10 On the generation of shared symbols

Arjen Stolk, Mark Blokpoel, Iris van Rooij, & Ivan Toni

Abstract Despite the multiple semantic ambiguities present in every utterance during natural language use, people are remarkably efficient in establishing mutual understanding. This chapter illustrates how the study of human communication in novel settings provides a window into the mechanisms supporting the human competence to rapidly generate and understand novel shared symbols, capturing the joint construction of meaning across interacting agents. In this chapter, we discuss empirical findings and computational hypotheses generated in the context of an experimentally controlled non-verbal interactive task that throw light on these fundamental properties of human referential communication. The neural evidence reviewed here points to mechanisms shared across interlocutors of a communicative interaction. Those neural mechanisms implement predictions based on presumed knowledge and beliefs of the communicative partner. Computationally, the generation of novel meaningful symbolic representations might rely on cross-domain analogical mappings. Those mappings provide a mechanism for systematically augmenting individual pre-existing representations, adjusting them to the current conversational context.

The communicative use of language

Referential communication is a complex and anomalous instance of biological social interactions (Dawkins & Krebs, 1978; Owings & Morton, 1998). Referential communication is anomalous because it relies on context-dependent behaviors designed to influence the mental state of specific addressees, rather than on stable traits designed by natural selection to reliably influence bystanders (Danchin *et al.*, 2004). Referential communication is complex because each of its behavioral vehicles can carry multiple meanings, and a given meaning can be conveyed by a variety of behaviors. A great deal of effort has been spent in understanding features and rules of the system most frequently used by humans for achieving referential communication, i.e. language (Chomsky, 1995; de Saussure, 1910–1911; Jackendoff, 2002). Although those efforts have undoubtedly improved our understanding of the cognitive structures intrinsic to the language faculty (Hauser,

Willems, R. M. (Ed.). (2015). Cognitive neuroscience of natural language use. Cambridge University Press.

Chomsky, & Fitch, 2002), considerably less emphasis has been given to defining the cognitive processes that support the communicative use of language (Clark, 1996; Levinson, 2006; Schilbach *et al.*, 2013; Wittgenstein, 1953/2001). This chapter steps into this gap by focusing on our ability to share the meaning of a *novel* symbol, independently from the conventions and additional complexities introduced by linguistic processing (de Ruiter *et al.*, 2010; Galantucci & Garrod, 2011). Although it is often assumed that pre-existing symbols can be shared across interlocutors by simply coding and decoding them, using those symbols requires a computational mechanism powerful enough to mutually negotiate them across communicators (Levinson, 2006). Studying the generation of novel shared symbols provides a *privileged window* into this mechanism: given that novel symbols lack a pre-existing shared representation, jointly establishing their meaning relies on *converging on a common ground of knowledge and beliefs across communicators*, even more so than the meaning of already known words and gestures. As elaborated in the next section, existing accounts cannot explain the exceptional flexibility of human referential communication (when compared to other forms of animal communication), which may underlie our ability to share meanings and create language in the first place (Levinson, 2006). This chapter elaborates on the mechanisms supporting this human faculty, addressing the question of how communicators can design and interpret effective communicative acts. Starting from the premise that the generation of shared symbols depends on inferred knowledge and beliefs of a communicative partner, i.e. conceptual knowledge that accumulates and is adjusted in our minds as we interact, we reason that these mechanisms should be shared by the interlocutors of the communicative exchange and involve conceptual predictions based on a dynamic conversational context. We introduce an interactive experimental platform that induces the generation of shared symbols, and we then discuss empirical findings and computational hypotheses on properties of human referential communication that appear relevant for understanding natural language use.

Existing accounts of human communication

In the late 1940s, communication was formalized by Claude Shannon as an instance of signal transmission (Shannon, 1948). In Shannon's framework, agents can communicate as long as they have the same set of predefined coding–decoding rules. However, that framework does not explain how agents can negotiate those rules. Natural selection can drive organisms towards shared coding–decoding rules across multiple generations (Danchin *et al.*, 2004), but this account does not explain how humans can rapidly disambiguate situations lacking predefined

coding–decoding rules. This is not an exceptional situation. In fact, we achieve this feat during most daily conversations, when learning a language as infants, or when communicating with others in the absence of a common idiom (Levinson, 2006; Noordzij *et al.*, 2010). Even words used during natural dialogue do not contain fixed meanings – they may provide us with clues to a communicative meaning – but are coordinated through an interactive process by which people in dialogue seek and provide evidence that they understand one another (Brennan, Galati, & Kuhlen, 2010; Hofstadter & Sander, 2013). For instance, when a customer asks a bartender "Could you prepare a Margarita?", the bartender is not likely to pause wondering why the customer is questioning his skills, and the customer would not be puzzled by a logically unrelated answer like "Happy Hour starts in five minutes."

Studies on natural dialogue and recent reports in controlled experimental situations (de Ruiter *et al.*, 2010; Galantucci, 2005; Scott-Phillips, Kirby, & Ritchie, 2009) have shown that humans quickly develop new symbols when they need to, for instance novel shapes on a digitizing pad to communicate to another agent a location within a digital environment (Galantucci, 2005). However, it remains to be explained how those new symbols can be generated in the absence of an *a priori* common code. Computer simulations, using reinforcement-learning algorithms, have shown that communication systems can arise without the presence of common knowledge (Barr, 2004; Kirby & Hurford, 2002; Puglisi, Baronchelli, & Loreto, 2008; Steels, 2003). For instance, two computer agents can share novel symbols by virtue of guesses and explicit performance feedback (Steels, 2003). However, establishing these arbitrary signal–meaning mappings required many thousands of pair-wise interactions. Accordingly, general-purpose learning algorithms like temporal difference (Behrens, Hunt, & Rushworth, 2009) or Hebbian learning (Keysers & Perrett, 2004) do not seem suitable to explain the human ability to quickly grasp a meaning or to design an action that can be understood from scratch (de Ruiter *et al.*, 2010) since those learning algorithms require many trials to converge on statistically relevant features. Other scholars have suggested that human referential communication relies on cognitive modules that are involved only when communication requires it, e.g. when having to "repair" a misunderstanding (Horton & Keysar, 1996; Keysar & Horton, 1998) or when a certain representation is primed (automatically) by the utterance of an interlocutor (Garrod & Pickering, 2004). Further simplifications of this approach have led other scholars to suggest that actions can convey communicative meanings "without any cognitive mediation," by virtue of an automatic sensorimotor mechanism ("mirroring") that link the mental representation of an observed action to the representation of an executed action, and the latter to its outcome

(Rizzolatti & Craighero, 2004). However, those accounts leave unspecified how humans can effectively repair, prime, or "mirror" a communicative action when required, and they remain silent on how we organize our behavior for conveying intentions. Automatic priming, reinforcement learning, or sensorimotor associations might be instrumental in finessing a solution once a communicative action has been drafted, but they do not seem suitable to explain how we can rapidly converge on a shared understanding of a novel symbol. Those symbols, being novel, do not have well-defined priors (Fodor, 2000; Levinson, 2006; Sperber & Wilson, 2001) and dedicated neuronal circuits for unpacking their references (Giese & Poggio, 2003; Peelen, Fei-Fei, & Kastner, 2009).

Accordingly, the generation of shared symbols requires a mechanism that allows us to rapidly converge on a shared meaning, constraining a potentially infinite cognitive search space of mappings between symbols and their possible interpretations (or meanings). We suggest that this mechanism should be shared by interlocutors of the communicative exchange and, in order to alter an interlocutor's mental state in a predictable manner, should involve predictions based on presumed knowledge and beliefs of that specific interlocutor, conceptual knowledge that needs to be continuously updated and sharpened according to the shared history of the interaction (Brennan *et al.*, 2010; Clark, 1996). This account is closely linked to accounts of human social abilities based on the theory-of-mind framework (Frith & Frith, 2006; Premack & Woodruff, 1978). In this framework, the assumption is that behavior is the observable product of mental states, and making inferences about these mental states ("mentalizing": Frith & Frith, 2012) requires knowledge of their content and relationship with behavioral responses (Nichols & Stich, 2003). This concept-based account of our mentalizing abilities has been linked with cerebral structures that are distinct from the sensorimotor system, and include the superior temporal sulcus, the temporo-parietal junction, the temporal poles, and the medial prefrontal cortex (Amodio & Frith, 2006; Frith & Frith, 1999, 2006; Grezes, Frith, & Passingham, 2004; Walter *et al.*, 2004). Unfortunately, the majority of imaging studies investigating theory-of-mind have been conducted in non-interactive settings. In those settings, participants read or view story scenarios that trigger reasoning about the mental state of story characters (Saxe *et al.*, 2004). To date, the specific functions of the so-called "theory-of-mind network" remain unknown. Furthermore, the theory-of-mind framework is theoretically heterogeneous (Carruthers, 1996; Leslie, Friedman, & German, 2004; Nichols & Stich, 2003) and it remains to be seen whether and how theory-of-mind mechanisms play a role in genuine social interaction (Schilbach *et al.*, 2013).

Novel shared symbols as a privileged window into communicative interactions

The central tenet of this chapter is that the study of the generation of shared symbols provides a privileged view into the mechanisms of human communication, capturing the joint construction of meaning across interacting agents, contingent on the interaction dynamics. Unveiling the fundamental properties of human referential communication requires experimental procedures that capture these principles of human interaction, rather than the use of conventional linguistic representations. One way to address this issue is to generate experimental situations in which people need to communicate independently from the speech and gestures that are often used as behavioral vehicles for those mental representations. Novel symbols, like new words and gestures, are tokens that may represent and be used to convey ideas and beliefs while their meaning becomes shared between interlocutors. Studying how people generate shared novel symbols (technically known as "experimental semiotics": Galantucci & Garrod, 2011) therefore may provide a window into the mechanisms supporting the human competence to rapidly generate and understand communicative actions.

An experimental platform suitable for studying human communicative interaction requires that it is simple enough to be abstracted in computational models and neurophysiological experiments. Yet, it needs to be sufficiently flexible to capture non-trivial aspects of human communication. Several human communicative games have been developed and studied (Camerer, 2003; Feiler & Camerer, 2010; Galantucci, 2005; Scott-Phillips *et al.*, 2009; Selten & Warglien, 2007), with the Tacit Communication Game (de Ruiter *et al.*, 2010) being one of the few that has been studied from both a computational and neuroscientific perspective (Blokpoel *et al.*, 2011; Noordzij *et al.*, 2009; Stolk, Verhagen *et al.*, 2013). In this communication game, interlocutors do not have access to pre-existing conventions (e.g. a common language, body emblems, facial expressions) that may provide clues to the meaning of a symbol. The only available communicative vehicle consists of geometric shape movements, controlled by and visible to both players on a game board. This novel medium enforces the participant pairs to mutually negotiate novel symbols over the course of the task, effectively creating a new communication system. Consequently, the same symbol can be used by different communicative pairs to negotiate different meanings. The same symbol can even be used to convey different meanings by the same pair at different points in time, and vice versa (for examples see movies in Stolk, Verhagen *et al.*, 2013). These observations emphasize how, in this task, a symbol acquires

meaning, in part, by virtue of the history of the communicative interactions within a given pair.

The goal of the communication game is for pairs of participants – labeled as a "Communicator" and an "Addressee" throughout this chapter – to jointly re-create a spatial configuration of two geometric shapes shown only to the Communicator (see the thought cloud in Plate 10.1A, and event 2 in Plate 10.1B – in color plate section). This requires the Communicator to use the movements of his shape (in blue, event 3 in Plate 10.1B) to indicate to the Addressee how she should configure her shape (in orange). There are no *a priori* correct solutions to this communicative task, nor exists a limited set of options from which the Communicator can choose. The Addressee cannot solve the communicative task by reproducing the movements of the Communicator's shape. Rather, she needs to disambiguate communicative and instrumental components of the Communicator's movements, and find some relationship between the shape movements, i.e. the symbol, and their meaning. Success in this game thus relies on the Communicator designing a symbol that can be understood by the Addressee (for instance a "wiggle" to indicate a shape's orientation: Plate 10.1B), and on the Addressee inferring the Communicator's intentions. Participants turn out to be remarkably successful communicators under these constrained conditions (de Ruiter *et al.*, 2010). Given that they do not have access to pre-existing conventions, the participant pairs need to take into account the presumed beliefs and knowledge of their interlocutors when selecting and interpreting novel symbols as Communicators and Addressees respectively. Manipulation of the task structure shows that game performance (i.e. number of spatial configurations successfully re-created by the two players) improves when Communicators are able to see the Addressees' behaviors (event 5 in Plate 10.1B), suggesting that they take into account how Addressees interpreted their messages (de Ruiter *et al.*, 2010). This interpretation is reinforced by another study (Blokpoel *et al.*, 2012) showing that changes in the Communicators' movement characteristics after a misinterpretation of the Addressees are dependent on the nature of the error made by the Addressee. If an Addressee had placed her shape in an incorrect location, but with correct orientation, the Communicator tended to pause relatively longer on the Addressee's goal location. Such change in behavior is intended to indicate that a long pause should be interpreted as being dissociated from the rest of the movement, making it in effect less ambiguous for the Addressee which of the locations on the board was marked by the Communicator as the Addressee's goal location. This behavior cannot be explained by an appeal to a simple heuristic "if location in error then pause longer," because Communicators did not

pause longer when both location and orientation were in error. Rather, in those cases, Communicators understood the error was produced by a different type of misunderstanding on the part of the Addressee, leading Communicators to adjust their movement differently.

In sum, the communication game induces the generation of symbols that pertain to the inferred knowledge of the communicative partner. Within this task, communicative difficulty is easy to manipulate, using different combinations of shapes (for examples see Blokpoel *et al.*, 2012). Furthermore, it allows manipulating common ground knowledge across communicators, by having pairs encounter problems for which they previously had jointly established a solution. In this chapter, we discuss empirical findings and computational hypotheses in the context of this interactive task that throw light on the fundamental properties of human referential communication. We start by giving an overview of evidence from patient studies which show how neurological lesions may lead to alterations of communicative abilities.

Neurological alterations of communicative interactions

Clinical and experimental observations have clearly indicated that patients with severe damage to the language system can retain their communicative abilities (Goodwin, 2006; Willems *et al.*, 2011). In contrast, patients with right-hemisphere damage have been reported to have difficulties with conversational components of language (Sabbagh, 1999). The latter difficulties pertain to language use that requires an appreciation of non-literal speaker's intentions, as in the cases of sarcasm, indirect requests, metaphors, and humor interpretation (thus outside the domains of standard syntactical, phonological, or semantical levels of linguistic processing).

Another source of information on neural structures supporting our social interactive abilities comes from patients suffering frontotemporal dementia (FTD), a deterioration of the ventral base of the frontal lobe progressing towards the anterior temporal lobes (Snowden, Neary, & Mann, 2002). The analyses of the neurobiology of these patients reveal that intrinsically motivated social relationships are affected when the frontal lobe and right temporal pole degenerate (Fiske, 2010). The behavioral variant of FTD (bvFTD), also referred to as frontal variant FTD (fvFTD), is associated with a lack of insight, as when patients fail to recognize that anything is wrong with their behaviors (Avineri, 2010). The patient is able to reason about social rules even though the same patient has difficulty putting these rules into action (Mikesell, 2010). In a similar vein, patients can understand that others can have different beliefs

but perform badly when asked about the emotional state of another person (Mates, 2010). Social deficits may be seen during conversational exchanges where a patient is unable to keep track of events occurring during the interaction, and thus is unable to hold a coherent conversation. Semantic dementia, also referred to as temporal variant FTD (tvFTD), is associated with predominantly temporal lobe atrophy, typically greater in the left than in the right hemisphere (Weder *et al.*, 2007). Studies involving semantic dementia patients show that the anterior temporal lobes are important for accessing knowledge of coherent concepts (Lambon Ralph *et al.*, 2010). When these patients are shown a picture of a cat and are asked to point out other related items from a list, they point to photos of animals sharing superficial features with the cat, rather than conceptual similarities. For instance, the patients might include furry and long-tailed animals, and exclude tigers and lions. Taken together, these observations might suggest a degree of specialization between temporal and frontal contributions to human communication. The anterior temporal lobes might be particularly relevant for processing coherent concepts, whereas the frontal cortex might be involved in putting this conceptual knowledge into action as when social behaviors are guided by mental models of other agents. This suggestion fits with evidence obtained in patients with lesions in the ventromedial prefrontal cortex (vmPFC), a brain region consistently found more activated in functional imaging studies by tasks that rely on theory of mind than tasks that do not (Amodio & Frith, 2006). Evidence from lesion studies corroborates these findings, indicating that the vmPFC is a critical part of a network of neural structures important for taking into account the mental states of other people during decision-making (Bechara *et al.*, 1994; Kalbe *et al.*, 2010; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009; Shamay-Tsoory *et al.*, 2003; Stone, Baron-Cohen, & Knight, 1998). Similarly to the consequences of frontal lobe atrophy in FTD, brain injury in the vmPFC is associated with a constellation of symptoms which include impulsivity, perseveration, and compulsive behaviors (Damasio, 1994), and vmPFC patients also seem unaware of their socially inappropriate behavior. Yet, the same patients are able to recognize that their behavior is inappropriate when they view their own behavior on video (Beer *et al.*, 2006). A related finding comes from another study in which participants needed to press a left or right key for discriminating between male/female names and strong/weak words. Healthy participants typically become slower in the incongruent condition in which the same key is mapped to stereotypically incompatible stimuli, e.g. male names and weak words, or female names and strong words. Patients with vmPFC lesion do not show this response bias on this implicit task, but their performance is matched to controls when making explicit

judgments regarding gender-related stereotypical attributes, suggesting that their stereotypical knowledge is still intact (Milne & Grafman, 2001).

These and other studies have led to the suggestions that the vmPFC serves a unitary underlying function, namely to access and mediate model-based representations of the (social) environment to infer meaning, which might be used by other brain regions involved in processes related to decision-making (Euston, Gruber, & McNaughton, 2012; Jones *et al.*, 2012; Krueger, Barbey, & Grafman, 2009; Roy, Shohamy, & Wager, 2012; Schoenbaum *et al.*, 2009). Accordingly, damage to the vmPFC region interferes with patients' avoidance of unsavory others (e.g. using the scarf of a busker – requiring stereotype representations), but not of contaminating objects (Ciaramelli *et al.*, 2013). The moral judgments of patients with vmPFC lesions also seem particularly sensitive to a harmful outcome of a social interaction, rather than to the underlying intention of the agent (Ciaramelli, Braghittoni, & di Pellegrino, 2012). This observation also fits with the notion that the vmPFC is crucial for incorporating model-based representations of the agent into the decision process. Accordingly, it might be expected that the vmPFC is crucially involved in adjusting behaviors to a mental model of an interlocutor during interaction. Using a version of the communication game outlined in the previous section, we tested whether vmPFC patients spontaneously adjust their behavior according to their beliefs of an Addressee's cognitive abilities. Healthy and lesion control participants spent longer on communicatively relevant locations of the game board when they believed to be interacting with a child, as compared to an adult Addressee. The vmPFC patients were able to communicate as effectively as the control groups, but they did not adjust their communicative behavior to the characteristics of the presumed Addressee (A. Stolk, D. D'Imperio, G. di Pellegrino, & I. Toni, unpublished data). Furthermore, although patients clearly detected communicative errors and adjusted to those errors by moving slower in the subsequent trial, they did not adjust their communicative behavior to the cause of the error. For instance, they failed to make the communicatively relevant location more discriminable from other visited locations of the game board to the Addressee. These findings suggest that patients are still able to produce communicative actions, but they are not able to take into account the inferred knowledge and beliefs of their interlocutor when doing so. In contrast, testing verbal communication in patients with vmPFC lesions did not reveal differences from healthy controls. Namely, there were similar reductions in time and words used for verbal referential descriptions during a collaborative referencing task (Gupta, Tranel, & Duff, 2012). Taken together, these findings suggest that the vmPFC is

necessary for using a mental model of an interlocutor. In contrast, interactions based on verbal material might bypass those models and rely on purely linguistic phenomena, e.g. an increased accessibility of syntactic and semantic nets sharpened by their recent use in the communicative interactions (Sass *et al.*, 2009; Segaert *et al.*, 2012).

Neural mechanisms of communicative interactions

Evidence from patient studies indicates an important role for the frontal and temporal lobes in supporting our communicative abilities. Functional imaging studies probing human theory-of-mind abilities corroborate these findings, showing consistent involvement of brain regions in the frontal and temporal lobes, including the superior temporal sulcus, the temporo-parietal junction, the temporal poles, and the medial prefrontal cortex (Amodio & Frith, 2006; Frith & Frith, 1999, 2006; Grezes *et al.*, 2004; Saxe *et al.*, 2004; Vogeley *et al.*, 2001; Walter *et al.*, 2004). As mentioned earlier, the majority of imaging studies that aimed to probe this theory-of-mind network has been conducted in non-interactive settings in which participants read or view story scenarios that trigger the participant to reason about the story character's belief (e.g. Saxe *et al.*, 2004). To date, the exact roles of the distinct brain regions involved in the theory-of-mind network during genuine communicative interaction remain unknown. Similar to sensorimotor simulation theory (Rizzolatti & Craighero, 2004), the theory-of-mind framework remains silent on how we organize our behavior for conveying intentions. For instance, it is still unknown whether, and if so how, these brain regions even support the human ability to generate novel shared symbols, a fundamental property of human referential communication. In this section, we discuss recent findings that may throw light on these issues.

To investigate whether the mechanisms supporting the human ability to share novel symbols are involved both when generating and understanding novel symbols, Noordzij and colleagues (Noordzij *et al.*, 2009) recorded brain activity with fMRI from one subject of each pair interacting within the communication game. As indicated in a previous section, in this interactive task pairs have to solve communicative problems involving the joint re-creation of a spatial configuration of two geometric shapes, shown to one of the players only. In a first experiment, participants either generated novel symbols to convey to Addressees where and how to position their shapes (as Communicators, event 2 in Plate 10.1B), or they generated identical symbols but with no communicative necessity (non-communicative control). Namely, in this experimental condition it was made explicit to them that their Addressees also saw the spatial goal

configuration so there was no need for them to consider their interlocutors when generating those symbols. In a second experiment, participants observed novel symbols generated by their Communicators to infer their meanings (as Addressees, event 3 in Plate 10.1B), or to keep track of the location where the Communicators last moved their shape twice (non-communicative control). Contrasting neural activation in each communicative condition with that evoked during their respective controls, they found that generating (by Communicators) and understanding novel symbols (by Addressees) relied on spatially overlapping portions of their brains (the right posterior superior temporal sulcus – pSTS). Furthermore, the hemodynamic response of this region was strongly modulated by the ambiguity in meaning of the communicative acts, but not by the sensorimotor complexity of those acts. This finding does not fit with the suggestion that our communicative abilities are supported by automatic sensorimotor resonances between a sender of a message and its receiver (Keysers & Perrett, 2004; Rizzolatti & Craighero, 2004). Instead, this study provides a first indication for a computational overlap between generating and understanding novel shared symbols, involving processes that fall outside the sensorimotor and linguistic domain.

To implement a more stringent and informative test of the computational overlap hypothesis, we used magnetoencephalography (MEG), a technique that allows one to characterize temporal and spectral dimensions of neural activity, besides the spatial distribution of that activity. Studying the neural dynamics of the predicted overlap between generating and understanding novel shared symbols also gives rise to the possibility of exploring whether these processes rely on a cognitive set implemented through tonic neural activity, or on phasic processes related to low-level features of the stimulus material. We therefore had pairs of participants engage in live communicative interactions. Communicative difficulty increased over the course of the experiment, in order to have the pairs continuously (re)negotiate meanings of symbols, a core element of daily dialogue. Neural activity was measured from one participant of each communicative pair, alternating between the role of Communicator and Addressee on a trial-by-trial basis, and distinguished from activity evoked during another interactive game that involved the same stimuli, responses, attention, and between-participant dependencies but no communicative necessities (Stolk, Verhagen *et al.*, 2013). Namely, in this non-communicative control interaction, the same participants solved their individual problems following learned rules. During communicative interactions, two brain regions exhibited significantly stronger signal power, most pronounced around 55–85 Hz (gamma band). This effect emerged from a broadband spectral change in neural activity over vmPFC

and right temporal cortex, and it was present when participants were generating as well as understanding novel shared symbols (Plate 10.2 in color plate section).

Further characterization of the overlap, using an absolute index of neural activity (i.e. source-reconstructed time-resolved estimates of gamma-band activity: Gross *et al.*, 2001), revealed three important features of the underlying neural dynamics. First, sharing the meaning of novel symbols relies on processes with a surprisingly matched phasic temporal dynamics as during non-communicative control interactions (cf. color and gray traces in Plate 10.2). This finding argues against computational modules that are exclusively and sufficiently dedicated to social cognition (Adolphs, 2009). Second, the tonic upregulation of neural activity across Communicator and Addressee was present well before the occurrence of a specific communicative problem (during baseline epochs), with baseline neural activity in the right temporal lobe (TL in Plate 10.2) predicting task performance in an upcoming event (see Figure 4 of Stolk, Verhagen *et al.*, 2013). This finding supports the notion that crucial cognitive elements of human communication are not stimulus-locked. Rather, conceptual knowledge abstracted from the history of communicative interactions needs to be continuously aligned to the current conversational context (Clark, 1996). Third, there were distinct temporal profiles of neural activity in those regions with overlapping increases in gamma-band activity during generation and comprehension of novel shared symbols. A ventrolateral portion of the right temporal lobe (TL, Plate 10.2) showed a tonic upregulation of neural activity, but without clear transient responses time-locked to the sensorimotor events occurring during those epochs. The ventromedial prefrontal cortex (vmPFC, Plate 10.2) showed decreases in neural activity when participants observed actions in both the communicative and non-communicative tasks and increases when participants started planning their actions. This pattern fits with the recent observation that this region is crucial for guiding our (communicative) decisions with inferred knowledge and beliefs of a communicative partner (A. Stolk, D. D'Imperio, G. di Pellegrino, & I. Toni, unpublished data). The right posterior superior temporal sulcus (pSTS, Plate 10.2) is sensitive to computational demands that occur early in planning and that rise during action observation, i.e. with presentation of new stimulus material.

Previous work has highlighted the right pSTS as an important element of the cerebral system supporting human referential communication, both for Communicators generating novel symbols as well as for Addressees trying to understand those symbols (de Langavant *et al.*, 2011; Gao, Scholl, & McCarthy, 2012; Mashal *et al.*, 2007; Noordzij *et al.*, 2009,

2010). However, the exact contributions of this region, let alone the necessity, to human communicative interaction remain unknown. Namely, the involvement of the right pSTS in establishing shared symbols is one among several contributions associated with this region, including the perception of biological motion and goal-directed actions, moral judgments, and mental state attribution (Arfeller *et al.*, 2013; Bahnemann *et al.*, 2010; Grossman, Battelli, & Pascual-Leone, 2005; Schultz *et al.*, 2005; Shultz *et al.*, 2011). Presumably, this heterogeneity might reflect superficial differences of an underlying unitary function. The neural dynamics indicate that right pSTS is also upregulated as a function of the cognitive set (already before the occurrence of stimulus material), but with clear transient responses to incoming visual information (see Plate 10.2), suggesting that this region might be involved in the integration of stimulus material with priors (Jakobs *et al.*, 2012). Accordingly, we reasoned that these priors could capture (1) statistical regularities of the sensory stimuli experienced by the participants (Iacoboni, 2005; Schippers *et al.*, 2010; Tognoli *et al.*, 2007; Turesson & Ghazanfar, 2011); (2) conceptual predictions based on semantic conventions (Schultz *et al.*, 2005; Wyk *et al.*, 2009; Young *et al.*, 2010); or (3) conceptual predictions based on a dynamic conversational context shared among communicators (Menenti, Pickering, & Garrod, 2012). To test these hypotheses, we used low-frequency repetitive transcranial magnetic stimulation (rTMS) to perturb functioning of this region while participants observed novel symbols generated by their communicators to infer their meanings (Boroojerdi *et al.*, 2000; Mottaghy *et al.*, 2002). We found that general task performance was not affected by rTMS, whereas task-learning was disrupted according to TMS site and task combinations. Namely, rTMS over pSTS led to a diminished ability to improve understanding of those novel symbols on the basis of the recent communicative history, while rTMS over MT+, a contiguous homotopic control region involved in integrating position information when viewing moving objects, perturbed improvement over trials in visual tracking (non-communicative control) of exactly the same time series of stimuli used in the communicative setting (Stolk, Noordzij, Volman *et al.*, 2013). This finding increases our understanding of the neural mechanisms of human communication by showing that the right pSTS, in contrast to MT+, is necessary for continuously adjusting conceptual predictions (hypothesis #3) according to the recent history of interactions of the communicators, over and above the statistical regularities of the sensory stimuli experienced by the participants (that are also present in the control task). The task-, region-, and learning-specific effect observed in this study suggests that human communicative abilities operate on

conceptual inferences, rather than sensorimotor brain-to-brain couplings (Hasson *et al.*, 2012), and that those conceptual inferences are continuously updated. It remains to be seen whether the right pSTS supports the dynamic updating of communicative inferences also when communication relies on linguistic material with strongly established semantic conventions (Mitchell *et al.*, 2009; van Ackeren *et al.*, 2012; Willems *et al.*, 2010).

The neural evidence reviewed thus far points to mechanisms that are shared by interlocutors of the communicative exchange, and that involve flexible conceptual priors based on the shared history of interactions of communicators rather than statistical regularities in the stimulus material. However, it remains unclear how this conceptual knowledge is shared among communicators in the first place. The neuronal computations supporting this ability might be synchronized across interlocutors by the symbols used during a communicative interaction (Hari *et al.*, 2013; Hasson *et al.*, 2012; Rizzolatti & Craighero, 2004). Alternatively, the conversational meta-knowledge shared across a pair of interlocutors might be neuronally implemented over temporal scales independent from individual communicative events (Stolk, Verhagen *et al.*, 2013). We addressed this issue in another study, using fMRI to simultaneously record brain activation in pairs of participants building a pair-specific conversational context across multiple communicative interactions (Stolk, Noordzij, Verhagen *et al.*, 2013). During these interactions, participants solved communicative problems for which the pairs already had established common ground and communicative problems in which common ground yet had to be established. We observed that as common ground emerged within a pair of interlocutors, activity in the right superior temporal gyrus (STG) also increased, during both production and comprehension of a communicative action. To investigate whether the emergence of common ground, neuroanatomically supported by right STG, was specific to the context and participants of the interaction, we applied a methodology originally refined in electrophysiology, spectral coherence analysis, to the fMRI time series and contrasted the joint neural dynamics evoked within pairs with those evoked in participants from different pairs. This analysis showed a significantly stronger within-than between-pair coherence at frequencies lower than the dominant experimental frequency and with zero phase-lag, indicating a temporal synchronization of blood-oxygen-level dependent (BOLD) changes in right STG that was specific for elements of a communicative pair, and over a timescale spanning several communicative interactions (25–100 seconds, whereas one interaction lasted ~20 seconds). These findings indicate that sharing conceptual knowledge among communicators relies on conceptual

operations shared across participants of a communicative pair, and superseding individual communicative symbols.

In sum, the work outlined thus far shows that participants can successfully share novel symbols in the absence of an *a priori* common code (e.g. a common language). This work also shows that the generation of shared symbols upregulates the same neuronal mechanism in the same brain regions across pairs of communicators, and over temporal scales independent from transient sensorimotor events. This finding indicates that the communicative meaning of a symbol arises from the (pair-specific) conversational context rather than from the stimulus material itself. Mechanistically, the meaning of novel shared symbols might be rapidly inferred by embedding those symbols in a conceptual space whose activation predates in time the processing of the symbols themselves (van Berkum *et al.*, 2008). This conceptual knowledge thus needs to be continuously aligned to the conversational context (Clark, 1996). Taken together, the current neural evidence suggests that jointly establishing meaning, a feature crucial for human communication, relies on *knowledge and beliefs knowingly shared and updated* between the communicators during the course of their interactions. Currently, it is still an open question how novel meaning is found, and mapped onto a symbol. In the next section, we elaborate on computational processes that might support our abilities to engage in communicative interaction and to share novel symbols.

Computational features of communicative interactions

The overall ability to generate and interpret novel communicative acts arguably relies on a large number of cognitive functions, ranging from object recognition to theory-of-mind. Thinking of the cognitive operations germane to the generation of novel shared symbols, three functions seem to stand out: *parsing*, *perspective-taking*, and *meaning-mapping*. To grasp a meaning, an Addressee minimally needs to parse a signal into communicative and instrumental parts and then infer the meaning of the communicative parts. To convey a meaning, a Communicator needs to reason about how the Addressee will parse and interpret a signal, through some form of perspective-taking (Blokpoel *et al.*, 2012; van Rooij *et al.*, 2011). These functions, parsing and perspective-taking, are arguably also involved in communicative interactions that rely on pre-existing conventions. The same holds for meaning-mapping, but this process seems to be more heavily taxed during communicative exchanges in which novel symbols have to be created and understood. In this section we will focus on this meaning-mapping process.

A characterizing feature of meaning-mapping, both in everyday conversation as well as in the communication game used in our neuroscientific research, is that addressees can often infer the intended meaning of a new symbol on first encounter or otherwise within a few trials (de Ruiter *et al.*, 2010; Volman, Noordzij, & Toni, 2012). This phenomenon cannot be easily accommodated by traditional reinforcement learning theories (Kaelbling, Littman, & Moore, 1996), game theories (Osborne, 2004), fast and frugal heuristics (Gigerenzer, 2008), or Bayesian models (Tenenbaum, Griffiths, & Kemp, 2006). Those models require either *a priori* internal models of all possible novel signals (i.e. the models have meanings of symbols built in as conventions), or an unrealistically large number of training trials (Steels, 2003). An alternative account that does not seem to suffer from these problems can be found in structure mapping theory, or analogical reasoning (Gentner, 1983; Gentner, 2003). In analogical reasoning, one uses representations of the relational structure of concepts to find *analogical matches* between different concepts (e.g. "the atom is like a solar system, with electrons circling the nucleus in much the same way as planets circle the sun"). Based on such matches, one can then *transfer knowledge* from a base concept to a target, generating new concepts (e.g. "perhaps the revolving of electrons is caused by the attraction of the nucleus like the revolving of the planets is caused by attraction of the sun"). This kind of reasoning seems to meet the computational requirements for fast, even one-trial, learning. Using a case study from the communication game, we will illustrate how analogical reasoning can, in principle, explain how Communicators can generate novel symbols whose meaning can be correctly inferred by Addressees on first encounter.

To be able to explain how movements made by Communicators in the communication game can take on novel meanings that Addressees can understand quickly, we will make the plausible assumption that players share considerable amounts of general world knowledge, i.e. everyday knowledge that the players have both acquired outside the context of the communication game. For instance, in our case study we will assume that players have basic geometric knowledge (e.g. concepts of "circle," "triangle," "frame of reference," "line," "point," etc.) as well as basic concepts of motion such as "direction," "speed," etc. To use this knowledge to infer the meaning of the movement depicted in Plate 10.3 (see color plate section) (bottom; what we refer to as a "wiggle"), multiple inferential steps are necessary. Each step in this inferential process gradually builds more sophisticated, abstract (and potentially novel) representations of the observed movements such that at some point the meaning becomes evident. The inferential steps involve what we call *analogical augmentations*: by finding analogies between specific observations and

non-game-specific knowledge, one can augment the raw observations into more and more abstract representations of location and orientation. For instance, we can represent the transition between two consecutive positions of the circle as "like (drawing) a line" (Plate 10.3, top). This analogy may seem quite trivial. In fact, it involves representing the positions of the circle and the relations between these positions, e.g. the circle is now "right_of" its previous position. Furthermore, this analogy requires the knowledge that a line is a relation between two positions that have a spatial relationship (e.g. "right_of"). Only then can an analogical match be found, viz. between the positions and relationships, to transfer the relation "line" onto the observed circle positions. Those abstract representations of location and orientation can eventually analogically match to the representation of the triangle, allowing the Addressee to infer its location and orientation. Note that, given different representations and knowledge, widely different meaning mappings become possible. This is consistent with the diversity of strategies observed in players of the communication game (Blokpoel *et al.*, 2012; de Ruiter *et al.*, 2010).

The computational account of meaning-mapping that we roughly outlined above and in Plate 10.3 illustrates that there are computationally sufficient mechanisms for generating and understanding novel symbols. Unlike most standard learning models, these mechanisms involve various forms of analogical reasoning. That is, to generate novel meaningful symbolic representations one needs to be able to systematically augment one's representations such that these representations support cross-domain analogical mappings. In the case of the communication game we illustrated that this involves analogical mappings between the domains of geometry and motion. Given that for any given pair of representations there may exist augmentation paths that lead to analogical matches, the model outlined here may best be seen as a meaning *hypothesizer*. That is, it defines the set of candidate meanings for a given signal, without specifying how people select the most plausible or probable meaning from the set of candidate meanings. Combining an analogy-based model with rational, probabilistic, or coherentist models might offer a more complete picture (Thagard, 1989; van Rooij *et al.*, 2011).

Conclusions

This chapter elaborates on neurobiological and computational mechanisms supporting the generation of novel shared symbols. Functional imaging data, supported by the observation of consequences of brain injury and transient interference with brain function, highlight a fundamental role for right temporal and ventromedial prefrontal brain regions in the

coordination among interlocutors during referential communication. Empirical evidence obtained in an interactive communicative setting shows that generating and comprehending novel shared symbols upregulates the same neuronal mechanisms in cortical regions known to be crucial for processing conceptual knowledge, across pairs of communicators, and over temporal scales independent from transient sensorimotor events (Stolk, Verhagen *et al.*, 2013). In fact, the neural dynamics observed in the right superior temporal gyrus suggest that those conceptual operations may span multiple communicative exchanges, temporally synchronized within a communicating pair, and modulated when novel knowledge is generated among the interlocutors (Stolk, Noordzij, Verhagen *et al.*, 2013). We suggest that the right posterior superior temporal sulcus supports our ability to benefit from recent communicative experiences with a communicative partner (Stolk, Noordzij, Volman *et al.*, 2013). The ventromedial prefrontal cortex seems crucial for taking into account inferred knowledge and beliefs of the interlocutor when choosing from a set of possible communicative options (A. Stolk, D. D'Imperio, G. di Pellegrino, & I. Toni, unpublished data). Taken together, the empirical findings and computational considerations suggest that the meaning of a novel symbol arises from a conceptual space dynamically defined by the ongoing interaction, rather than from the stimulus material itself. Plate 10.4 (see color plate section) summarizes these considerations and the main issues addressed in this review.

This review raises a number of outstanding issues that deserve further investigations. The ability to quickly converge on a common ground of knowledge and beliefs across communicators, efficiently building new and reconfiguring existing semiotic conventions, emerges at different levels of human communication, from infants learning a language without access to the local communicative conventions, to adults with purportedly limited communicative means (shape movements on a game board) as in the studies outlined above. The present work indicates that the meaning of novel shared symbols might be rapidly inferred by embedding those symbols in a conceptual space whose activation predates in time the processing of the symbols themselves (van Berkum *et al.*, 2008). Even during a simple conversation, we continuously update and sharpen our (conceptual) priors according to the recent history of the communicative interaction. We present a draft of a computational model that taps directly into the mystery of how the human mind constrains the inferential process that leads to action selection and understanding within communicative interaction. Future studies might shed light on the mechanisms of how representations are constructed from, and integrated with, incoming stimulus material.

Currently, there is a debate as to whether our theory-of-mind abilities can be subdivided in a cognitive component, supporting our abilities to take into account knowledge and beliefs of another agent, and an affective component, supporting our abilities to take into account the feelings of another agent (Gupta *et al.*, 2012; Shamay-Tsoory *et al.*, 2009). Recent investigations from our lab seem to initially support such a dissociation, as when measures of fluid intelligence and systemizing abilities, but not empathy and reward-related tendencies, have been shown to account for significant portions of inter-subject variability in the ability to quickly grasp novel communicative meanings according to recent communicative interactions (Stolk, Noordzij, Volman *et al.*, 2013; Volman *et al.*, 2012). In contrast, empathy scores appear to be more closely related to audience design abilities (Newman-Norlund *et al.*, 2009). Taken together, we suggest that while pro-social attitudes (approximately indexed by empathy) might provide the motivational drive necessary for adjusting communicative behavior to a given agent (Tomasello, 2008), other general-purpose cognitive abilities (approximately indexed by fluid intelligence) might provide the computational tools necessary to cope with the complexity of human referential communication (van Rooij *et al.*, 2011). Studying human development might provide a relevant handle for understanding how those motivational drives and cognitive abilities are implemented and coordinated (Stolk, Hunnius *et al.*, 2013). In a first attempt to address these issues, we have investigated children's ability to influence the mental state of others, and whether these abilities are influenced by the extent and nature of children's social interactions (Carpendale & Lewis, 2004; de Rosnay & Hughes, 2006; Dunn & Shatz, 1989; Hrdy, 2009; Lewis *et al.*, 1996; Perner, Ruffman, & Leekam, 1994). The rationale and focus of this study is quite different from a large body of existing developmental work that has focused on our ability to attribute mental states to others (Baron-Cohen, Leslie, & Frith, 1985; Wellman, Cross, & Watson, 2001). In a nutshell, our work suggested that referential communicative abilities might be bootstrapped within social interaction itself: 5-year-olds' internally generated communicative adjustments to their mental model of an addressee were shaped by their early social experience with other cognitive agents (Stolk, Hunnius *et al.*, 2013). Those findings open the way for systematic and sensitive investigations into the contribution of early social experiences towards children's communicative abilities, raising the possibility to chart the developmental trajectories generated by different sources of social interaction through longitudinal studies with objective measures of the time spent on those interactions. It is known that, in adults, social network size has a positive impact on neural circuits deemed relevant for social cognition, e.g. vmPFC, pSTS, anterior

cingulate cortex, and amygdala (Bickart *et al.*, 2011; Kanai *et al.*, 2012; Lewis *et al.*, 2011; Sallet *et al.*, 2011). Accordingly, it appears relevant to explore how brain development is influenced by early social experiences that have an impact on our communicative abilities, and whether such effects are long-lasting.

Finally, it should be emphasized that this review has largely focused on empirical observations obtained in the context of a highly controlled experimental setup, designed to capture one crucial element of communicative interaction, namely sharing meanings of novel symbols extended over several seconds. It remains open for discussion whether this approach is adequate for understanding the theoretical components and the cerebral mechanisms supporting human communication in more naturalistic settings. Certainly, it will be important to test how the present findings generalize to other communicative materials (e.g. linguistic and/or gestural), and to interactive situations where communicative roles can be frequently exchanged, as during natural dialogue.

References

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annu Rev Psychol*, *60*, 693–716.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci*, *7*(4), 268–277.

Arfeller, C., Schwarzbach, J., Ubaldi, S., Ferrari, P., Barchiesi, G., & Cattaneo, L. (2013). Whole-brain haemodynamic after-effects of 1-Hz magnetic stimulation of the posterior superior temporal cortex during action observation. *Brain Topogr*, *26*(2), 278–291.

Avineri, N. (2010). The interactive organization of "insight": clinical interviews with frontotemporal dementia patients. In A. W. Mates, L. Mikesell, & M. S. Smith (eds.), *Language, Interaction and Frontotemporal Dementia: Reverse Engineering the Social Mind* (pp. 115–138). London: Equinox.

Bahnemann, M., Dziobek, I., Prehn, K., Wolf, I., & Heekeren, H. R. (2010). Sociotopy in the temporoparietal cortex: common versus distinct processes. *Soc Cogni Affect Neurosci*, *5*(1), 48–58.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Barr, D. J. (2004). Establishing conventional communication systems: is common knowledge necessary? *Cogn Sci*, *28*(6), 937–962.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15.

Beer, J. S., John, O. P., Scabini, D., & Knight, R. T. (2006). Orbitofrontal cortex and social behavior: integrating self-monitoring and emotion–cognition interactions. *J Cogn Neurosci*, *18*(6), 871–879.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *Science*, *324*(5931), 1160–1164.

Bickart, K. C., Wright, C. I., Dautoff, R. J., Dickerson, B. C., & Barrett, L. F. (2011). Amygdala volume and social network size in humans. *Nat Neurosci*, *14*(2), 163–164.

Blokpoel, M., Kwisthout, J., Wareham, T., Haselager, P., Toni, I., & Van Rooij, I. (2011). The computational costs of recipient design and intention recognition in communication. Paper presented at the Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Austin, TX.

Blokpoel, M., van Kesteren, M., Stolk, A., Haselager, P., Toni, I., & van Rooij, I. (2012). Recipient design in human communication: simple heuristics or perspective taking? *Frontiers Hum Neurosci*, *6*.

Boroojerdi, B., Prager, A., Muellbacher, W., & Cohen, L. G. (2000). Reduction of human visual cortex excitability using 1-Hz transcranial magnetic stimulation. *Neurology*, *54*(7), A400.

Brennan, S. E., Galati, A., & Kuhlen, A. (2010). Two minds, one dialog: coordinating speaking and understanding. In B. Ross (ed.), *Psychology of Learning and Motivation* (Vol. 53, pp. 301–344). Burlington, MA: Academic Press.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends Cogn Sci*, *7*(5), 225–231.

Carpendale, J. I., & Lewis, C. (2004). Constructing an understanding of mind: the development of children's social understanding within social interaction. *Behav Brain Sci*, *27*(1), 79–96; discussion 96–151.

Carruthers, P. (1996). Simulation and self-knowledge: a defence of theory-theory. In P. Carruthers & P. K. Smith (eds.), *Theories of Theories of Mind* (pp. 22–38). Cambridge: Cambridge University Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Ciaramelli, E., Braghittoni, D., & di Pellegrino, G. (2012). It is the outcome that counts! Damage to the ventromedial prefrontal cortex disrupts the integration of outcome and belief information for moral judgment. *J Int Neuropsychol Soc*, *18*(6), 962–971.

Ciaramelli, E., Sperotto, R. G., Mattioli, F., & di Pellegrino, G. (2013). Damage to the ventromedial prefrontal cortex reduces interpersonal disgust. *Soc Cogn Affect Neurosci*, *8*(2), 171–180.

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

Damasio, A. R. (1994). *Descartes' Error*. New York: Putnam.

Danchin, E., Giraldeau, L. A., Valone, T. J., & Wagner, R. H. (2004). Public information: from nosy neighbors to cultural evolution. *Science*, *305*(5683), 487–491.

Dawkins, R., & Krebs, J. (1978). Animals signals: information or manipulation. In J. R. Krebs & N. B. Davies (eds.), *Behavioural Ecology: An Evolutionary Approach* (pp. 282–309). Oxford: Blackwell.

de Langavant, L. C., Remy, P., Trinkler, I., McIntyre, J., Dupoux, E., Berthoz, A., & Bachoud-Levi, A. C. (2011). Behavioral and neural correlates of communication via pointing. *PLoS ONE*, *6*(3).

de Rosnay, M., & Hughes, C. (2006). Conversation and theory of mind: do children talk their way to socio-cognitive understanding? *Br J Devel Psychol*, *24*, 7–37.

de Ruiter, J. P., Noordzij, M. L., Newman-Norlund, S., Newman-Norlund, R., Hagoort, P., Levinson, S. C., & Toni, I. (2010). Exploring the cognitive infrastructure of communication. *Interaction Stud*, *11*(1), 51–77.

de Saussure, F. (1910–1911). *Cours de linguistique générale* [Course in General Linguistics]. Paris: Payot.

Dunn, J., & Shatz, M. (1989). Becoming a conversationalist despite (or because of) having an older ibling. *Child Devel*, *60*(2), 399–410.

Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron*, *76*(6), 1057–1070.

Feiler, L., & Camerer, C. F. (2010). Code creation in endogenous merger experiments. *Econ Inquiry*, *48*(2), 337–352.

Fiske, A. P. (2010). Dispassionate heuristic rationality fails to sustain social relationships. In A. W. Mates, L. Mikesell, & M. S. Smith (eds.), *Language, Interaction, and Frontotemporal Dementia: Reverse Engineering the Social Mind* (pp. 199–242). London: Equinox.

Fodor, J. A. (2000). *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Frith, C. D., & Frith, U. (1999). Interacting minds: a biological basis. *Science*, *286*(5445), 1692–1695.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annu Rev Psychol*, *63*, 287–313.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cogn Sci*, *29*(5), 737–767.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers Hum Neurosci*, *5*, 11.

Gao, T., Scholl, B. J., & McCarthy, G. (2012). Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. *J Neurosci*, *32*(41), 14 276–14 280.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends Cogn Sci*, *8*(1), 8–11.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cogn Sci*, *7*(2), 155–170.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (eds.), *Language in Mind: Advances in the Study of Language and Thought* (pp. 195–235). Cambridge, MA: MIT Press.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci*, *4*(3), 179–192.

Gigerenzer, G. (2008). Why heuristics work. *Perspect Psychol Sci*, *3*(1), 20–29.

Goodwin, C. (2006). Human sociality as mutual orientation in a rich interactive environment: multimodal utterances and pointing in aphasia. In N. J. Enfield & S. C. Levinson (eds.), *Roots of Human Sociality* (pp. 97–125). New York: Berg.

Grezes, J., Frith, C. D., & Passingham, R. E. (2004). Inferring false beliefs from the actions of oneself and others: an fMRI study. *NeuroImage*, *21*(2), 744–750.

Gross, J., Kujala, J., Hamalainen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proc Natl Acad Sci USA*, *98*(2), 694–699.

Grossman, E. D., Battelli, L., & Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision Res*, *45*(22), 2847–2853.

Gupta, R., Tranel, D., & Duff, M. C. (2012). Ventromedial prefrontal cortex damage does not impair the development and use of common ground in social interaction: implications for cognitive theory of mind. *Neuropsychologia*, *50*(1), 145–152.

Hari, R., Himberg, T., Nummenmaa, L., Hamalainen, M., & Parkkonen, L. (2013). Synchrony of brains and bodies during implicit interpersonal interaction. *Trends Cogn Sci*, *17*(3), 105–106.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends Cogn Sci*, *16*(2), 114–121.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Hofstadter, D., & Sander, E. (2013). *Surfaces and Essences*: *Analogy as the Fuel and Fire of Thinking*. New York: Basic Books.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117.

Hrdy, S. B. (2009). *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Cambridge, MA: Belknap Press of Harvard University Press.

Iacoboni, M. (2005). Neural mechanisms of imitation. *Curr Opin Neurobiol*, *15*(6), 632–637.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jakobs, O., Langner, R., Caspers, S., Roski, C., Cieslik, E. C., Zilles, K., . . . , Eickhoff, S. B. (2012). Across-study and within-subject functional connectivity of a right temporo-parietal junction subregion involved in stimulus-context integration. *NeuroImage*, *60*(4), 2389–2398.

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, *338*(6109), 953–956.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: a survey. *J Artific Intell Res*, *4*, 237–285.

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., . . . , Kessler, J. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex*, *46*(6), 769–780.

Kanai, R., Bahrami, B., Roylance, R., & Rees, G. (2012). Online social network size is reflected in human brain structure. *Proc R SocLond B*, *279*(1732), 1327–1334.

Keysar, B., & Horton, W. S. (1998). Speaking with common ground: from principles to processes in pragmatics: a reply to Polichak and Gerrig. *Cognition*, *66*(2), 191–198.

Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends Cogn Sci*, *8*(11), 501–507.

Kirby, S., & Hurford, J. (2002). The emergence of linguistic structure: an overview of the iterated learning model. In A. Cangelosi & D. Parisi (eds.), *Simulating the Evolution of Language* (pp. 121–148). London: Springer.

Krueger, F., Barbey, A. K., & Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends Cogn Sci*, *13*(3), 103–109.

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proc Natl Acad Sci USA*, *107*(6), 2717–2722.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends Cogn Sci*, *8*(12), 528–533.

Levinson, S. C. (2006). On the human interactional engine. In N. Enfield & S. Levinson (eds.), *Roots of Human Sociality* (pp. 39–69). Oxford: Berg.

Lewis, C., Freeman, N. H., Kyriakidou, C., Maridaki-Kassotaki, K., & Berridge, D. M. (1996). Social influences on false belief access: specific sibling influences or general apprenticeship? *Child Devel*, *67*(6), 2930–2947.

Lewis, P. A., Rezaie, R., Brown, R., Roberts, N., & Dunbar, R. I. (2011). Ventromedial prefrontal volume predicts understanding of others and social network size. *NeuroImage*, *57*(4), 1624–1629.

Mashal, N., Faust, M., Hendler, T., & Jung-Beeman, M. (2007). An fMRI investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain Lang*, *100*(2), 115–126.

Mates, A. W. (2010). Using social deficits in frontotemporal dementia to develop a neurobiology of person reference. In A. W. Mates, L. Mikesell, & M. S. Smith (eds.), *Language, Interaction and Frontotemporal Dementia: Reverse Engineering the Social Mind* (pp. 139–166). London: Equinox.

Menenti, L., Pickering, M. J., & Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers Hum Neurosci*, *6*, 185.

Mikesell, L. (2010). Examining perservative behaviors of a frontotemporal dementia patient and caregiver responses: the benefits of observing ordinary interactions and reflections on caregiver stress. In A. W. Mates, L. Mikesell, & M. S. Smith (eds.), *Language, Interaction and Frontotemporal Dementia: Reverse Engineering the Social Mind* (pp. 85–114). London: Equinox.

Milne, E., & Grafman, J. (2001). Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping. *J Neurosci*, *21*(12), RC150.

Mitchell, J. P., Ames, D. L., Jenkins, A. C., & Banaji, M. R. (2009). Neural correlates of stereotype application. *J Cogn Neurosci*, *21*(3), 594–604.

Mottaghy, F. M., Keller, C. E., Gangitano, M., Ly, J., Thall, M., Parker, J. A., & Pascual-Leone, A. (2002). Correlation of cerebral blood flow and treatment effects of repetitive transcranial magnetic stimulation in depressed patients. *Psychiat Res Neuroimaging*, *115*(1–2), 1–147.

Newman-Norlund, S. E., Noordzij, M. L., Newman-Norlund, R. D., Volman, I. A., Ruiter, J. P., Hagoort, P., & Toni, I. (2009). Recipient design in tacit communication. *Cognition*, *111*(1), 46–54.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Clarendon Press.

Noordzij, M. L., Newman-Norlund, S. E., de Ruiter, J. P., Hagoort, P., Levinson, S. C., & Toni, I. (2009). Brain mechanisms underlying human communication. *Frontiers Hum Neurosci*, *3*, 14.

Noordzij, M. L., Newman-Norlund, S. E., de Ruiter, J. P., Hagoort, P., Levinson, S. C., & Toni, I. (2010). Neural correlates of intentional communication. *Frontiers Hum Neurosci*, *4*, 188.

Osborne, M. J. (2004). *An Introduction to Game Theory*. New York: Oxford University Press.

Owings, D., & Morton, E. (1998). *Animal Vocal Communication: A New Approach*. New York: Cambridge University Press.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid *natural* scene categorization in human visual cortex. *Nature*, *460*(7251), 94–97.

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: you catch it from your sibs. *Child Devel*, *65*(4), 1228–1238.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav Brain Sci*, *1*(4), 515–526.

Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proc Natl Acad Sci USA*, *105*(23), 7936–7940.

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

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cogn Sci*, *16*(3), 147–156.

Sabbagh, M. A. (1999). Communicative intentions and language: evidence from right-hemisphere damage and autism. *Brain Lang*, *70*(1), 29–69.

Sallet, J., Mars, R. B., Noonan, M. P., Andersson, J. L., O'Reilly, J. X., Jbabdi, S., . . . , Rushworth, M. F. (2011). Social network size affects neural circuits in macaques. *Science*, *334*(6056), 697–700.

Sass, K., Krach, S., Sachs, O., & Kircher, T. (2009). Lion – tiger – stripes: neural correlates of indirect semantic priming across processing modalities. *NeuroImage*, *45*(1), 224–236.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, *42*(11), 1435–1446.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *BehavBrain Sci*, *36*(4), 393–414.

Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. *Proc Natl Acad Sci USA*, *107*(20), 9388–9393.

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat Rev Neurosci*, *10*(12), 885–892.

Schultz, J., Friston, K. J., O'Doherty, J., Wolpert, D. M., & Frith, C. D. (2005). Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. *Neuron*, *45*(4), 625–635.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension:n FMRI study. *Cereb Cortex*, *22*(7), 1662–1670.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proc Natl Acad Sci USA*, *104*(18), 7361–7366.
Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*(3), 617–627.
Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: the role of the right ventromedial prefrontal cortex. *J Cogn Neurosci*, *15*(3), 324–337.
Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.
Shultz, S., Lee, S. M., Pelphrey, K., & McCarthy, G. (2011). The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. *Soc Cogn Affect Neurosci*, *6*(5), 602–611.
Snowden, J. S., Neary, D., & Mann, D. M. (2002). Frontotemporal dementia. *Br J Psychiatry*, *180*, 140–143.
Sperber, D., & Wilson, D. (2001). *Relevance*: *Communication and Cognition*. Oxford: Blackwell.
Steels, L. (2003). Evolving grounded communication for robots. *Trends Cogn Sci*, *7*(7), 308–312.
Stolk, A., Hunnius, S., Bekkering, H., & Toni, I. (2013). Early social experience predicts referential communicative adjustments in five-year-old children. *PLoS ONE*, *8*(8), e72667.
Stolk, A., Noordzij, M. L., Verhagen, L., Volman, I., Schoffelen, J.-M., Oostenveld, O., . . . , Toni, I. (2013). Cerebral coherence between communicators marks the emergence of meaning. Paper presented at the 43rd annual meeting of the Society for Neuroscience, San Diego.
Stolk, A., Noordzij, M. L., Volman, I., Verhagen, L., Overeem, S., van Elswijk, G., . . . , Toni, I. (2013). Understanding communicative actions: a repetitive TMS study. *Cortex*, *51*, 25–34.
Stolk, A., Verhagen, L., Schoffelen, J. M., Oostenveld, R., Blokpoel, M., Hagoort, P., . . . , Toni, I. (2013). Neural mechanisms of communicative innovation. *Proc Natl Acad Sci USA*, *110*(36), 14 574–14 579.
Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *J Cogn Neurosci*, *10*(5), 640–656.
Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn Sci*, *10*(7), 309–318.
Thagard, P. (1989). Explanatory coherence. *Behav Brain Sci*, *12*(3), 435–467.
Tognoli, E., Lagarde, J., DeGuzman, G. C., & Kelso, J. A. (2007). The phi complex as a neuromarker of human social coordination. *Proc Natl Acad Sci USA*, *104*(19), 8190–8195.
Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.
Turesson, H. K., & Ghazanfar, A. A. (2011). Statistical learning of social signals and its implications for the social brain hypothesis. *Interaction Stud*, *12*(3), 397–417.
van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S. A. (2012). Pragmatics in action: indirect requests engage

theory of mind areas and the cortical motor network. *J Cogn Neurosci*, *24*(11), 2237–2247.
van Berkum, J. J., van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *J Cogn Neurosci*, *20*(4), 580–591.
van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional communication: computationally easy or difficult? *Frontiers Hum Neurosci*, *5*, 52.
Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., . . . , Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*(1), 170–181.
Volman, I., Noordzij, M. L., & Toni, I. (2012). Sources of variability in human communicative skills. *Frontiers Hum Neurosci*, *6*, 310.
Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B. G. (2004). Understanding intentions in social interaction: the role of the anterior paracingulate cortex. *J Cogn Neurosci*, *16*(10), 1854–1863.
Weder, N. D., Aziz, R., Wilkins, K., & Tampi, R. R. (2007). Frontotemporal dementias: a review. *Ann Gen Psychiatry*, *6*, 15.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Devel*, *72*(3), 655–684.
Willems, R. M., de Boer, M., de Ruiter, J. P., Noordzij, M. L., Hagoort, P., & Toni, I. (2010). A dissociation between linguistic and communicative abilities in the human brain. *Psychol Sci*, *21*(1), 8–14.
Willems, R. M., Benn, Y., Hagoort, P., Toni, I., & Varley, R. (2011). Communicating without a functioning language system: implications for the role of language in mentalizing. *Neuropsychologia*, *49*(11), 3130–3135.
Wittgenstein, L. (1953/2001). *Philosophical Investigations*. Oxford: Blackwell.
Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. A. (2009). Action understanding in the superior temporal sulcus region. *Psychol Sci*, *20*(6), 771–777.
Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc Natl Acad Sci USA*, *107*(15), 6753–6758.