19 Key Issues and Future Directions: Interactional Foundations for Language

STEPHEN C. LEVINSON AND IVAN TONI

1. The Design Features of Interactive Abilities That Lie behind Successful Language Use

The faculty for language is so bound up with the success of our species that without it we would still probably be a middle-sized inarticulate ape in Africa. Certainly, there would be no large-scale polities, no elaborated culture and technology, and no science. It may then seem churlish to question its preeminence, its ability to single-handedly transform us into cultural animals with a collective consciousness. But that is what this part of this volume sets out to do.

Consider for a moment a list of the failings of language (where "language" is understood just as the meaning of the words expressed), many of which have been usefully exposed by attempts to make machines comprehend natural languages (Jurafsky, 2003):

a. Natural language vocabularies are limited, hence most words are semantically generally over many meanings (consider *Aunt* which might denote one's father's sister, or mother's sister, or father's brother's wife, or mother's brother's wife, not to mention the many possible denotations of *great aunt*), or just plain vague (How big is a *heap*? When does a *mist* become *fog*, or a *bush* become a *tree?*).
b. Virtually all natural language sentences are multiply ambiguous, as in *He saw her duck* or *Visiting academics can be tiresome.*
c. Virtually all utterances require a good dose of common-sense understanding under interpretive rules of thumb: *It's not impossible he'll still come* suggests it is unlikely (Levinson, 2000), *The room was huge and the speaker couldn't be heard* suggests that the speaker was in that room and the size of it was the reason he couldn't be heard (Clark, 1977).
d. Language has distinct limits: it seems ill equipped to express shapes, precise hues, faces, or emotions (Levinson & Majid, 2014).
e. The main point of an utterance, the speech act expressed (e.g., advising, offering, requesting), is rarely encoded directly, but rather contextually inferred (Levinson, 2013).

The list could be multiplied, but the point is clear: language only works because it is supplemented by a great deal of further information, including conventions or heuristics under which it is used. One of the enduring insights of pragmatics—the study of language usage—is that language presupposes, indeed rides on, a vast infrastructure of human capacities for communication, partially independent of language. This is not simply the insight that there are forms of human communication independent of language—pictorial signs, gesture and "body language," music, and so forth—but rather the insight that language could not possibly work without a prior framework of mutual assumptions, interpretative heuristics, and interactional norms. Most observers and practitioners of the language sciences hugely overrate the independent efficacy of language alone—it is a seemingly miraculous ability, but the miracle lies very largely in these unseen foundations that allow it to operate.

This section of the book seeks to throw some light on these underlying, unseen foundations on which our linguistic abilities rest. These foundations play a crucial role in nearly every current use of language, but also in the creative abilities and the cooperative assumptions under which new languages have arisen, and they throw a great deal of light on the great unsolved mystery of language, namely how it ever evolved in the first place. But what are these foundations exactly?

The chapters gathered here agree that the focus should be on interactive uses of language, for conversation and task-oriented interactive language use forms the great bulk of ordinary language usage and the context in which language is learned by children and almost certainly arose in evolution. So what are the hidden background properties that make language

work so well in this interactional niche? One way to think about this is to entertain the following thought experiment: What would we need to build into a robot, beyond and above linguistic abilities themselves, for such a machine to be able to use language in a humanoid way? Table 19.1 provides a provisional and partial list (drawn partly from Levinson, 2019). The list starts off by pointing out that natural interactive language usually comes in a multimodal format: there is first of all the prosody (the duration, pitch, amplitude not intrinsically specified by the linguistic system) of the vocal signal; then there are the gestures of the hands, the raised eyebrows and furrowed brow of the face, the nods and even the blinks (Hömke, Holler, & Levinson, 2017), that are all packaged up to construct a single complex multimodal message (see Clark, chapter 17 of this volume, for illustrations of the expressive power of multimodal communication). In some cases, the entire burden of communication is shifted off the voice to the hands (see Goldin-Meadow, chapter 16).

The list continues with aspects of the context of verbal interaction, which canonically takes place under the umbrella of cooperative assumptions, with far-reaching inferential consequences (Grice, 1989). Interactive language usage normally takes place in the context of, and indeed constitutes, a form of joint action: participants have some shared goals, and to that end they each contribute to them. Next comes the physical context in its epistemic wrapper as it were: the mutual knowledge of what is salient in the environment, affording, for example, pointing to it, and more broadly the shared knowledge that each participant will assume the other shares. Arguably, the whole point of communication is primarily to convert a knowledge asymmetry into symmetrical, matched epistemic states (see Toni & Stolk, chapter 18). Then comes the discourse context—the parsed sequential environment in which an individual utterance is enacted. For interactive language use, the alternating turns at talk provide a crucial environment for response interpretation—whether some utterance is an answer,

Table 19.1

Some ingredients, beyond language, for a language-using robot (after Levinson, 2019)

MEDIA	
(Function: communication)	Multimodal signals
	Language
COOPERATIVE UMBRELLA	Joint action: shared goals with distributed sub-goals;
	Additional motivations, e.g. politeness, signaling of appropriate affect
EPISTEMIC CONTEXT	Salient properties of physical context; presumed 'common ground' and background knowledge
DISCOURSE CONTEXT, ACTION SEQUENCES	
(a) Alternating turns (Functions include legibility, opportunities for repair)	Speech act mapped to language
(b) Action sequences (Functions include structuring exchange)	Adjacency pairs (e.g. Q-A)
	Complex sequences: e.g. Insert pairs (e.g. Q-Q-A-A)
(c) Simultaneous actions (Function: coordination, ritual)	e.g. shaking hands, laughing
META-COMMUNICATION (Function: check communication)	Repair
(Function: confirm message receipt)	Feedback tokens (e.g. *uhuh*)
TIMING (Function: indicates state of processing is 'on time')	Turn-taking timing
(Function: 'clock speed' check)	Synchronicity (as in shaking hands)
(Function: 'message received now')	Timing of feedback
LEGABILITY	
(a) of attention (Function: indicates current focus of processing)	e.g. gaze readability
(b) of intention (Function: aid predictive processing)	e.g. of gesture signal vs. instrumental action

an assertion, a correction, and so on, will depend on the prior utterance. This involves a complex mapping of utterances onto speech acts, for which we currently lack deep understanding (Levinson, 2013). Then there is a sort of "grammar" of action sequences, which can be quite elaborate (see Levinson, chapter 14). The whole system only works because there are ways of invoking "repair" when a prior utterance cannot be construed—this mechanism of repair is fundamentally metacommunicative, as is the system of back-channel signals that acknowledge receipt without intruding on the speaker's turn. All of this is conducted under tight timing constraints, for example, turns at talking are usually separated by no more than 200 ms, literally near the duration of a blink of the eye, and delays of 600 ms or more in response begin to be interpreted as a hearer's problem (see Bögels & Levinson, 2017). Finally, and especially problematic for a robot, all these signals have to be legible—thus if my gaze is directed at the bottle when I say "this," the robot would have to be able to "read" the direction of my gaze. Most complex of all is that understanding an utterance may require "mind reading"—the ability to recover not only the surface content of an utterance, but the underlying role it is meant to play in the joint action in which we are engaged.

Summarizing, we can say that building an interactive robot with these properties would be extremely challenging (see Gluck and Laird, 2019), not least because for many of these properties we have only the haziest idea about the underlying cognitive mechanisms that instantiate them. We return to the future challenges in section 3.

2. The Chapters in This Part

The individual chapters in this part can be read as contributing to the overall picture of this large, implicit basis for human communication. Clark points to a fundamental gap in our understanding of human communication. We tend to think about communication as the transfer of propositional information—a view that founders on many rocks in fact, including the fact that what is traded in interactional exchange is speech acts, not propositions. Clark notes that many utterances *depict*, that is, convey an iconic representation of a described situation, as when a gesture indicates the manner in which an action was done, or a facial expression depicts the horror of a witnessed scene. In an elaborate taxonomy of such depictions, he notes that sometimes they fill a slot in a sentence frame, as in *Suddenly the car went grrrrrrunk!* where the acoustic depiction fills an adverbial slot. We have no adequate theory about how such depictions, acoustic or gestural, can be unified with the rest of the message—depictions are analog, "pictorial" elements that can nevertheless contribute to the content of utterances in crucial ways.

Goldin-Meadow looks at the "big-bang" of language origins, as discernable in the microcontexts of a deaf child shielded from institutional help, growing up with hearing parents who do not command sign language. Under these circumstances, infants actually create a mini sign language, one that may be understood by the parents but not symmetrically mastered by them. Known as *homesign*, these systems are idiosyncratic, and yet they show some striking similarities in design and usage. They have a strong basis in the iconicity and enactment that Clark has also noted as characteristic of natural speech. These systems over time come to exhibit grammatical categories such as noun versus verb, hierarchical structure, sentence modifiers such as negation and interrogation, and other properties typical of language. Since the parental input, typically speech and gesture, has none of these properties discernible to the deaf child, the child appears to create them de novo—there is some interaction between human nature and the communicative situation that builds these mini systems. Without cultural elaboration they remain highly limited, however, but the transformation of such systems into much richer communication systems can be witnessed when multiple homesigners are brought together, as in the birth of the conventional Nicaraguan sign language.

Homesigners and their caretakers succeed in understanding one another to the extent that they do because they build on a fundamental incremental mechanism, the generation of "common ground," things we know each other knows. Toni and Stolk sketch recent developments in our understanding of this "conceptual alignment," the sharing of intentions, mental models, and interaction history: having once used a phrase like *the squiggle* to successfully refer, we know we can continue to do so. They point to the new field of experimental semiotics, where participants have to create successful signals de novo, just like the homesigners, but in controlled experimental circumstances that give insight into the underlying reasoning and its neurocognitive implementation. An interesting new development is the demonstration that oxytocin, a relatively simple neuropeptide, can induce improved performance in these communicative tasks. This observation suggests that relatively simple motivational factors, such as those mobilized and modulated by hormones, might have substantial cognitive consequences (Theofanopoulou, Boeckx, & Jarvis, 2017).

Levinson sketches a great slew of interactional assumptions and behaviors that we bring to verbal exchanges,

including the motivational system that promotes intense and prolonged communication. He argues that these are deeply embedded in human nature, and are thus strongly universal across all cultures and languages. They consist of elements such as the exchange of short bursts of nonoverlapping speech with rapid transitions of speakers, of strikingly similar action types and sequences, and even the rates and types of interactive repair. He argues that these have a prelinguistic origin, since they can be partially seen very early on in infancy, and even in part across other primate species. It is this strong preexisting background of interactional skills and expectations, and indeed motivations, that makes it possible for infants to bootstrap themselves up into the local language with all of its conventions. This "interaction engine" points to a likely deep phylogenetic background, and hints that there may be strong continuities with other species underlying the perceived only-human Rubicon of linguistic abilities.

Rossano (chapter 15) provides much further evidence for parallels in interactional patterns between humans and the great apes, our nearest cousins. He argues that studies of animal communication have been hampered by an information theoretic framework; instead he argues we should adopt the same kind of framework that conversation analysts (like Schegloff, 2007) have applied to human interaction, namely thinking about communication as the exchange of social actions (cf. speech acts like requesting, offering, refusing, and such). He argues that by adopting this perspective, one sees strong commonalities across human and ape interactions, down to the very kinds of social action and their temporal properties. The interactional perspective thus serves to bridge the Rubicon between human and animal communication systems, and in so doing points to the evolutionary origins of our communicational prowess.

3. Future Challenges

As we said in the introduction, for most of the properties that characterize efficient language use we have little understanding of the underlying cognitive mechanisms that drive them. Nor do we understand the origins of the often quite striking uniformities across languages and cultures—native ethology, rational agency, or emergent adaptation of tools for the communicative job. Grice (1989) outlined some principles that he felt followed from the cooperative nature of communicative exchange (see Levinson, 2000, for an account in terms of simple heuristics), which are responsible for a wide range of inferences beyond what is actually said. But how many such principles are there? And, again, where do they come from? The elephant in the room here is perhaps the nature of communicative intention recognition. Philosophical reconstructions of this process seem psychologically unrealistic (e.g., threatening an infinite regress of my thinking about what you will think my intent is, based on your thinking about my thinking about your thinking…), even though some mirroring of design processes seems to show up in neuroimaging studies (Stolk et al., 2013). Computational models seem to demonstrate the intractability of intention reconstruction outside a narrow range of goals (Blokpoel, 2015). It will be easier to make progress on some of the more superficial features in table 19.1, for example, it is currently unknown how multimodal signals are composed into a single message—we know that the bits that go together often do not precisely align in time, so we have here a substantial "binding" problem both from a comprehension perspective (Habets, Kita, Shao, Özyürek, & Hagoort, 2011; Holler et al., 2015) and, even more challenging, a production perspective (see Chu & Hagoort, 2014). We have recently made good progress on some of the timing questions (see, e.g., Levinson, 2016), the meta-communicative systems (see, e.g., Dingemanse et al., 2015, Hömke et al., 2017), and the action-sequencing systems (Schegloff, 2007; Kendrick et al., 2014), although only in the first do we have a sketch of the cognitive processes involved.

Future work could be usefully directed at the following targets:

- Developing tasks and experimental models that can illuminate the cognitive processing behind the recognition of intentions (cf. de Ruiter et al., 2010) and the creative creation of one-off signals and depictions.
- Understanding the *binding problem* in multimodal signals—determining what bits in the parallel streams of signals belong together, how participants know, and how communicators orchestrate these multiple channels.
- Building processing models for multimodal processing, common ground updating, signal disambiguation that can make predictions in the neurocognitive domain.
- Determining how context is brought to bear so rapidly to disambiguate and resolve reference, grammatical ambiguity, speech act assignment, and the like.
- Determining how much of the "interaction engine" is native endowment, how much is learned, and how much is naturally emergent in the context of interactive communication.
- Examining how best can we pursue the leads that seem to point to deep underlying homologies

across at least some of our nearest cousins, the other primate species.

- Using human neuroscience methods with high signal-to-noise (e.g., electrocorticography), it might become possible to reliably capture neuronal traces supporting conceptual alignment across interlocutors, that is, transient and nonstationary events contingent on the shared history of interaction during a communicative exchange.

REFERENCES

Blokpoel, M. (2015). *Understanding understanding: A computational-level perspective* (Doctoral dissertation), Donders Graduate School for Cognitive Neuroscience Series 195. Retrieved from Radboud Repository of the Radboud University Nijmegen, http://hdl.handle.net/2066/144897.

Bögels, S., & Levinson, S. C. (2017). The brain behind the response: Insights into turn-taking in conversation from neuroimaging. *Research on Language and Social Interaction*, *50*(1), 71–89.

Chu, M., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, *143*(4), 1726–1741.

Clark, H. H. (1977). Inferences in comprehension. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 243–263). Hillsdale, NJ: Lawrence Erlbaum.

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 Studies*, *11*(1), 51–77.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, *10*(9), e0136100.

Gluck, K. A. & Laird, J. E. (Eds.). (2019). *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*. Cambridge, MA: MIT Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*, 1845–1854.

Holler, J., Kokal, I., Toni, I., Hagoort, P., Kelly, S. D., & Ozyurek, A. (2015). Eye'm talking to you: Speakers' gaze direction modulates co-speech gesture processing in the right MTG. *Social Cognitive and Affective Neuroscience*, *10*, 255–261.

Hömke, P., Holler, J., & Levinson, S. C. (2017). Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, *50*(1), 54–70.

Jurafsky, D. (2003). Pragmatics and computational linguistics. In L. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 578–604). Oxford: Blackwell.

Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., & Levinson, S. C. (2014). Sequence organization: A universal infrastructure for action. Paper presented at the *4th International Conference on Conversation Analysis,* University of California at Los Angeles.

Levinson, S. C. (2000). *Presumptive meanings*. Cambridge, MA: MIT Press.

Levinson, S. C. (2013). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), *The handbook of conversation analysis* (pp. 103–130). Chichester, UK: Wiley-Blackwell.

Levinson, S. C. (2016). Turn-taking in human communication, origins, and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14.

Levinson, S. C. (2019). Natural forms of purposeful interaction among humans: What makes interaction effective? In K. A. Gluck & J. E. Laird (Eds.), *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*. Cambridge, MA: MIT Press.

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind and Language*, *29*(4), 407–427.

Schegloff, E. A. (2007). *Sequence organization in interaction*. Cambridge: Cambridge University Press.

Stolk, A., Verhagen, L., Schoffelen, J.-M., Oostenveld, R., Blokpoel, M., Hagoort, P., … & Toni, I. (2013). Neural mechanisms of communicative innovation. *PNAS*, *110*(36), 14574–14579.

Theofanopoulou, C., Boeckx, C., & Jarvis, E. D. (2017). A hypothesis on a role of oxytocin in the social mechanisms of speech and vocal learning. *Proceedings of the Royal Society B*, *284*(1861), 20170988. doi:10.1098/rspb.2017.0988.