# 2 *Human Communication and the Interaction Engine*

## 2.1 The Puzzle of Language Diversity

There are over 7,000 languages in the world today. Just as species can be grouped into taxa with their known relatives, languages can be grouped into language families, which are the largest groups where their member languages can be shown to be related. The 7,000-odd languages belong to over 400 such families – the relation between these families goes so far back in time that we cannot easily recover it.[1] In general we can track relatedness of languages through their vocabulary or structure to about 10,000 years back, although occasionally we may be able to go a little farther. Over the last quarter of a million years, an estimate for the time depth of anatomically modern humans, there have been perhaps half a million languages. Languages do not just differ superficially, as if they were the same basic structure in a different set of sound clothing as it were. Instead, they differ in every possible way: some have as few as a dozen distinctive sounds, others twelve times as many (depending a bit on how you count), some have such complex morphology (ways of building words from words) that a whole English sentence can be expressed in a single word, while others have no morphology. The literature is full of claims that all languages exhibit some structure, but these claims are based on inadequate samples. We now know for example that basic sentence structure is highly variable, and that all possible phrase orders or even no set phrase order can be found in different languages. If structural diversity is like a wild garden, so is meaning: languages differ wildly in the concepts they choose to lexify (encode as words) or grammaticalize. Even the human body, one of the few universal objects, is segmented quite differently

[1] Data from Glottolog (Hammarström *et al.* 2021), which lists 7,606 languages in 425 families or isolates (languages without any relatives, of which there are 181). Some authors think they can discern many connections between these families, but the scientific basis for these larger groupings is weak.

into named parts in different languages. Imagine a language with no words for numbers greater than three, or no words for left and right, or no words for relatives beyond parents and siblings – these are all attested. Instead of strong universals, we find enormous diversity, and instead of a limited set of alternatives, relatively unconstrained variation, albeit with tendencies for structural coherence.[2]

Now contrast all known animal communication systems. In most cases these have a finite set of signals with an instinctive basis, triggered by recurrent events in the environment (for example, threats) or biological needs (such as advertising for a mate or defending a territory). In some species, including songbirds, the shape of the signals may be partially learned, but the 'meaning', the function or triggering events, are fixed. In short, they lack the structural complexity (like the use of meaningless elements to construct meaningful ones, and complex hierarchical patterning), the indefinitely extended meanings, and critically, the deep variation across groups that human languages exhibit.

Examining the structural diversity of human languages shows a number of things. First, despite their cultural nature, languages seem to evolve or change remarkably like biological evolution, inheriting traits faithfully across generations, inventing new structures, losing old ones, although they hybridize more like plants than animals. Secondly, if we imagine a 'design space' constructed from all the known parameters of linguistic variation, languages can be shown to have spread out to explore many of the far corners of this space; only related languages are likely to cluster closely together.[3] What this suggests is that there are relatively few constraints on the directions that languages can evolve in, providing they retain learnability for the next generation. Most extraordinary of all, human languages can flip from the oral mode to the gestural as in the sign languages of the deaf, without losing expressive finesse.

The unparalleled variability of human communication systems does not argue against a biological basis for language: it is patently clear

[2] See Evans & Levinson 2009, Hammarström 2010, Levinson 2003b. On some recurrent patterns and biases, though, see Verkerk *et al.* in press.

[3] Harald Hammarström and I performed this experiment (in an unpublished work) on a sample of languages from the Nijmegen Typological Survey, but this can now be replicated on much larger typological samples, as in Skirgård *et al.* 2023.

that we have evolved over deep time a complex vocal tract with gymnastic tongue and the whole panoply of muscular and neural control that goes along with that, together with the associated brain adaptations. For example, although apes and humans share a very similar brain anatomy, the arcuate fasciculus – the white matter 'wiring' from Broca's area to Wernicke's area in the brain – is slightly extended in humans in a way that is probably crucial for speech production, and likely plays a role in our ability to vocally imitate by constructing a quick loop between what we say and what we hear.[4] But what the variability of languages does indeed argue is that the whole system has extraordinary degrees of freedom to vary, in a way that is unique to humans. There is no extended 'innate grammar' or 'language of thought' able to dictate the detailed structure and meaning of sentences, as used to be supposed. What we will see later is that this variability of language contrasts fundamentally with much stronger constraints on how it is typically used.

## 2.2 Communication without Language

It is a common enough experience to find that limited communication is possible even when individuals do not share a common language – we have all probably experienced this when making our way in a foreign land. In many of these cases of course the context is restricted and the likely messages guessable. But there have been historical occasions in which 'radical translation' across languages and cultures has occurred. I mentioned in Chapter 1 Thomas Henry Huxley's voyage on HMS *Rattlesnake*, one of countless early voyages as the world opened up to exploration, and how he managed to barter with the inhabitants of Nimowa and Rossel Island in New Guinea.[5] A much better-documented example occurred in the twentieth century: the Highlands of New Guinea were for over a hundred years imagined to be uninhabited, and it was only when adventurous gold prospectors flew over the mountainous areas and landed in 1933 that first contact was made. The first contact was filmed and makes interesting viewing – the miners mimed their connection to the aeroplane

[4] Rilling *et al*. 2008.

[5] Although Huxley was unable to land on Rossel Island, he later recognized the distinctive canoe type 100 miles away and made contact with the crew.

and indicated their peaceful intentions with open hands and they soon established trading relations.[6] And again in Chapter 1, I described my own epiphany of an encounter with a deaf Pacific islander. Clearly some systematic communication is possible not only without language but also without a shared cultural background.

A more controlled experiment is the observation of children born deaf to hearing parents in circumstances where the children are deprived of both cochlear implants and contact with institutional sign languages. In these circumstances children and their carers develop from scratch a 'home sign' system – a gestural system capable of communicating an open-ended range of messages. There are some notable resemblances across these systems, and although they do have their expressive limitations, they nevertheless serve as proof of the existence of surprising possibilities of communication in the absence of a conventional language.[7] I have briefly studied one such system on Rossel Island, the remote island off Papua New Guinea, analyzing the communication between a profoundly deaf adult called Kpémuwó (the source of my epiphany) and people in surrounding villages (see Figure 2.1).

One of the most surprising findings was that Kpémuwó was able to communicate not only what he was going to do today, but also abstract ideas like the view that a women's illness was caused by a god specialized in sorcery retribution (see Figure 2.2). However, even his own brother could not always be sure of his intended messages when they were less obviously grounded in the context.

Despite some limitations, these cases of communication without an established language raise the question of how this kind of communication can possibly work. On the view that the meaning of expressions is established by convention, it is indeed a puzzle. Nevertheless, there is a quite compelling account that circumvents conventional meaning. Consider the following circumstance: I am in a seminar when a friend comes in late with a cappuccino 'moustache' of frothy white milk on her upper lip. I signal to her by catching her eye and rubbing my upper lip vigorously. She thinks what on earth could I be doing, and starts to wonder about her own upper lip and wipes it clean! How does this work? She notes that my rubbing of my lip is too vigorous to be purely instrumental (no itch would warrant that); she knows that I am

[6] Connolly & Anderson 1987. [7] Goldin-Meadow 2003.

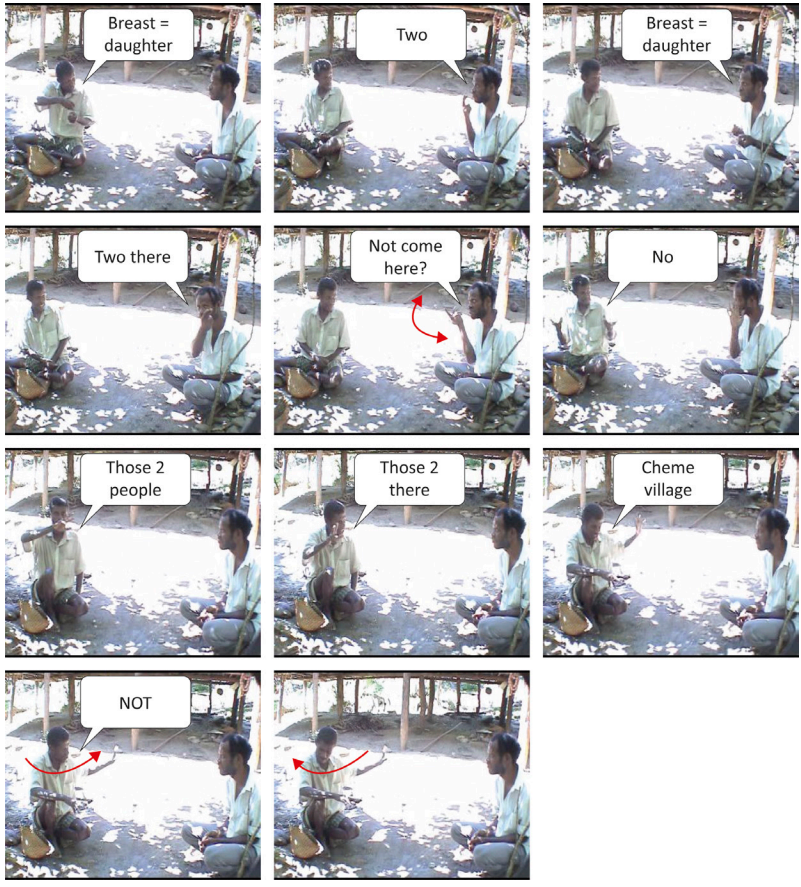


**Figure 2.1 Interactions between deaf-mute Kpémuwó (left) and a member of a nearby village (right) on Rossel Island concerning whether two daughters of a women will visit her on her sick bed.** (Stills from video shot by the author to be read left to right and top to bottom.)

looking at her, so engaging her attention; she knows that in the middle of a seminar people are reluctant to talk across the invited speaker. So, she thinks I might be trying to tell her something by my gesture. What connection could there be with an upper lip? She realizes she has just had a foamy coffee so perhaps the residue is on her lip….

The general form of this inference can be formalized as follows:

The signaller S means something Z by doing action A to the recipient R if


**Figure 2.2 Kpémuwó (left) communicates abstract ideas to his interlocutor (right), miming the eagle avatar of the anti-sorcery god.** (Stills from video shot by the author to be read left to right and top to bottom.)

(i) S intends the action A to cause the thought Z in R
(ii) S intends (i) to be achieved by R recognizing that (i) is the main motivation for A

This is the philosopher Paul Grice's theory of meaning – or theory of 'non-natural meaning' in his terminology.[8] The idea is that when I mean something successfully by an action I get you to have the thought I intended, and your recovery of the thought exhausts the purpose of my action (it has no instrumental function). Here, meaning is no longer a conventional relation between an arbitrary expression and a thought, but is instead a psychological notion: I'm trying to get

[8] There is a large secondary literature on Grice 1957, mostly concerned with the possibility of infinite regress: the recipient might be thinking what the sender is thinking the recipient would be thinking the sender would intend…

you to think about why I would have done the action, and once you recover the thought the action has exhausted its purpose. But how do I get you to successfully recover my thought?

First, note that we are quite good at thinking what the other would be thinking outside of communicative settings. The Nobel laureate Thomas Schelling showed experimentally that if we both have to come up independently with the same number to earn $100, we can easily beat the apparently infinite odds.[9] Just as we solve that problem by means of finding some number that I think you will think that I think is salient, so my rubbing my lip in the case of the cappuccino moustache has a salient solution (there is something on your lip). Secondly, it is clear that once we have successfully navigated this novel puzzle once, we can use the same clue or signal again unerringly – next time I wipe my lip when you come into the seminar late you will know immediately what I mean. This then is the basis for a theory about how conventions arise, conventional meanings amongst them.[10] Turning this idea back to the case of the deaf children and home sign systems, one can see immediately how it might be that child and caregiver can stabilize on the sign 'finger in mouth means I'm hungry' or the like.

This then gives us a theoretical reconstruction of how it is possible to communicate without language or indeed any conventional sign system. It is also possible to show experimentally that we can routinely perform this miracle. Suppose for example we devise a computer game where I can move my cursor around a nine-square board – my job is both to move my cursor to a given position and to signal where you have to put your cursor solely by means of my novel moves. I might inventively move my cursor to the square I want you to go to, then jiggle it back and forth, before going to my position. We can complicate the game and make it necessary to also signal orientation of the piece you have to move. Again, we can solve this. Meanwhile we can scan the brains of both sender and receiver, and what we will find is that during the planning of my move I activate especially an area of the brain involved in action interpretation, namely the posterior superior temporal sulcus (just behind the temple), and during the time that you are observing my signal you will activate the same part of the brain as part of a broader area.

[9] Schelling 1960. [10] See Lewis 1969.

This overlap of brain activation between sender and receiver fits with Grice's idea that the receiver must recover the sender's plan by thinking it through.[11]

Thinking about what the other is thinking is a crucial part of our ability to coordinate actions with one another, and communication is a kind of mental coordination. It is possible to trace the development of joint attention in infants – already within their first year of life and well before they master much language, infants are aware that both the caregiver and themselves are focused on some object or some joint action. The psychologist Mchael Tomasello and colleagues have shown that even adult apes only rarely achieve this kind of mental coordination, perhaps because they just lack the motivation and the interest in others' mental lives.[12] Yet being able to think 'I am aware that you are aware that we are both focusing on that' has the quite magical consequences that Schelling pointed out: if we lose each other in a giant department store (and only one of us has a mobile phone) we each have to think what the other would think about where to go in the hope of meeting, and the chances are we'll successfully coordinate, such as on the door where we came in. So, this ability extends beyond face-to-face interaction, but it is in interaction that it opens up the door to communication through mental coordination. It is this ability that makes it possible for children to acquire their first language. Although this account of communication focuses on the inference of others' plans or intentions, a coordination of feelings is also the root of empathy, which may have played an important part in the evolution of communication (a theme explored in Chapter 3).

## 2.3 The Interactional Niche for Language Use

We live in the first era in which a significant amount of the language we imbibe has been encoded remotely, either in the form of written language or in the form of broadcast or telecommunicated speech. But this should not distract us from the fact that the primary form of language use is in face-to-face communication – this is the forum where, crucially, language is learnt by infants and

[11] See Noordzij *et al*. 2010. [12] Tomasello 2022.

where the great bulk of human communication occurs (indeed until recently it was the sole arena for spoken language use). We each speak on average about 15,000 words a day,[13] but this is typically distributed in a pattern characteristic of human communication: we alternate little bursts of communication, first you take a turn, then me, then you, and so on. Each turn is on average about 2 seconds (or up to ten English words) long, so we each produce about 1,500 turns a day.[14] These machine-gun bursts of words are produced in rapid alternation, a matter that will occupy us in Chapter 3. Figure 2.3 illustrates this rapid alternation for a language of completely different cultural heritage than English, namely a Mayan language spoken in Mexico.

This very special context for normal language usage has striking properties. The face-to-face character allows the vocal signal to be embedded in a multi-modal display – the hands gesture, the face expresses, the eyes gaze at the recipient, and the whole body moves to express indignation, empathy, shyness, humour, or affection as appropriate. For speech, the obvious articulators are the tongue and the lips, but in fact over a hundred muscles are involved, including those that control breathing, glottal constriction, the position of the velum, and so on. Thirty-four muscles control the movements of each hand, twenty-six muscles in the neck control head movement, forty-three muscles control facial expressions, six muscles control each eye, at least seven muscles control arm and shoulder movement, and then there are muscles involved in the posture of the trunk. It is clear that the typical expressive use of the body in human communication involves the unconscious coordination of a veritable orchestra of muscles. A single gesture like a shrug is likely to involve many of these and coordinate with a facial expression and a hand gesture. Of all of these articulators, the hands, eye gaze, and facial expressions are probably the most important sources of communicative information beyond the voice. Interestingly, it has been shown that multimodal signals influence what you think you hear: for example a *ba* sound heard with lips visibly shaped to make a *ga* sound is more likely to be

[13] Estimate based on a cross-cultural sample in Mehl *et al.* 2007.

[14] These `s (derived from Yuan, Liberman, & Ciceri 2006; Levinson & Torreira 2015) obviously hide a great deal of variation, both individual and cultural. On the cross-cultural variation see Chapter 3.

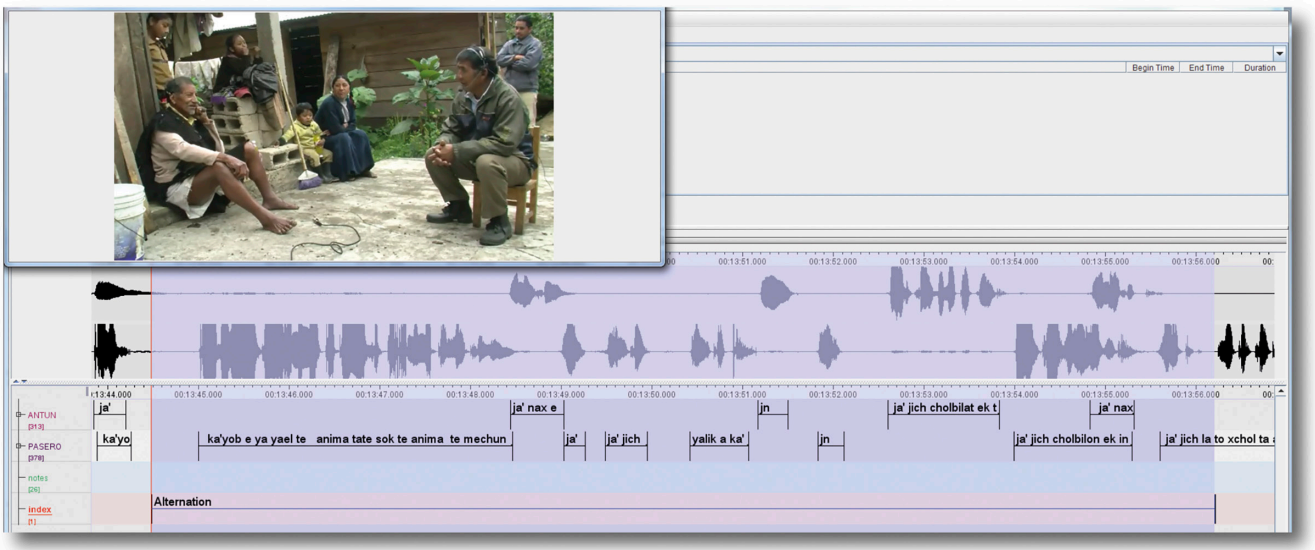


**Figure 2.3 Rapid alternation between two speakers of a Mayan language, Tzeltal.** Using separated audio channels for each interlocutor (top and bottom speech waves), we can use machine processing to get precise measurements of timing across two speakers in many languages. (Still from video shot by the author displayed in ELAN software.[15])

[15] Sloetjes & Wittenburg 2008; downloadable from the Max Planck Institute for Psycholinguistics, https://archive.mpi.nl/tla/elan

heard as *ga*, and a syllable is more likely to be heard as stressed when it cooccurs with a beat-like gesture.[16]

Multimodal communication thus implies simultaneous signalling in multiple streams of information, and interpretation involves a massive exercise of parallel processing. One abiding puzzle about multimodal signalling is that a single coherent message may be distributed not only over different articulators but also over a wide temporal span (as with a nod, a wink, and a smile accompanying a tease). There is therefore a 'binding problem', knowing which signals belong together, which is rarely solved by synchrony alone (see Figure 2.4).[17]

Much has been written about gesture, and the mysteries of what exactly motivates and shapes it.[18] When we speak we gesture; listeners scarcely ever gesture, rather it is part of the communicative performance of the speaker's role. We know that gesture is much more closely integrated with language production than used to be appreciated, matching or elaborating specific linguistic expressions, especially those describing spatial relations. Consequently, when languages differ in the way in which they package spatial information, they will also differ in where and how the gestures are coordinated.[19] Gesture can carry important complementary information missing from the speech stream, and experiments show that when people are asked to remember what was said, they in fact remember the gist derived from both gesture and speech.[20] There are good reasons to think that gesture may have been the primary flexible form of communication during early hominin evolution, for this is something we share with the apes, and specifically the chimpanzees (see Section 4.2). On the other hand, although pre-linguistic children point, they do not use the kind of small expressive iconic or mimicking gestures typical of adult communication, which constitutes a puzzle for the evolutionary story.

Multimodal communication depends a great deal on the visual channel. Consider for a moment human gaze. The white human sclera (the white of the eyes) has evolved since our lineage split from the nearest apes some 6 million years ago; it seems designed to make it easy to

[16] Bosker & Peeters 2021. [17] Holler & Levinson 2019.
[18] See, e.g. McNeill 2000, Seyfedinnipur & Gullberg 2014.
[19] Özyürek *et al.* 2008. [20] Kelly *et al.* 1999.

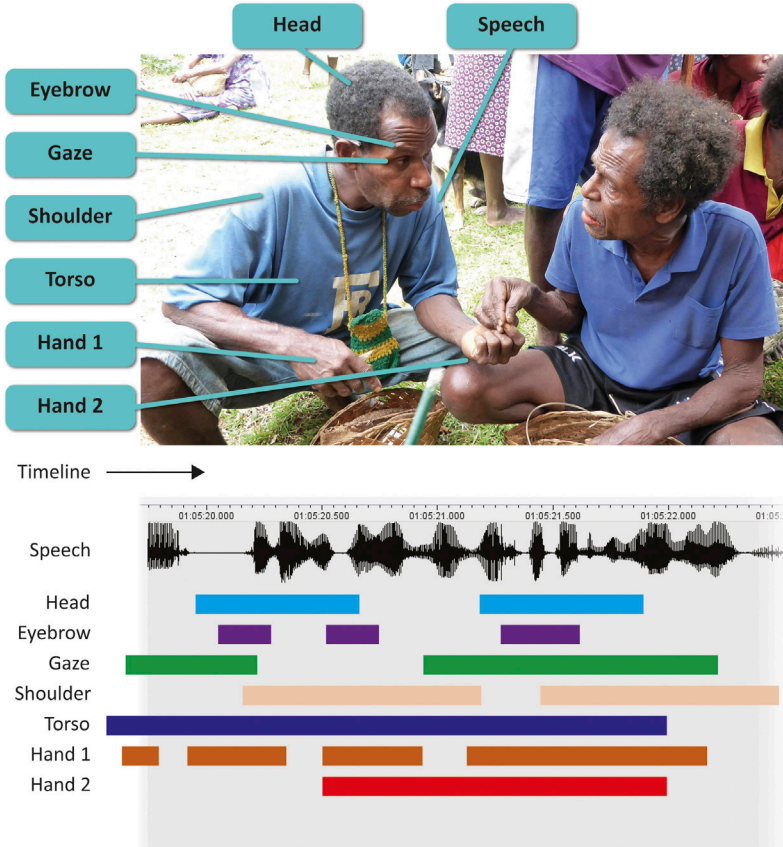


**Figure 2.4 The 'binding problem' for multimodal signals (Holler & Levinson 2019).** Speakers communicate using many overlapping bodily signals – how are these combined to form a single coherent message? (Still from video shot by the author on Rossel Island.)

track the other's gaze.[21] The human sclera contrasts with that of most of the apes, as shown in Figure 2.5. Interestingly, the bonobo sclera is slightly lighter than that of the more aggressive chimpanzees. The greater sociability and cooperation of bonobos compared to chimpanzees is also reflected in more mutual gaze.[22]

[21] Kobayashi & Kohshima 2001. Recent studies show the human-ape differences are actually more gradient than earlier assumed (Clark *et al.* 2023).
[22] Mulholland *et al.* 2020. See also Perea-García *et al.* 2019.


**Figure 2.5 The coloured sclera of apes versus the white sclera of humans** (from Kobayashi & Kohshima 2001).

As a result of the white sclera of the human eye, we can judge whether someone is looking at us with remarkable accuracy, within about 5 degrees of arc, and this is an ability apparent in neonates.[23] For many species this would be deleterious – my eyeing of your food or mate would be cause for battle. It must have arisen specifically to afford coordination and communication, perhaps in an evolutionary stage where hominins were primarily gesturers (see Chapter 4). Today it allows us to signal our attention to a speaker, and for a speaker to judge whether he or she has an attending recipient. There are cross-cultural variations in gaze behaviour, but even though they are striking to the traveller or ethnographer, the differences are quantitative in character and relatively small.[24] It is the availability of gaze, of course, that empowers the whole orchestra of multimodal signals which otherwise might pass unnoticed.

The importance of the other's gaze to us is reflected in some unobvious recent findings. For example, pupil dilation reflects both cognitive effort and empathy. We monitor others' pupils, and we tend to mimic their pupil dilation, and although chimpanzees do so too, in that case it is more confined to mother–infant interaction.[25] A second finding of interest is that human blinking behaviour performs many subtle functions in interaction, for example, blinks coordinate with the end of the other's turn and signal that the blinker has understood the point of the ongoing utterance by another. This can be shown by programming an

[23] Farroni *et al.* 2002, Mareschal *et al.* 2013.
[24] See Rossano, Brown, & Levinson 2009.
[25] Kret, Tomonaga, & Matsuzawa 2014.

avatar to blink at different points, with the result that long blinks by the listening avatar lead to speakers truncating their utterances.[26] So gaze is a crucial human instrument, used with great sensitivity, both for collecting information from the multimodal channels and for signalling interactive engagement in various ways.

A feature of the interactional niche is, as introduced in Chapter 1, the contingency between one utterance and the next. Utterances can be thought of as performing actions, 'speech acts' as they have been called. They may question, answer, request, promise, greet, swear, and the like. These functions may be partly independent of the form or the literal meaning of the utterance – 'It's five o'clock' might be an answer to a question, an excuse for hurrying off, a suggestion to go to the pub, or an announcement that the meeting is starting, all according to context and circumstance. In general, the form of the utterance constrains the possibilities but rarely uniquely individuates the action. Looking at the grammars of languages suggests that there may be some universal speech acts, like greetings, questions, and requests. Unfortunately, there has never been a proper survey to establish to what degree this is the case. Even where the functions may be generally the same, as with questions, the details vary – for example some languages rely heavily on an 'X or Y?' structure, while some rarely if ever use such a structure; some languages like English have a good inventory of questions words like *when*, *what*, *who*, *how*, and so on, but others have just one (glossing as 'which', so allowing the expression of 'which time?', 'which person?', and so on); some languages have an answer system where 'yes' in response to 'He's not there?' means he is not there, whereas in others it would mean he is indeed there.

But it is clear that many speech acts are culturally bound, because they rest on specific cultural institutions like marriage, religion, legal frameworks and the like. Consider the following exchange, translated from Yélî Dnye spoken on Rossel Island, offshore from Papua New Guinea (the island mentioned in Chapter 1). A and B are observing a man talking into a megaphone:

<1>
A: 'He's yelling into a bit of bush knife'
B: 'He's yelling under a mango tree'

[26] Hömke, Holler, & Levinson 2017.

To understand the exchange you need to know that the locals have a genre of pointed father-in-law jokes (they are a matrilineal people), and B's father-in-law killed his wife with a bush knife, while A's father-in-law died falling from a mango tree. It's a ritual exchange of jokes, and one of these father-in-law jokes should be retaliated with another, to the amusement of both parties.[27] All this is opaque to us. But the following, extracted from an interview with a prospective apartment renter in Los Angeles, would be equally opaque to them:[28]

<2>
A: 'I have a fourteen-year-old son'
B: 'Well, that's alright'
A: 'I also have a dog'
B: 'Oh I'm sorry'

Clearly, then, the contingency we are interested in holds not between the form of the utterances but between the underlying function or action in its discourse context. Especially prominent are pairs of actions like question-answer, greeting-greeting, complement-acceptance, request-compliance, and so forth. Called 'adjacency pairs' in the literature, they are typically, but not necessarily, adjacent, and the first part sets up the expectation for a relevant second part regardless.[29] They are not necessarily adjacent because principled further actions can intervene, as in:[30]

<3>
A: 'May I have a bottle of Mich?'
B: 'Are you twenty-one?'
A: 'No'
B: 'No'

Here a question-answer sequence is inserted into a question-answer sequence, and the original answer is thus postponed. The structure is what computer programmers call a 'push-down stack': the first question is followed by a second and then the answers match up with the questions from the inside out. This kind of embedding can be indefinitely repeated, with sequences inside sequences; and

[27] Levinson 2005. [28] Sacks 1992:757.
[29] Levinson 1983: chapter 6 offers a brief introduction; Schegloff 2007, and Clift 2016 provide detailed accounts.
[30] From Merritt 1976:333.

naturally-occurring cases go at least six embeddings deep.[31] This property of centre embedding is a paradigm example of recursion, once thought to be a hallmark of grammar (as in *The girl who Bill saw was Annie*), but in fact recursion in grammar is much shallower than the equivalent found in dialogue,[32] suggesting that our recursive abilities may actually have their origins in interaction structure. This is a property of interaction we take up in more detail in Chapter 3.

Adjacency pairs are central building blocks in interactional structure – they also operate in non-verbal interaction as when we greet each other with a wave, or as I gesturally offer to refill your glass and you put yours forward to accept. They can be elaborated by building out in front and behind, as in:

<4> (after Schegloff 2007:30)
N: 'Whatcha doin?' <-Pre-invitation
C: 'Not much' <-Go-ahead
N: 'Y'wanna drink?' <-Invitation
C: 'Yeah' <--Acceptance

Here the first question checks whether the conditions for an invitation obtain; that is shown by the nature of the response, which is not a simple account of what the recipient is doing, but rather a 'go ahead' signal. The invitation and acceptance – the core adjacency pair – follow.

This kind of exchange suggests another way of thinking about how interaction is structured. Recollect that experiments have shown that when planning inventive non-verbal communication, the sender thinks about what the recipient will think given a novel signal, and the recipient tries to reconstruct the sender's thinking (Section 2.2). The same kind of advance planning has been employed by the speaker in the exchange in Example 4: the plan to invite the recipient for a drink depends on her availability; questioning the availability makes clear to the recipient the nature of the overall plan, hence her 'not much' response, inviting the proposal. By giving the go-ahead the recipient also indicates that other things being equal she might be open to the

[31] Levinson 2013a.

[32] Karlsson 2007 shows that the maximum number of true centre-embeddings of this sort in spoken language is two, and in written language three. Here is one of his examples of a doubly embedded sentence: *[A lot of the housing [C-1 that the people [C-2 that worked in New Haven] lived in] was back that way.]*.

forthcoming proposal. So, from the initial utterance the recipient has recovered the speaker's plan, along with its likely subsequent steps, and by giving the go-ahead the recipient has also signalled her inclination to accept the proposal, other things being equal. His planned proposal and her planned acceptance are gently pre-adumbrated.

What this suggests is that coordinated interaction involves both deep planning and, more intriguing, the inference of the other's likely plans lying behind what has been said – an important type of 'mind reading'. That's how we can, for example, carry a table together down some steps, me guessing when and how you will likely move in what direction. Similarly, team sports rely on each team member guessing how the other will act, so achieving the rapid prospective coordination that might win the game. To the extent that it has been possible using artificial intelligence to construct conversational agents with human-like mentation, these systems also assume likely plans in the tight constraints of the functional encounters that are modelled. Note too that there need be nothing entirely mysterious about such mind reading: when an infant chimpanzee raises its arms, it may signal the wish to be carried, and thus initiate a sequence of actions between mum and infant – the signal itself evolves within the relationship from 'ritualization' of the sequence, so a part can stand for the whole sequence and effectively initiate it.

All this suggests of course that human communicative interaction is a species of cooperative behaviour. It is worth noting the similarities and dissimilarities between cooperative behaviour and its opposite, antagonistic behaviour. We can contrast, for example, cooperative hunting of game on the one hand, to the behaviour of predator and game on the other. Notice that both have quite a few properties in common. Antagonistic interaction also has quick responses that are contingent, so that when the fleeing rabbit veers left, the fox does too. Because the actions are responsive in antagonistic interaction, they are successive, just as they are in cooperative ones. Antagonistic interaction may also involve goal or plan reconstruction, as when the fox tries to intervene between the rabbit and bushes that would give it cover. Where they differ of course is that in the case of antagonistic interaction the goals of the two participants are inverse or zero-sum, and if there is any signalling, the signal is likely to be false (the rabbit may feint to the left but jog to the right). Less obviously perhaps, cooperative interaction has many properties that follow from the shared goals

of the participants – each action is done specifically to make its proximate purpose transparent, and its structure simple and timely. These properties are reflected deeply in the structure of language and communicative interaction.[33] Because the 'mind reading' involved in cooperative interaction is actually fraught with risk, various safeguards are provided by the structure of conversation. First, the contingent nature of adjacency pairs allows participants to see if the initial utterance was taken correctly – if your response indicates that you had a different 'John' in mind than the one I intended in my utterance, I can correct the interpretation. Second, there are procedures for indicating failure to hear or understand, both the blanket 'huh?' and the finely targeted 'He did what did you say to Anne?'. Conversation can thus be self-correcting, sequentially zeroing in on mutual understanding.

## 2.4 Key Design Features of Human Communicative Interaction

In this chapter we have reviewed some of the crucial features of human communicative interaction. Its core niche is face-to-face interaction, which allows a full range of multimodal signals. Interaction is based on sequences of contingent actions produced alternately by two or more participants. Language encodes these actions indirectly for the most part, with a many-to-many correspondence between utterance shapes and action types. These sequences can have elaborate structure of their own, and their understanding involves a reconstruction of goals behind proximate intents – a question may be a vehicle for a challenge or a complaint.

Table 2.1 lists in summary form some of these properties and the functional role they may play in communicative interaction. The table is a crude summary of the necessary ingredients that any constructor of an interacting robot would need to carefully mimic. The following paragraphs explain the role that each of these factors have in facilitating communicative interaction.

First, interaction opens multiple channels (row 1 in Table 2.1) and the multiple streams of information on different articulators allow the efficiency of coding, for example, of spatial angles by gesture, of

[33] These are studied in the branch of linguistics called 'pragmatics', along lines originally sketched by Grice 1975. See, for example, Levinson 2000, 2024.

**Table 2.1** *Interactional design features*

| | Property | Mode of expression | Functions |
|---|---|---|---|
| 1 | Channels | Kinesic-visual: hands, face, eyes, torso<br>Vocal-auditory: language, non-linguistic vocalizations | Communication of action, proposition, and attitude in multiple streams on different articulators |
| 2 | Action | Speech acts indicated in language in conjunction with context, sequence, kinesics | Performed in such a way as to aid intention recovery, while respecting social mores |
| 3 | Action Sequences | Adjacency pairs<br>Expansions of adjacency pairs<br>Opening/closing sequences and kinesics | Building contingent action sequences for joint action |
| 4 | Alternation across speakers | Linguistic turns, with multimodal extensions beyond turns | Short turns are the building blocks for sequences; alternation reveals any comprehension errors and enables fast repair |
| 5 | Simultaneous actions | Laughter, crying, greeting, interjections | Ritual functions |
| 6 | Meta-communication | Linguistic and kinesic feedback, continuers, repair initiators | Serves to show understanding, allows truncation, or signals need for repair, and signals (lack of) seriousness |
| 7 | Timing | Inter-turn vocal timing<br>Intra-turn pausing | Pauses can signal problems; fast timing can signal 'mental synchronization' |

emotion and attitude on the face, of precise proposition by language, and of the relation to other participants by body posture. Second (row 2 in Table 2.1), the inferential coding of action – determining whether this utterance is a question or request for example – allows language to perform indefinitely diverse actions, and allows some of them to be hinted at rather than stated, which plays a special role in human interaction (see Chapter 6). Nevertheless, whereas in antagonistic interaction intents may be disguised, in cooperative interaction actions must be designed to wear their goals on their sleeves. Third (row 3 in Table 2.1), contingencies across the actions done in alternating turns set up the very framework for cooperation, often in the form of highly structured sequences – a question expects an answer, an offer an acceptance or rejection, and so forth. As we have seen, sequences can have partially stereotypic structure, as with greetings, or they can be freely elaborated. Fourth (row 4 in Table 2.1), the alternations – taking turns – serve a number of functions: first, they leave the signal unmasked by a competing signal; secondly, they allow the prior speaker to see from the response whether his or her action was correctly interpreted; thirdly the short turns make for an efficient sharing of a limited channel. Fifth (row 5 in Table 2.1), although the taking of turns is a central property of human communication, there are exchanges of specific kinds of signal that are produced intentionally either simultaneously or in overlap, getting part of their special quality from the contrast with alternation. The simultaneous signals may be relics of an earlier, more primitive communicative system, since they include the reflexes of laughter and tears, cheers, howls, cries, expressive vocalizations of various sorts, and greeting behaviours – all communicative signals which can be recognized in other species. In addition, singing as an activity that bonds groups is in its simplest form done in unison. These signals serve largely 'ritual' functions, satisfying social requirements, and they stand in contrast to the 'one at a time' character of conversational turn-taking.

Sixth (row 6 in Table 2.1), a system of this sort dependent on 'mind reading' or intention recovery is very prone to the loss of mutual understanding; this is reduced by the existence of a special metacommunicative channel, which allows the priority signalling of hearing or understanding problems and their swift resolution. Finally (row 7 in Table 2.1), human communication exhibits a remarkable finesse

of fine timing. The vocal tract can produce sounds differentiated in just tens of milliseconds; the timing of turns, as we will see, is also measurable in small fractions of a second. Variation in timing can serve to indicate rapid understanding, or alternatively incomprehension, and precise timing is used in the prediction of next actions – for example, a late response to an invitation is predictive of a likely refusal.

These then are the features that mark out human communicative interaction and in part differentiate it from the communicative behaviour of other species (in addition of course to the unique properties of human language). We earlier characterized 'the interaction engine' in terms of the four key properties of multimodality, contingency, intent recognition, and fine timing, but Table 2.1 gives a fuller picture of the whole ensemble as we now understand it. The contributing elements are different enough that different evolutionary paths may have led to each of them. It is easy to imagine systems that would be strikingly different, for example featuring only overlapping or simultaneous signals, or only directly coded signal-to-function mapping instead of inferential communication, or which lacked extended highly structured sequences. Indeed, many examples of this type can be found elsewhere in the animal kingdom.

There might then be doubts that all of these properties make a package. Here some evidence is provided by the very complexity of the system, so that some infants are born with systematic deficits. The key syndrome here is autism, which is a spectrum of disabilities that typically involves systematic interactional impairment. Many autistic individuals have relatively high IQ and functioning language skills with large vocabularies but avoid face-to-face gaze, have poor timing, lack social smiling, hardly use gesture, have difficulty taking the point of view of the other, and above all have poor intention recognition.[34] But these are precisely difficulties involving a substantial portion of the properties of 'the interaction engine'. In contrast, Down's syndrome individuals may have poor language skills and lower IQ, but are well functioning interactants. These two syndromes serve to 'double dissociate' the bundle of skills that together make up the interaction engine (see Chapter 6).

[34] Frith 2008 provides an excellent introduction. See also Baron-Cohen, Leslie, & Frith 1985.