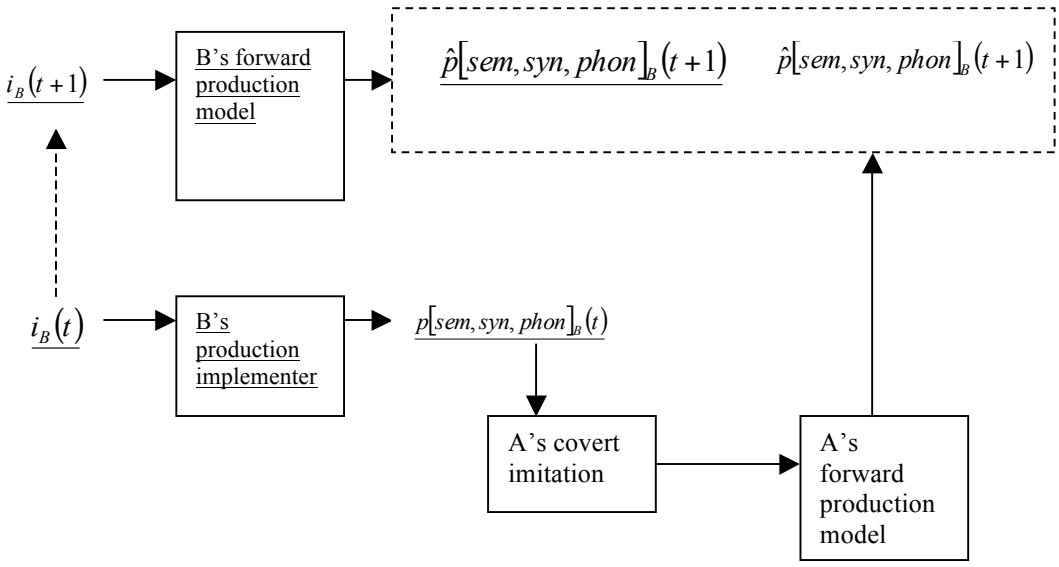

Figure 7. *A* and *B* predicting *B*'s forthcoming utterance (with *B*'s processes and representations underlined). *B*'s production command $i_B(t)$ feeds into *B*'s production implementer and leads to *B*'s utterance $p[sem, syn, phon]_B(t)$. *A* covertly imitates *B*'s utterance and uses *A*'s forward production model to predict *B*'s forthcoming utterance (at time *t*+1). *B* simultaneously constructs the next production command (the dotted line indicates that this command is causally linked to the previous action command for *B* but not *A*) and uses *B*'s forward production model to predict *B*'s forthcoming utterance. If *A* and *B* are coordinated, then *A*'s prediction of *B*'s utterance and *B*'s prediction of *B*'s utterance (in the dotted box) should match. Moreover, they may both match *B*'s forthcoming (actual) utterance at time *t*+1 (not shown).

Our account can explain how interlocutors can be so well coordinated, for example why intervals between turns are so close to 0 ms (Sacks et al., 1974; M. Wilson & Wilson, 2005) and why interlocutors are so good at using the content of utterances to predict when they are likely to end (de Ruiter et al., 2006). Moreover, it accords with the treatment of dialogue as coordinated joint activity, in which partners are able to take different roles as appropriate (Clark, 1996). It can also explain the existence and speed of completions, overt imitation (e.g., Branigan et al., 2000; Fowler et al., 2003; Garrod & Anderson, 1987), and (assuming links between intentions) rapid complementary responses (as in answers to questions).

We illustrate with the following extract (from Howes et al. 2011).

2a ---- A: …and then we looked along one deck, we were high up, and down below there were rows of, rows of lifeboats in case you see

2b ---- B: there was an accident

2c ---- A: of an accident

In (2b-c), *B* speaks at the same time as *A* and has a similar understanding to *A*. *B* interrupts *A*, and it is clear that *B* must be as ready to contribute as *A*. Because *B* completes *A*'s utterance without delay, it would not be possible for *B* to produce (2b) by comprehending (2a) and then preparing a response "from scratch," as traditional "serial monologue" accounts assume (see Fig. 1). Instead, we assume that *B* covertly imitates *A*'s utterance, determines *A*'s current production command, determines *A*'s forthcoming production command, and produces an overt completion (see Overt Responses in Fig. 6). Thus *B*'s response is time-locked to *A*'s contribution. In fact, (2b) is different from *A*'s own continuation (2c). The two continuations are syntactically different (though both grammatical) but semantically equivalent, thereby indicating that prediction can occur differently at different linguistic levels. Note that prediction-by-association might allow *B* to predict *A*'s continuation, but would not explain the rapidity of *B*'s response, as *B* would also have to produce the continuation "from scratch".

During conversation, interlocutors tend to become aligned with each other at different linguistic levels, and such alignment appears to underlie mutual understanding (Pickering & Garrod, 2004). Our account can help explain this process, because the close link between production and comprehension leads to tightly yoked representations for comprehension and production, and allows those representations to

be used extremely rapidly (see Garrod & Pickering, 2009). But the relationship also works the other way. Prediction during comprehension is facilitated when the interlocutors are well-aligned, because the comprehender is more likely to predict the speaker accurately (and the speaker is more likely to predict the comprehender's response, as in question-answering). One effect of this is that *B*'s prediction of what *A* is going to say is more likely to accord with what *B* would be likely to say if *B* spoke at that point. In other words, *B*'s prediction of *B*'s completion becomes a good proxy for *B*'s prediction of *A*'s completion, and so there is less likelihood of an egocentric bias.[14] In fact, linguistic joint action is more likely to be successful and well-coordinated than many other forms of joint action, precisely because the interlocutors communicate with each other and share the goal of mutual understanding.

4 GENERAL DISCUSSION

Our accounts of comprehension and dialogue assign a central role to simulation. We discuss three aspects of simulation: the relationship between "on-line" and "off-line" simulation, between prediction-by-simulation and prediction-by-association, and between simulation and embodiment. We conclude by explaining how our account provides an integrated theory of production and comprehension.

We have focused on "on-line" simulation, when the comprehender wishes to predict the speaker in real time. However, our notion of simulation is compatible with the simulation theory of mind-reading (Goldman, 2006; see Gordon, 1986), which is primarily used to explain "off-line" understanding of others. In our account, the

comprehender "enters" the simulation during covert imitation, and "exits" after constructing the predicted utterance percept (see Fig. 6). As in our account, Goldman assumes that people covertly imitate as though they were character acting – attempting to resemble their target as much as possible, and then running things forward as well as they can. This means that the derived action command is "supposed" to be the action command of the target, but it incorporates any changes that are required because of bodily differences. (I can walk like Napoleon by putting my hand inside my jacket and seeing how this affects my gait, but I cannot shrink.) In addition, the perceiver may fail to derive the actor's action command correctly, in which case her covert imitation is biased toward her own proclivities.

The important difference between such accounts and ours is that they do not assume forward models and therefore assume that covert imitation uses the action implementer (but inhibiting overt responses). This may be appropriate for "off-line" reasoning but is too slow for prediction (see Goldman, 2006, pp. 213-217; Hurley, 2008b). Goldman's account uses simulation as an alternative to constructing a theory of the other person's mind. In contrast, our account uses simulation to facilitate processing, which is particularly important when behavior is rapid (as in Grush, 2004; Prinz, 2006). Clearly this is the case for language processing.

However, prediction-by-simulation can also be applied "off-line" as part of the process of thinking and planning (as indeed can prediction-by-association). For example, a speaker might think about the likely consequences of producing a particular utterance, both for her own subsequent utterances and perhaps more importantly for the responses that addressees are likely to produce. She might do this

by constructing a predicted utterance percept, using forward modeling. She could also construct an utterance percept (without articulating), using the production implementer and comprehension implementer (see top right of Fig. 5 and discussion in Section 3.1), as she would typically have enough time to do so. Assuming co-activation, "off-line" predictions may often involve both the production implementer and forward modeling. See Pezzulo (2011) for a related discussion.

Our account assigns a central role to prediction-by-simulation, but it assumes that language comprehension and dialogue also involve prediction-by-association. We propose that comprehenders will emphasize simulation when they are (or appear to be) similar to the speaker because simulation will tend to be accurate. These similarities might relate to cultural or educational background or dialect, or alternatively to speed or style of language processing. In addition, simulation will be emphasized during dialogue because the interlocutors will tend to become aligned (Pickering & Garrod, 2004), and simulation will tend to persist among those in close relationships (who continue to be aligned). In addition, simulation may also be primed during dialogue, because the fact that the comprehender also has to speak may activate mechanisms associated with production. In contrast, prediction-by-association will be emphasized when the comprehender is less similar to the producer, as for example when the comprehender is a native adult speaker of the language and the producer is a non-native speaker or a child, or when the comprehender does not have the opportunity to speak (as in reading).

We therefore assume that comprehenders emphasize whichever route is likely to be more accurate (given that they should both be fast enough). It may also be that

prediction-by-association is more accurate for simple, "one-step" associations between a current and a subsequent state. For example, people can straightforwardly predict that a person who looks confused is likely to respond slowly. In contrast, prediction-by-simulation is likely to be more complex, because it makes use of the structure inherent in the speaker's own production mechanisms.

Of course, comprehenders may combine prediction-by-simulation and prediction-by-association. They make use of the same representational vocabulary and hence the mental states are the same; the association route simply involves a different (and more straightforward) set of mappings than the simulation route. Informally, if I see that you are about to speak, I can predict your utterances by combining my experiences of how people like you have spoken and my experiences of how I have spoken under similar circumstances.

There is a lot of current interest in the extent to which language is embodied (see Barsalou, 1999; Fischer & Zwaan, 2008). Such literature focuses on embodiment of content, in which the conceptual content of language is represented in "modal" (i.e., action-based or perceptual) terms (e.g., *kick* is represented in terms of the movements associated with kicking). It is supported by strong evidence from behavioral experiments (e.g., Glenberg & Kaschak, 2002) and cognitive neuroscience (e.g., Desai et al., 2010). In contrast, our account is concerned with embodiment of form, which Gallese (2008) called the vehicle level. It assumes that comprehension involves aspects of production, which is a form of action; by definition, production is embodied at the form level. Interestingly, Glenberg and Gallese (in press) used covert imitation and prediction in an account primarily concerned with content embodiment.

Of particular interest, it explains why representational gesture tends to co-occur with speech by arguing that speaking activates the corresponding action and that the need to perform the action of articulation prevents the inhibition of related gestural actions (see Hostetter & Alibali, 2008).

Both our account and embodied accounts seek to abandon the "cognitive sandwich" (Hurley, 2008a). Our account assumes that producers use comprehension processes and comprehenders use production processes; whereas embodied accounts assume that producers and comprehenders use perceptual and motor representations associated with the meaning of what they are communicating. Our account does not require such embodiment but is compatible with it.

5. CONCLUSION

Traditional accounts of language assume separate processing "streams" for production and comprehension. They adopt the "cognitive sandwich", a perspective which is incompatible both with the demands of communication and with extensive data indicating that production and comprehension are tightly interwoven. We therefore proposed an account of language processing that abandons the cognitive sandwich. This account assumes a central role to prediction in language production, comprehension, and dialogue. By building on research in action and action perception, we propose that speakers use forward models to predict aspects of their upcoming utterances and listeners covertly imitate speakers and then use forward models based on their own potential utterances to predict what the speakers are likely to say. The account helps explain the rapidity of production and comprehension and

the remarkable fluency of dialogue. It thereby provides the basis for a psychological account of human communication.

ACKNOWLEDGEMENTS

We thank Dale Barr, Martin Corley, Chiara Gambi, and Laura Menenti for their comments, and acknowledge support of ESRC Grants no. RES-062-23-0376 and RES-060-25-0010.

REFERENCES

Adank, P. & Devlin, J.T. (2010) On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. NeuroImage 49: 1124–1132.

Adank, P., Hagoort, P. & Bekkering, H. (2010) Imitation improves language comprehension. Psychological Science 21: 1903-1909.

Allopena, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998) Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language 38: 419–439.

Altmann, G. T. M. & Kamide, Y. (1999) Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition 73: 247–264.

Altmann, G. T. M. & Mirkovic, J. (2009) Incrementality and prediction in human sentence processing. Cognitive Science 33: 583-609.

Barsalou, L. 1999: Perceptual symbol systems. Behavioral and Brain Sciences 22: 577–600.

Bavelas, J.B., Coates, L. & Johnson, T. (2000) Listeners as co-narrators. Journal of Personality and Social Psychology 79: 941-952.

Ben Shalom, D. & Poeppel, D. (2008). Functional anatomic models of language: assembling the pieces. The Neuroscientist 14:119-27.

Blackmer, E. R. & Mitton, J. L. (1991) Theories of monitoring and the timing of repairs in spontaneous speech. Cognition 39: 173-194.

Blakemore, S-J., Frith,C.D. & Wolpert, D.M. (1999) Spatio-temporal prediction modulates the perception of self-produced stimuli. Journal of Cognitive Neuroscience 11: 551-559.

Bock J.K. (1996) Language production: Methods and methodologies. Psychonomic Bulletin & Review 3: 395–421.

Bock, J. K. & Levelt, W. J. M. (1994) Language production: Grammatical encoding. In: Handbook of psycholinguistics, ed. M. A. Gernsbacher. Academic Press.

Bock, K. & Miller, C.A (1991) Broken agreement. Cognitive Psychology 23: 45-93.

Branigan, H.P., Pickering, M.J. & Cleland, A.A. (2000) Syntactic coordination in dialogue. Cognition 75: B13-B25.

Brown, R. & McNeill, D. (1966) The “tip of the tongue” phenomenon. Journal of Verbal Learning and Verbal Behavior 5: 325-337.

Chang, F., Dell, G.S., & Bock, K. (2006) Becoming syntactic. Psychological Review 113: 234-272.

Chartrand, T.L. & Bargh, J.A. (1999) The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology 76: 893-910.

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

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

Corley, M, Brocklehurst, PH, & Moat, HS (2011). Error biases in inner and overt speech: Evidence from tongue twisters. Journal of Experimental Psychology: Learning, Memory, and Cognition 37: 162-75.

Csibra, G. & Gergely, G. (2007) 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. Acta Psychologica 124: 60-78.

D'Ausilio, A., Jarmolowska, J., Busan, P., Bufalari, & Craighero, L. (2011) Tongue corticospinal modulation during attended verbal stimuli: Priming and coarticulation effects. Neuropsychologia 49: 3670-3676.

D'Ausilio, A., Pulvermuller, F., Salmas, P., Bufalari, I., Begliomini, C. & Fadiga, L. (2009) The motor somatotopy of speech perception. Current Biology 19: 381–385.

Davidson, P.R. & Wolpert, D.M. (2005) Widespread access to predictive models in motor system: a short review. Journal of Neural Engineering 2: S313-S319.

DeLong, K.A., Urbach, T.P. & Kutas, M. (2005) Probabilistic word pre-activation during comprehension inferred from electrical brain activity. Nature Neuroscience 8: 1117-1121.

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

Dell, G. S. (1986) A spreading-activation theory of retrieval in sentence production. Psychological Review 93:283-321.

Dell G. S. (1988) The retrieval of phonological forms in production: Tests of predictions from a connectionist model. Journal of Memory and Language 27: 124–142

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. Psychological Review 104: 801–838

Desai, R. H., Binder, J. R., Conant, L. L. & Seidenberg, M. S. (2010) Activation of sensory-motor areas in sentence comprehension. Cerebral Cortex 20: 468–478.

Dijksterhuis, A. & Bargh, J. A. (2001) The perception-behavior expressway: Automatic effects of social perception on social behavior. In: Advances in experimental social psychology, vol. 33, ed. M. P. Zanna. Academic Press.

Dikker, S. & Pylkkänen, L. (2011) Before the N400: Effects of lexical-semantic violations in visual cortex. Brain & Language 118: 23-28.

Dikker, S., Rabagliati, H. & Pylkkänen, L. (2009) Sensitivity to syntax in visual cortex. Cognition 110: 293-321.

Dikker, S., Rabagliati, H., Farmer, T.A., & Pylkkänen, L. (2010) Early occipital sensitivity to syntactic category is based on form typicality. Psychological Science 21: 629-634.

Di Pellegrino G, Fadiga L, Fogassi L, Gallese V & Rizzolatti G. (1992) Understanding motor events: A neurophysiological study. Experimental Brain Research 91:176–80.

Echterhoff, G., Higgins, E. T. & Levine, J. M. (2009) Shared reality: Experiencing commonality with others' inner states about the world. Perspectives on Psychological Science 4: 496-521.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolati, G. (2002) Speech listening specifically modultes the excitability of tongue muscles: A TMS study. European Journal of Neuroscience 15: 399-402.

Federmeier, K. D. (2007) Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology* 44: 491-505.

Ferreira, F. (2003) The misinterpretation of noncanonical sentences. Cognitive Psychology 47: 164-203.

Ferreira, V. S. (1996) Is it better to give than to donate? Syntactic flexibility in language production. Journal of Memory and Language 35: 724-755.

Fodor, J.A. (1983) The Modularity of Mind. Cambridge, MA: MIT Press.

Fowler, C. A., Brown, J., Sabadini, L. & Weihing, J. (2003) Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. Journal of Memory and Language 49: 296-314.

Frazier, L. (1987) Sentence processing: A tutorial review. In: M. Coltheart (Ed.), Attention and performance XII: The psychology of reading (pp. 559-586). Erlbaum.

Freyd, J.J. & Finke, RA. (1984) Representational momentum, Journal of Experimental Psychology: Learning, Memory, and Cognition 10: 126–132

Gallese, V. (2008) Mirror neurons and the social nature of language: The neural exploitation hypothesis. Social Neuroscience 3: 317–333.

Garrett, M. (1980) Levels of processing in speech production. In: Language production, vol. 1, ed. B. Butterworth. Academic Press

Garrett, M. (2000) Remarks on the architecture of language production systems. In Language and the Brain: Representation and Processing. In: Y. Grodzinsky & L.P. Shapiro, eds, pp. 31–69. Academic Press

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

Garrod, S. & Pickering, M.J. (2004) Why is conversation so easy? Trends in Cognitive Sciences 8: 8-11.

Garrod, S. & Pickering, M.J. (2009) Joint action, interactive alignment and dialogue. Topics in Cognitive Science 1: 292-304.

Gaskell, G. (2007) Oxford Handbook of Psycholinguistics. Oxford University Press.

Gergely, G., Bekkering, H. & Király, I. (2002) Developmental psychology: Rational imitation in preverbal infants. Nature 415: 755.

Gibson, E. (1998) Linguistic complexity: Locality of syntactic dependencies. Cognition 68: 1–76.

Glenberg, A.M. & Gallese, V. (in press) Action-based language: A theory of language acquisition, comprehension, and production. Cortex.

Glenberg, A.M. & Kaschak, M.P. (2002) Grounding language in action. Psychonomic Bulletin & Review 9: 558-565.

Goldberg, A.E. (1995) Constructions: A Construction Grammar approach to argument structure. University of Chicago Press.

Goldman, A. I. (2006) Simulating minds. The philosophy, psychology, and neuroscience of mindreading. Oxford University Press.

Gordon, R. (1986) Folk psychology as simulation. Mind and Language 1: 158-171.

Graf, M., Schütz-Bosbach, S. & Prinz, W. (2010) Motor involvement in object perception: Similarity and complementarity. In: G. Semin & G. Echterhoff, eds., Grounding sociality: Neurons, minds, and culture, 27-52. Psychology Press.

Gregoromichelaki, E., Kempson, R., Purver, M., Mills, J.G., Cann, R., Meyer-Viol, W. & Healey, P.G.T. (2011) Incrementality and intention-recognition in utterance processing, Dialogue and Discourse 2: 199–233.

Grush, R. (2004) The emulation theory of representation: Motor control, imagery, and perception. Behavioral and Brain Sciences 27: 377-435.

Häberle, A., Schütz-Bosbach, S., Laboissière, R. & Prinz, W. (2008) Ideomotor action in cooperative and competitive settings. Social Neuroscience 3: 26–36.

Hale, J. (2006) Uncertainty about the rest of the sentence. Cognitive Science 30: 609-642.

Hanna, J. E., Tanenhaus, M. K. & Trueswell, J. C. (2003) The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language 49:43–61.

Harley, T. (2007) The psychology of language: From data to theory, 3rd Edition. Psychology Press.

Hartsuiker, R. J., & Kolk, H. H. J. (2001) Error monitoring in speech production: A computational test of the perceptual loop theory. Cognitive Psychology 42: 113-157.

Haruno, M., Wolpert, D.M. & Kawato, M. (2001) MOSAIC Model for sensorimotor learning and control. Neural Computation 13: 2201-2220.

Haruno, M., Wolpert, D.M. & Kawato, M. (2003) Hierarchical MOSAIC for movement generation. International Congress Series, 1250, 575-590.

Haueisen, J. & Knösche, T.R. (2001) Involuntary motor activity in pianists evoked by music perception. Journal of Cognitive Neuroscience 13: 786-792.

Hawkins, J.A. (1994) A performance theory of order and constituency. Cambridge University Press.

Heim, S., Opitz, B., Müller, K. & Friederici, A.D. (2003) Phonological processing during language production: fMRI evidence for a shared production-comprehension network. Cognitive Brain Research 12: 285-29.

Heinks-Maldonado, T.H., Nagarajan, S.S. & Houde, J.F. (2006) Magnetoencephalographic evidence for a precise forward model in speech production. NeuroReport 17: 1375–1379.

Hostetter, A.B. & Alibali, M.W. (2008) Visual embodiment: Gesture as simulated action. Psychonomic Bulletin & Review 15: 495-514.

Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. (2001) The theory of event coding (TEC): A framework for perception and action planning. Behavioral and Brain Sciences 24: 849-878.

Howes, C., Purver, M., Healey, P.G.T., Mills, G.J. & Gregoromichelaki, E. (2011) Incrementality in dialogue: Evidence from compound contributions. Dialogue and Discourse 2: 279-311.

Huettig, F. & McQueen, J. M. (2007) The tug of war between phonological, semantic and shape information in language-mediated visual search. Journal of Memory and Language 57: 460−482.

Huettig, F., Rommers, J. & Meyer, A.S. (2011) Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica 137: 151-171.

Huettig, F. & Hartsuiker, R. J. (2010) Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. Language and Cognitive Processes 25: 347−374.

Hurley, S. (2008a) The shared circuits model (SCM): How control, mirroring and simulation enable imitation, deliberation, and mindreading. Behavioral and Brain Sciences 31: 1-22.

Hurley, S. (2008b) Understanding simulation. Philosophy and Phenomenological Research 77: 755-774.

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

Ito, T., Tiede, M. & Ostry, D.J. (2009) Somatosensory function in speech perception. Proceedings of the National Academy of Sciences 106: 1245-1248.

Jaeger, T. F. (2010) Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology 61: 23-62.

Kaiser, E. & Trueswell, J.C. (2004) The role of discourse context in the processing of a flexible word-order language. Cognition 94, 113-47.

Kamide, Y., Altmann, G.T.M., & Haywood, S.L. (2003) Prediction and thematic information in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language 49: 133-156.

Keysar, B., Barr, D.J., Balin, J.A. & Brauner, J.S. (2000) Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science 11:32-38.

Kilner, J.M., Paulignan,Y. & Blakemore, S.-J. (2003) An interference effect of observed biological movement on action. Current Biology 13: 522-525.

Kim, A. & Lai, V. (in press) Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*.

Knoblich, G. & Flach, R. (2001) Predicting action effects: Interaction between perception and action. Psychological Science 12: 467-472.

Knoblich, G., Seigerschmidt, E., Flach, R. & Prinz, W. (2002) Authorship effects in the prediction of handwriting strokes: Evidence for action simulation during action perception. Quarterly Journal of Experimental Psychology 55A: 1027–1046.

Knoeferle, P., Crocker, M. W., Scheepers, C. & Pickering, M. J. (2005) The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. Cognition 95: 95-127.

Kutas, M., DeLong, K.A., Smith, N.J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In: M. Bar (Ed.), *Predictions in the Brain: Using Our Past to Generate a Future* (pp. 190-207), Oxford University Press.

Lakin, J., & Chartrand, T.L. (2003) Using nonconscious behavioral mimicry to create affiliation and rapport. Psychological Science 14: 334-339.

Lau, E., Stroud, C., Plesch, S. & Phillips, C. (2006) The role of structural prediction in rapid syntactic analysis. Brain and Language 98: 74–88.

Laver, J. D. M. (1980) Monitoring systems in the neurolinguistic control of speech production. In: V. A. Fromkin, ed., Errors in linguistic performance: Slips of the tongue, ear, pen, and hand. Academic Press.

Levelt, W. J. M. (1983) Monitoring and self-repair in speech. Cognition 14:41-104

Levelt, W. J. M. (1989) Speaking: From intention to articulation. MIT Press.

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999) A theory of lexical access in speech production. Behavioral and Brain Sciences 22: 1-75.

Levy, R. (2008) Expectation-based syntactic comprehension. Cognition 106: 1126-1177.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994) The lexical nature of syntactic ambiguity resolution. Psychological Review 101: 676-703.

MacKay, D.G. (1982) The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behaviors. Psychological Review 89: 483-506.

Mar, R.A. (2004) The neuropsychology of narrative: Story comprehension, story production and their interrelation. Neuropsychologia 42: 1414-34

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10: 29-63.

Menenti, L., Gierhan, S.M.E., Segaert, K. & Hagoort, P. (2011) Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by fMRI. Psychological Science 22: 1173-1182

Metzing, C. & Brennan, S. E. (2003) When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. Journal of Memory and Language 49: 201-213.

Miall, R.C., Stanley, J., Todhunter, S., Levick, C., Lindo, S. & Miall, J.D. (2006) Performing hand actions assists the visual discrimination of similar hand postures. Neuropsychologia 44: 966-976.

Motley, M.T., Camden, C.T. & Baars, B.J. (1982) Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. Journal of Verbal Learning and Verbal Behavior 21: 578-594.

Mottonen, R. & Watkins, K. E. (2009) Motor representations of articulators contribute to categorical perception of speech sounds. Journal of Neuroscience 29: 9819–9825.

Mukamel, R., Ekstrom, A.D., Kaplan, J., Iacoboni, M. & Fried, I. (2010) Single-neuron responses in humans during execution and observation of actions. Current Biology 20: 750-756.

Neda, Z., Ravasz, Y., Brechet, T. Vicsek, T. & Barabasi, A.L. (2000) The sound of many hands clapping. Nature 403: 849.

Newman-Norlund, R. D., van Schie, H. T., van Zuijlen, A. M. J. & Bekkering, H.

(2007) The mirror neuron system is more active during complementary compared with imitative action. Nature Neuroscience 10: 817–818.

Nozari, N., Dell, G.S., & Schwartz, M.F. (2011) Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. Cognitive Psychology 63: 1-33.

Oppenheim, GM, & Dell, GS (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. Cognition 106: 528‑537.

Oppenheim, GM, & Dell, GS (2010). Motor movement matters: the flexible abstractness of inner speech. Memory & Cognition 38: 1147-60.

Pardo, J.S. (2006) On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America 119: 2382-2393.

Paus, T., Perry, D.W., Zatorre, R.J., Worsley, K.J., & Evans, A.C. (1996) Modulation of Cerebral Blood Flow in the Human Auditory Cortex During Speech: Role of Motor-to-sensory Discharges. European Journal of Neuroscience. 8, 2236-2246.

Peterson, R. R., Burgess, C., Dell, G. S. & Eberhard, K. A. (2001) Dissociation between syntactic and semantic processing during idiom comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition 27: 1223-1237.

Pezzulo, G. (2011) Grounding procedural and declarative knowledge in sensorimotor anticipation. Mind and Language 26: 78-114.

Pickering, M.J. & Garrod, S. (2004) Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27: 169-225.

Pickering, M.J., & Garrod, S. (2007) Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences 11: 105-110.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In: B. MacWhinney (Ed.), The emergence of language (pp. 381–415). Erlbaum.

Postma, A. (2000) Detection of errors during speech production. A review of speech monitoring models. Cognition 77: 97–131.

Prinz, W. (2006) What renactment earns us. Cortex 42: 515-518.

Pulvermüller, F. & Fadiga, L. (2010) Active perception: Sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11: 351-360.

Pulvermüller, F., Huss, M., Kheri, F., Moscoso del Prado Martin, F., Hauk, O. & Shtyrov, Y. (2006) Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences 103: 7865-7870.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. Psychological Review, 107: 460-499.

Rapp, B., & Goldrick, M. (2004). Feedback by any other name is still interactivity: A reply to Roelofs (2004). Psychological Review, 111: 573-578.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R.L. & Schmidt, R.C. (2007) Rocking together: Dynamics of intentional and unintentional interpersonal coordination. Human Movement Science 26: 867-891.

Roelofs, A. (2004). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). Psychological Review, 111: 561-572

Sacks, H., Schegloff, E.A. & 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., & Halgren, E. (2009) Sequential processing of lexical, grammatical, and articulatory information within Broca's area. Science 326: 445-449.

Sams, M., Möttönen, R. & Sihvonen, T. (2005) Seeing and hearing others and oneself talk. Cognitive Brain Research 23: 429-435.

Sanford, A. J. & Garrod, S. C. (1981) Understanding written language. Wiley

Schlenck, K.-J., Huber, W. & Willmes, K. (1987) "Prepairs" and repairs: Different monitoring functions in aphasic language production. Brain and Language 30: 226-244.

Schober, M. F. & Clark, H. H. (1989) Understanding by addressees and over-hearers. Cognitive Psychology 21: 211-232.

Schriefers, H., Meyer, A.S. & Levelt, W.J.M. (1990) Exploring the time course of lexical access in language production: Picture-word interference studies. Journal of Memory and Language 29: 86-102.

Scott. S. & Johnsrude, I.S. (2003) The neuroanatomical and functional organisation of speech perception. Trends in Neurosciences 26: 100-107.

Scott, S., McGettigan, C. & Eisner, F. (2009) A little more conversation, a little less action – candidate roles for the motor cortex in speech perception. Nature Reviews Neuroscience 10: 295-302.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006a) Joint action: Bodies and minds moving together. Trends in Cognitive Sciences 10: 70-76.

Sebanz, N., Knoblich, G., Prinz ,W. & Wascher, E. (2006b) Twin peaks: An ERP study of action planning and control in coacting individuals. Journal of Cognitive Neuroscience 18: 859–870.

Sebanz, N., & Knoblich, G. (2009) Prediction in joint action: What, when, and where. Topics in Cognitive Science 1: 353–367.

Segaert, K., Menenti, L., Weber, K., Petersson, K.M., & Hagoort, P. (in press). Shared syntax in language production and language comprehension – an fMRI study. *Cerebral Cortex*. doi:10.1093/cercor/bhr249

Shockley, K., Santana, M.V. & Fowler, C.A. (2003) Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance 29: 326-332.

Stanley, J., Gowen, E. & Miall, R.C. (2007) Effects of agency on movement interference during observation of a moving dot stimulus. Journal of Experimental Psychology: Human Perception and Performance 33: 915-926.

Staub, A. & Clifton, C., Jr (2006) Syntactic prediction in language comprehension: Evidence from *either ...or*. Journal of Experimental Psychology: Learning, Memory, and Cognition 32: 425-436.

Stephens G.J., Silbert, L.J. & Hasson U. (2010) Speaker-listener neural coupling underlies successful communication. Proceedings of the National Academy of Sciences 107: 14425–14430.

Swinney, D. (1979) Lexical access during sentence comprehension: (Re) consideration of context effects. Journal of Verbal Learning and Verbal Behavior 18: 645-659

Tian, X. & Poeppel, D. (2010) Mental imagery of speech and movement implicates the dynamics of internal forward models. Frontiers in Psychology 1: 166.

Tourville, J.A., Reilly, K.J. & Guenther, F.K. (2008) Neural mechanisms underlying auditory feedback control of speech. Neuroimage 39: 1429-1443.

Tourville, J.A. & Guenther, F.K. (2011) The DIVA model : A neural theory of speech acquisition and production. Language and Cognitive Processes 26: 952-981.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994) Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language 33: 285-318.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C. & Rizzolatti, G. (2001) I know what you are doing: A neurophysiological study. Neuron, 32: 91-101.

Van Berkum, J.J.A., Brown, M.C., Zwitserlood, P., Kooijman, V. & Hagoort, P. (2005) Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition 31: 443–467.

Van den Bussche, E., Van den Noortgate, W. & Reynvoet, B. (2009) Mechanisms of masked priming: A meta-analysis. Psychological Bulletin 135: 452-477.

Van Wijk, C. & Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. Cognitive Psychology 19: 403-440.

Van Schie, H. T., van Waterschoot, B. M., & Bekkering, H. (2008) Understanding action beyond imitation: Reversed compatibility effects of action observation in imitation and joint action. Journal of Experimental Psychology: Human Perception and Performance 34: 1493–1500.

Vigliocco, G., Antonini, T. & Garrett, M. F. (1997) Grammatical gender is on the tip of Italian tongues. Psychological Science 8: 314-17

Vigliocco & Hartsuiker (2002) The interplay of meaning, sound, and syntax in sentence production. Psychological Bulletin 128: 442-472.

Vigneau, M., Beaucousin, V., Herve, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B. & Tzourio-Mazoyera, N. (2006) Meta-analyzing left hemisphere

language areas: Phonology, semantics, and sentence processing. NeuroImage 30: 1414 – 1432

Vissers, C. T., Chwilla, D. J. & Kolk, H. H. (2006) Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. Brain Research 1106: 150-163.

Watkins, K. & Paus, T. (2004) Modulation of motor excitability during speech perception: The role of Broca's area. Journal of Cognitive Neuroscience 16: 978-987.

Watkins, K., Strafella, A. P. & Paus, T. (2003) Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia, 41, 989-994. Psychology Press.

Weber, A., Grice, M., & Crocker, M. W. (2006) The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. Cognition 99: B63-B72.

Wheeldon, L. R., & Levelt, W. J. M. (1995) Monitoring the time course of phonological encoding. Journal of Memory and Language 34: 311–334

Wilson, M. & Wilson, T.P. (2005) An oscillator model of the timing of turn-taking. Psychonomic Bulletin and Review 12, 957-968.

Wilson, S.M., Saygin, A.P., Sereno, M.I. & Iacoboni, M. (2004) Listening to speech activates motor areas involved in speech production. Nature Neuroscience 7: 701-702.

Wijnen, F. & Kolk, H.H.J. (2005). Phonological encoding, monitoring, and language pathology: conclusions and prospects. In R.J. Hartsuiker, R. Bastiaanse, A. Postma & F. Wijnen (Eds.), Phonological encoding in normal and pathological speech (pp. 283-304). Psychology Press

Wohlschläger, A. (2000) Visual motion priming by invisible actions. Vision Research 40: 925-930.

Wolpert, D.M. (1997) Computational approaches to motor control. Trends in Cognitive Sciences 1: 209-216.

Wolpert, D.M., Ghahramani, Z. & Flanagan, J.R. (2001) Perspectives and problems in motor learning. Trends in Cognitive Sciences 5: 487-494.

Wolpert, D.M., Doya, K. & Kawato, M. (2003) A unifying computational framework for motor control and social interaction. Philosophical Transactions of the Royal Society of London B 358: 593-602.

Wright, B. & Garrett, M.F. (1984) Lexical decision in sentences: Effects of syntactic structure. Memory & Cognition 12: 31-45.

Yoshida, M., Dickey, M.W. & Sturt, P. (in press) Predictive processing of syntactic structure: Sluicing and ellipsis in real-time sentence processing. Language and Cognitive Processes.

Yuen, I., Davis, M.H., Brysbaert, M. & Rastle, K. (2010) Activation of articulatory information in speech perception. Proceedings of the National Academy of Sciences 107: 592-597.

Fischer, M.H. & Zwaan, R.A. (2008) Embodied language: A review of the role of the motor system in language comprehension. Quarterly Journal of Experimental Psychology 61: 825-850.

Endnotes

[1] The meat is "amodal", in the sense that its representations are couched in terms of abstract symbols rather than in terms of bodily movements (see General Discussion).

[2] Nothing hinges on this particular "traditional" set of levels. For example, it may be correct to distinguish logical form from semantics, or phonetics from phonology.

[3] Note that a mapping from semantics to phonology would be a production process, and a mapping from phonology to semantics would be a comprehension process. Some researchers argue that levels can be "skipped" in comprehension (e.g., F. Ferreira, 2003). But mappings between phonology and semantics also occur for other reasons, for example to express the relationship between emphasis (represented in the message level) and phonological stress, or between meaning and sound in sound symbolism.

[4] We assume that prediction is separate from action or perception – that the processes involved in predicting action or perception can at least in principle be distinguished from action or perception itself. In this respect our account differs from some theories such as Elman (1990).

5 Our *forward action model* corresponds to Wolpert's *forward dynamic model* and our *forward perception model* corresponds to his *forward output model*.

[6] The perceiver also has to accommodate to differences in perspective (e.g., when the actor is facing the perceiver). This type of accommodation is less relevant to (spoken) language, so we do not refer to it again.

[7] Mirror neurons fire during both action and perceiving an action (Di Pellegrino et al. 1992), and they are of course compatible with covert imitation during perception. Most evidence for mirror neurons is indirect in humans (e.g., activation of action

areas during perception), but Mukamel et al. (2010) used intercranial electrodes to demonstrate widespread mirror activity in Broca's area of an epileptic patient.

[8] We assume that speakers implement a level of semantics during production that is distinct from the production command. The production command includes a situation model that incorporates non-linguistic information whereas semantics is more akin to an "LF" level of representation (e.g., incorporating quantifier scope).

[9] In fact, Wijnen and Kolk (2005) briefly speculate about the possible use of forward and inverse models in monitoring, making reference to Wolpert's proposals.

[10] Note that Levelt (1989) assumed that there is appropriateness monitoring that takes place over semantic representations, and that there is no loop based on syntactic representations.

[11] The predicted utterance percept must be represented similarly to the utterance percept, in order that they can be compared. Thus we might expect speakers to have some awareness of the predicted utterance percept as well as the utterance percept. One possibility is that tip-of-the-tongue states constitute awareness of the forward model (in cases when the production implementer fails completely) rather than incompletely implemented production. For example, the speaker may compute the forward model for the first phoneme (e.g., Brown & McNeill, 1966) or grammatical gender (Vigliocco et al., 1997).

[12] Some evidence suggests that inner speech may be impoverished (Oppenheim & Dell, 2008, 2010; though cf. Corley, Brocklehurst, & Moat, 2011). An intriguing possibility is that such impoverishment reflects forward modeling rather than an abstract phonological representation constructed by the production implementer.

[13] Note that Kim and Lai interpret their results as involving interaction during early stages of lexical access, but this is not necessary.

[14] In fact, our account can explain why completions can be compatible with the perspective of either of the interlocutors. In (1), *B* said *But have you...* and *A* completed with *burned myself*?, *A*'s completion takes *A*'s perspective (myself). However *A* could have alternatively said *burned yourself*?, thus taking *B*'s perspective (see Section 1.1).