Demonstratives as attention tools: Evidence of mentalistic representations within language

Julian Jara-Ettinger\textsuperscript{a,b,c,1} and Paula Rubio-Fernandez\textsuperscript{a,k,1}

Published August 1, 2024.

Linguistic communication is an intrinsically social activity that enables us to share thoughts across minds. Many complex social uses of language can be captured by domain-general representations of other minds (i.e., mentalistic representations) that externally modulate linguistic meaning through Gricean reasoning. However, here we show that representations of others’ attention are embedded within language itself. Across ten languages, we show that demonstratives—basic grammatical words (e.g., “this”/”that”) which are evolutionarily ancient, learned early in life, and documented in all known languages—are intrinsic attention tools. Beyond their spatial meanings, demonstratives encode both joint attention and the direction in which the listener must turn to establish it. Crucially, the frequency of the spatial and attentional uses of demonstratives varies across languages, suggesting that both spatial and mentalistic representations are part of their conventional meaning. Using computational modeling, we show that mentalistic representations of others’ attention are internally encoded in demonstratives, with their effect further boosted by Gricean reasoning. Yet, speakers are largely unaware of this, incorrectly reporting that they primarily capture spatial representations. Our findings show that representations of other people’s cognitive states (namely, their attention) are embedded in language and suggest that the most basic building blocks of the linguistic system crucially rely on social cognition.

Significance

Our work sheds light on the interface between social cognition and language. We show that representations of interlocutor attention are embedded into one of the most basic word classes that appear across all languages: demonstratives. In ten languages spanning five language families, eight genera, and both written and nonwritten traditions, demonstratives internally encode when joint attention is established, and how to correct interlocutor attention when it is not. Our work also shows through computational modeling that this form of attention manipulation cannot be explained via forms of Gricean reasoning that are external to the linguistic system, suggesting that mentalistic representations are embedded in a universal component of language.

Successful communication routinely requires us to represent other people’s mental states (known as mentalistic representations)\textsuperscript{(1, 2)}, but determining the exact way in which mentalistic representations interact with language remains elusive. On the one hand, many complex social meanings expressed in everyday communication (such as when people speak ironically, use metaphors, or make an indirect request) can be explained through frameworks where domain-general social cognition is applied to language. In this approach—which we broadly refer to here as Gricean reasoning—people follow general principles about how a social agent ought to use language cooperatively, supported by mentalistic representations, but these representations are not internal to language\textsuperscript{(1, 3)}. At the same time, social inferences appear to permeate communication in real time\textsuperscript{(4, 5)}. Because of the speed and granularity with which social inferences shape interactive communication, representations of the interlocutor’s mind might not only be external to language but also operate internally within language itself\textsuperscript{(6–10)}.

Inspired by theories on the origins of language\textsuperscript{(11)}, we hypothesized that mentalistic representations might be embedded within a core component of language: demonstratives. Demonstratives (e.g., “here” and “there,” or “this” and “that” in English) are thought to have emerged remarkably early in the evolution of language (as their origins cannot be traced back to other types of expressions;\textsuperscript{12, 13}) , they are thought to be among the first grammatical words that children learn (together with pointing;\textsuperscript{14}) , and they appear in all world languages\textsuperscript{(15)}. Traditionally, however, demonstratives are thought to be grounded on spatial cognition\textsuperscript{(16)}: Their meanings encode distance relative to a physical reference point—the speaker’s own position in some languages (e.g., “there” means “far from me” in English), and both the speaker’s and listener’s positions in others (e.g., “ahi” means “far from me but close to you” in Spanish;\textsuperscript{17–19}).

By contrast, given their pervasive social use, we hypothesized that demonstratives go beyond their spatial representations and also internally encode mentalistic representations to support two key social activities: marking when a referent is in joint attention, and guiding the listener’s attention when it is not. Under this view, in identical physical situations with identical referential intent, speakers should use different demonstratives depending on where the listener is looking, even when the referent’s location remains constant. Moreover, this phenomenon should be challenging to explain as a form of domain-general external social cognition applied to demonstrative spatial meanings.
Cross-linguistic fieldwork studies have extensively documented that demonstratives are universally used to redirect the listener and establish joint attention during face-to-face communication (12, 15, 20, 21). However, representations of the listener’s attention are not believed to be part of the meaning of demonstratives: Their primitive indicating function is supposed to be “useful for drawing attention, but attention-direction is not a semantically specified function” (22).

Here, we took a cross-linguistic approach to test whether representations of the listener’s attention are embedded within demonstrative meanings (Fig. 1). If not only spatial but also mentalistic representations are part of the conventional meaning of demonstratives, we expect to find three sources of evidence. First, speakers should be sensitive to the listener’s focus of attention when selecting which demonstrative to use. Second, there should be cross-linguistic variation in the relative encoding strength of spatial and mentalistic meanings (given that meanings are conventionalized). By contrast, if attention direction effects emerge from simple forms of Gricean reasoning (which are thought to be universal), their qualitative structure should be cross-linguistically stable. Finally, we should find that people’s demonstratives use to direct listener attention cannot be explained by applying external mentalistic representations to demonstratives with purely spatial internal meanings, which we test via computational modeling.

Results

Experiment 1 sought to replicate the observational findings from fieldwork studies that demonstratives are used for attention direction, using a tightly controlled cross-linguistic study. This enabled us to quantify the exact ways in which people use demonstratives to direct attention in different languages and test for both cross-linguistic stability and variability. Experiment 1 first tested native speakers of six languages (Mandarin Chinese, English, Hebrew, Hindi, Italian, and Polish) with binary demonstrative systems hypothesized to encode distance from the speaker’s position (e.g., “this” and “that” in English; 17). Participants saw simple abstract depictions of a referential event that included both speaker and listener attention (see Fig. 2 for examples) and they were asked to select the most suitable demonstrative for the speaker’s referent. To make the task as conservative as possible, participants were told that listener’s attention was irrelevant and they were allowed to use a marked distal (e.g., “that one over there” in English), because these are thought to further enhance spatial contrasts (26).

Under the traditional view, demonstratives encode a referent’s position in space (e.g., “this one” refers to an object close to the speaker) with the pragmatic goal of aligning the speaker and listener visual perspectives on the referent (12, 16). Therefore, in standard semantic analyses, the listener’s initial attention is irrelevant to demonstrative meaning and choice (27, 28). Contrary to this view, demonstrative choice in experiment 1 not only encoded the referent’s spatial position but also the listener’s attention (Figs. 2 and 3). This was evident in three patterns of cross-linguistically stable results. First, proximal demonstratives signal objects in joint attention anywhere along the table (joint attention meaning: “the one we are both looking at”; $P = 3.41, P < 0.0001$), but this social meaning competes with the distal’s spatial meaning when the referent is farther away, leading to decreased use (Fig. 3A). Second, proximal demonstratives were also used when the listener was looking too far (i.e., pulling...
Fig. 3. Experiment 1 results collapsed across languages (see SI Appendix for by-language results). (A) Demonstrative choices in joint attention trials. Below each barplot is a schematic of the event: The speaker (S) is in front of the right-hand object, and the listener (L) is always standing in front of the target referent while looking at it. These results show how the proximal marks joint attention, but in farther locations, this use competes with the spatial meaning of the distal and marked distal demonstratives, reducing the preference for the proximal. (B) Demonstrative choices on the nine misaligned trials, split by referent location. Below each figure is the event schematic with the three attention locations. Target object is marked in light blue, and lines spanning from the listener (L) mark the different attention positions. Each barplot shows the distribution of demonstrative choices when the listener’s attention is in the location immediately below.

attention meaning: “look closer,” $\beta = 1.64; P < 0.0001$ (controlling for spatial location). This means the proximal is polymorphous in all languages (i.e., has two related meanings). Finally, distal demonstratives were used when the listener was looking too close (i.e., pushing attention meaning: “look farther,” $\beta = 0.52; P < 0.001$ (controlling for spatial location); Fig. 3).

Experiment 2 replicated and extended experiment 1 using real-world first-person pictures (Fig. 4), to ensure that our results were stable across experimental design choices and language systems. We first replicated experiment 1 results with a subset of the tested languages (Mandarin Chinese, English, and Polish), removing the marked distal choice to ensure this did not affect our conclusions (Experiment 2a). We once again found a strong sensitivity to attention (Fig. 4B). Beyond encoding space, the proximal demonstrative signaled joint attention ($\beta = 3.89; P < 0.0001$) and that the listener’s attention was too far ($\beta = 1.48; P < 0.0001$, controlling for spatial location), whereas the distal demonstrative signaled that the listener was looking too close ($\beta = 1.48; P < 0.0001$, controlling for spatial location).

Experiment 2b next tested whether our results generalize to languages with three-way demonstrative systems (i.e., with a proximal, medial, and distal demonstrative), which are hypothesized to encode distance relative to both the speaker and listener rather than the speaker alone (testing Japanese, Portuguese, and Spanish speakers).Similar to languages with two demonstratives, the proximal signaled that the listener was looking too far relative to the target ($\beta = 1.47; P < 0.0001$ controlling for spatial location), and the distal signaled that the listener was looking too close ($\beta = 1.85; P < 0.0001$ controlling for spatial location; see Fig. 4B). But in these systems, the medial now signaled joint attention ($\beta = 2.53; P < 0.0001$).

Our sample (Fig. 1) so far has revealed attention sensitivity in two-way and three-way demonstrative systems in Romance languages (Spanish, Portuguese, and Italian), as well as other Indo-European languages from different genera: Germanic (English), Indic (Hindi), and Slavic (Polish). These effects further appear in other language families, including Afro-Asiatic (Hebrew), Japonic (Japanese), and Sino-Tibetan (Chinese). Despite the typological diversity of our sample (so far including nine languages from seven language genera), these languages are unified in having a written tradition. Thus, in Experiment 2c, we ran our online task in-person in The Gambia, where we recruited native speakers of Mandinka (a West-African language from the genus Western-Mandé). Unlike the other languages in our sample, Mandinka is an oral language without a standardized written form. Participants had therefore not been formally taught Mandinka grammar, since they had not been schooled in their native language. Replicating the pattern from experiments 1 and 2, Mandinka speakers were significantly more likely to use the proximal demonstrative to mark joint attention ($\beta = 3.58; P < 0.0001$). When attention was misaligned, speakers used the proximal form to pull attention and the distal form to push attention ($\beta = 0.73; P = 0.015$; see SI Appendix for trial-by-trial results).

Stability and Variability in Attention Sensitivity. Gricean reasoning is thought to rely on foundational language-independent universal aspects of social cognition that are thought to be cross-linguistically stable. On the other hand, semantic meanings are conventions within a language group, and they are therefore thought to have greater arbitrariness (29). If mentalistic representations of listener attention were internally encoded in language, we would expect them to vary cross-linguistically in strength, whereas if they emerged from external Gricean reasoning alone, they should exhibit lesser cross-linguistic variation.

Fig. 5 visualizes the strength of attention effects on demonstrative use documented in Experiments 1 to 3. This figure highlights the universal pattern we found, where the proximal is used to pull attention toward the speaker when joint attention has not been established (negative values on the x axis), and the distal

Fig. 4. Experiment 2 visual materials and results. (A) Two sample trials of the experiment with Spanish prompts. Numbers indicate coded positions but were not shown to participants. (B) Experiment results. The Top row shows experiment 2a and Bottom row shows experiment 2b. Each figure shows demonstrative choice (each dot representing a participant selection on the corresponding trial) as a function of listener attention (x axis) and referent position (y axis), which was always in front of the listener. Border colors on panel (A) identify the corresponding trials for the Spanish plot in panel (B) (applying to the same positions on all other language plots).
and marked distal are used to push attention farther away from the speaker. Moreover, the proximal also marked joint attention (positive values on the y axis) in two-way demonstrative systems, but this effect was reduced in three-way systems, where the medial now marked joint attention.

At the same time, this figure also highlights the relatively high cross-linguistic variability observed in our data. To analyze this variability, we calculated the mutual information between the listener’s attention and people’s demonstrative choice, for each potential referent location (thus having four mutual information estimates for each dataset), given by

$$ I(D, A) = \sum_{d \in D} \sum_{a \in A} p(d, a) \log \left( \frac{p(d, a)}{p(d)p(a)} \right) $$  \hspace{1cm} [1]$$

where $D$ is the set of demonstratives available for the interlocutor and $A$ is the set of locations where the listener might be attending. Thus, given a particular target referent, mutual information calculates how much information is revealed about what demonstrative a speaker will use given where the listener is looking. Fig. 6 shows the average mutual information per dataset. This figure reveals two findings. First, there is indeed variability in how much sensitivity to attention speakers of different languages show, with Hebrew showing the strongest sensitivity to attention and English showing lowest. These findings therefore also highlight the risk of research relying on a single language—typically English—as representative of how people acquire and use language (30). The second finding here is that for all three languages used in experiments 1 and 2a (Polish, Chinese, and English), the availability of a marked distal (experiment 1) appears to increase sensitivity to attention. This suggests that, in two-way demonstrative systems, marked distals might be used to further guide listener attention, rather than to create more fine-grained spatial distinctions.

**Modeling Demonstrative Meaning.** Results so far indicate that people use demonstratives to mark joint attention (using the proximal in two-way system languages, and the medial in three-way system languages) and to mark referent position relative to the listener’s attention, rather than in absolute spatial terms.

To test whether these effects could be explained purely via Gricean reasoning that is external to language, we next formalized different theories of linguistic meaning through computational modeling (31, 32). Each model captures a different theoretical combination of demonstrative meanings and external Gricean reasoning: 1), allowing us to test which accounts can explain the three key effects: a spatial sensitivity to referent position, a mentalistic sensitivity to joint attention, and a mentalistic sensitivity to the direction in which listener attention must be corrected. For simplicity, we focus on the results of two-way demonstrative systems (Chinese, English, and Polish from experiment 2a and Mandinka in experiment 2c). Additional results with three-way systems—leading to identical conclusions—are available in SI Appendix. Throughout, we model Gricean reasoning through the Rational Speech Act (RSA) framework (3)—the dominant computational theory of Gricean pragmatics. RSA implements Gricean reasoning as a process of recursive social inference, where the listener aims to infer the speaker’s communicative intention with an utterance, and speakers choose their utterances accordingly. Through this process, a simple nonsocial lexicon can be used by social agents to generate complex pragmatic phenomena, such as hyperbole, irony, and scalar implicature (33-35).

We evaluated each model based on its ability to replicate the key qualitative effects and to quantitatively capture people’s linguistic preferences (see SI Appendix for full implementation details). We first implemented the spatial bootstrapping account, which tested whether purely spatial demonstratives enhanced with RSA pragmatics could produce attention-sensitivity effects. In this model, the proximal and distal meanings were implemented as simple probability distributions that increased (for the distal) or decreased (for the proximal) as a function of the listener’s distance to the referent. The spatial bootstrapping model captured some broad structure in people’s linguistic choices ($r = 0.43$, CI$_{95\%} = 0.17$ to 0.64; Top Left, labeled RSA, Fig. 7A). However, the qualitative analysis (Top Left, labeled RSA, in Fig. 7B) reveals that the model failed to produce any
attention-sensitivity effects. While this model showed a sensitivity to spatial location (visualized as decreased use of proximal as target on the x-axis increases in Fig. 7B), it did not capture the use of the proximal to mark joint attention, or any attention correction effects (visualized as constant preferences in each column, indicating lack of sensitivity to listener attention).

We next considered three possible implementations of the mentalistic meanings account, which vary in how much mentalistic content is embedded in demonstrative meanings. We first integrated joint attention into their meaning (Top Right plot, labeled RSA+JA in Fig. 7 A and B). We achieve this by adding a second meaning where the proximal can also refer to any object in joint attention (not only those in near space). Adding joint attention increased the quantitative fit to data ($r = 0.78$, CI$_{95\%}$=0.63 to 0.87). This model successfully showed a dual use of the proximal demonstrative to signal spatial proximity (decreased use of proximal as target distance increases) and joint attention (consistent use of proximal demonstrative in the diagonal; Top Right plot in Fig. 7B). However, this model did not produce attention correction effects (in the same figure, all columns show the same propensity, excluding joint attention cases).

Finally, two alternative approaches generated attention correction, one via Gricean reasoning and the other from demonstrative meaning. The Gricean reasoning extension consisted of introducing an assumption that the listener’s attention reveals their referential expectations (i.e., the listener is looking at whatever object they think the speaker is talking about; Bottom right plot, labeled Graded RSA+JA, in Fig. 7 A and B). In this approach, speakers pragmatically favor using the proximal when the listener attention is far away to counter the listener’s low prior belief that the referent is close to the speaker. Conversely, they favor the use of the distal when the listener attention is close (to counter the listener’s low prior belief that the referent is far from the speaker). This resulted in both a high quantitative fit ($r = 0.79$, CI$_{95\%}$ = 0.65 to 0.88) and evidence of all key qualitative phenomena (sensitivity to location, sensitivity to attention correction, and sensitivity to joint attention; Fig. 7B).

The second way to achieve the attention correction effects was by implementing them into the meanings themselves. This was achieved by having the proximal map onto objects between the speaker position and the listener’s attention and the distal map onto objects that were more distant from the speaker relative to the listener’s attention. This model was also able to capture the qualitative pattern of results and also produced a high quantitative fit to data ($r = 0.81$, CI$_{95\%}$ = 0.69 to 0.89; Bottom Right plots, labeled RSA+JA&AC, in Fig. 7 A and B).

Disentangling between Mentalistic Priors and Mentalistic Meanings for Attention Correction. The modeling results suggest that demonstratives internally encode spatial location and joint attention. However, they reveal that attention correction could, in principle, emerge either from within the meaning of demonstratives or purely through Gricean reasoning. The key difference between these two models is that the external Gricean pragmatic account predicts that attention correction effects will only appear when the speaker believes that the listener’s attention reveals their referential expectations.

Experiment 3 sought to distinguish between these two possibilities by manipulating the validity of an attention-based prior. Given the stability of the effects across all languages that we tested, in the remainder of the paper, we focused on Spanish as a representative system where people are showing spatial and mentalistic uses of demonstratives. The first condition was identical to experiment 1 (no distractor condition) and sought to replicate the attention correction effects. The second condition (distractor condition) showed the listener’s phone on the table (with dotted arrows showing the listener looking at their phone; see Fig. 8A, for example). Participants were told...
that, in these events, the listener was distracted looking at their phone. This created a situation where the listener’s attention needed correction, without revealing any referential expectations. If attention correction is internally encoded in demonstratives, it should appear in both conditions. That is, speakers should continue using demonstratives to correct attention, even when the listener is looking at an object that could not possibly be the referent. However, if attention correction is a pragmatic outcome that reflects an attention-based referential expectation, its use to redirect attention should vanish in the distractor condition. To ensure that any differences were not due to differences across samples, the aligned trials in both conditions were identical (i.e., with no phone), providing a baseline set of trials where we expected identical performance.

Participants’ demonstrative choice was nearly identical across conditions \( (r = 0.97; CI_{95\%} : 0.94 \text{ to } 0.99; \text{ Fig. 8B}) \). In both conditions, participants used the medial demonstrative to signal joint attention \( (\beta = 1.12, P < 0.0001 \text{ by a mixed-effects model; see SI Appendix}) \), with no difference across conditions \( (\beta = 0.07; P = 0.82) \), confirming similar effects. A posttest manipulation check confirmed that participants did not think that the listener’s attention revealed referential expectations in the distractor condition \( (M_{\text{distractor}} = 8.73 \text{ on a scale from 0 to 100, where 100 indicates full confidence that listener attention reveals their referential expectation}) \), but were significantly more likely to accept this possibility in the no distractor condition \( (M_{\text{no-distractor}} = 42.33, t(157.93) = -8.45; P < 0.0001 \text{ by two-tailed } t \text{ test}) \).

Critically, attention correction emerged in both conditions. Participants in the distractor condition used the proximal demonstrative to pull attention \( (\beta = 1.16, P < 0.001 \text{; controlling for location}) \), and the distal demonstrative to push attention \( (\beta = 0.81, P < 0.05 \text{; controlling for location}) \), hence revealing that attention correction is not a pragmatic phenomenon dependent on the listener’s referential expectation. At the same time, speakers used the proximal to pull attention significantly more in the no distractor condition \( (\beta = 1.15; P < 0.01) \), although we did not find an equivalent interaction for pushing attention with the distal \( (\beta = 0.29; P = 0.52) \). These results show that, like joint attention, attention correction cannot be explained as a Gricean pragmatic phenomenon and is instead best explained as an internal feature of demonstratives.

**Meta-Linguistic Awareness.** Having established that listener attention is internally encoded in demonstratives, we sought to test people’s meta-linguistic knowledge of this meaning. Unlike vocabulary words, such as nouns and verbs, grammatical words, such as demonstratives, encode implicit meanings that are difficult for speakers to explicitly access (and hence struggle to define; 36). In experiment 4, we therefore explored people’s awareness that demonstratives encode listener attention. We first showed native Spanish speakers a schematic of the scene in experiment 2 (i.e., the table with four objects and two silhouettes with no attention information, the speaker’s in position 4 and the listener’s in position 3) and asked them to explain when they would use the proximal (“este”), medial (“ese”), and distal (“aquel”) demonstratives in that situation. Over 80% of participants spontaneously reported that demonstratives express distance from the speaker (82%, 86%, and 84% for proximal, medial, and distal, respectively; see Fig. 9A). The next three most frequent definitions were i) the medial expresses distance from the listener (16%), ii) the distal expresses distance from the listener (14%), and iii) the proximal is used with pointing (14%). No other feature was mentioned more than 10% of the time. Critically, listener attention was not mentioned by any participant, and speaker attention was only mentioned in 4% \( (n = 2) \) of responses for the proximal demonstrative and in 2% \( (n = 1) \) of responses for the medial and distal demonstratives.

Participants next completed four misaligned trials with the strongest attention correction effects from the demonstrative choice task from Experiment 2b. After completing this task, participants were explicitly asked whether their decisions were influenced by the speaker’s distance to the referent, the listener’s distance, and the listener’s attention. While participants continued to report that referent distance was the primary cue for demonstrative choice, more participants now reported that listener’s attention mattered (Fig. 9B). Critically, participants who engaged in more attention correction during the demonstrative-choice task were more likely to report that listener attention...
attention was internal to the meaning ($\beta = 0.92; P < 0.01$). This suggests that, immediately after having used demonstratives to correct attention, and when explicitly prompted to consider whether listener attention affected their demonstrative choice, participants were more likely to recognize its role.

**Discussion**

At a high level, all linguistic communication directs our interlocutor’s attention to the content of our message, which might be grounded in the physical context (e.g., when we request an object) or recreated in the interlocutors’ minds (e.g., when we tell a story). However, many of these phenomena can be captured via mentalistic representations that are external to language. By contrast, here we showed that some of the most foundational mentalistic representations are embedded into one of the most basic words we use for social engagement: demonstratives.

We specifically found that demonstratives internally encode both spatial and mentalistic representations. Across all ten languages that we tested, people used demonstratives to redirect listener attention toward a referent and to signal when joint attention was successfully established (e.g., in two-way demonstrative systems, “this one” was polysemous between marking joint attention and the need to pull the listener’s attention). Using computational modeling, we found that these effects could not be explained via forms of Gricean reasoning alone, where domain-general social cognition is applied to demonstratives with internal spatial meanings. Our work points to a deep connection where representations of other people’s attention are intrinsic to language, embedded in the meaning of demonstratives.

We further found that, despite the universal use of demonstratives for attention correction, the strength and degree to which speakers do this varied cross-linguistically. Hebrew, Portuguese, and Polish were among the languages with the highest use of demonstratives to mark joint attention and engage in attention correction and English, Spanish, and Italian among lowest. This is a common feature of semantic meanings, as they are arbitrary mappings within a linguistic community (29). We do not know, however, what cultural forces boost or suppress attention correction across languages, and this is a question we hope future work will pursue.

At the first sight, the idea that demonstratives encode mentalistic representation might not seem so surprising since many languages have words that explicitly invoke mentalistic representations (e.g., mental-state verbs like “thinking” and “knowing” or emotion vocabulary like “happy” and “anxious”). However, these words occur at the lexical level: Word classes (such as verbs, nouns, and adjectives) that are representational (and can hence be defined; e.g., what is the meaning of “dog” or “barking?”), and which are in a constant flux of creation, modification, and loss (hence called an open-class system 37–39). Words at this level vary widely across languages and are not necessary for the core functioning of the linguistic system, therefore failing to provide evidence that mentalistic representations are internal to language (40).

However, language is also built upon a grammatical level: Linguistic devices such as prepositions (e.g., “in” vs. “on”), articles (e.g., “a” vs. “the”), and most central to our work, demonstratives. The grammatical level encodes nonrepresentational information, which is hard to conceptualize and define (e.g., what is the meaning of “the?” 36) and is typically procedural, implicit, and accessed automatically (41). Although many words at the grammatical level evolved from the lexical level, their meaning appears to “fossilize” once they become grammatical elements (13, 42), showing more static word classes with stable meanings (hence called a closed-class system). Given their early acquisition and pervasive use in everyday communication (43), the information encoded in grammatical words is likely to reflect basic cognitive representations that are essential for language users. Thus, the evidence we find of internally encoded mentalistic representations in demonstratives points to a fundamental role for social cognition at the foundations of language.

These results are consistent with other work showing that mentalistic representations can be embedded at the grammatical level. These include grammatical elements like articles (such as the Spanish definites and indefinites, “el” vs. “un,” which encode whether the incoming information is familiar or new to the listener; 44, 45), evidentials (such as Japanese particles “yo” and “tte,” which mark direct knowledge or hearsay, respectively; 46), and conjunctions (such as the English “but,” which prepares the listener for something unexpected). However, these and similar grammatical elements appear only in a minority of the world’s languages (47–49). In addition, since articles, evidentials, and conjunctions are complex grammatical elements that are thought to have emerged relatively late in language evolution, their reliance on mentalistic representations need not point to a foundational embedding of social cognition in language. Thus, past work suggested that it is possible for mentalistic representations to become internal to language, but this is not necessarily a general property of linguistic systems, which could have emerged late in the evolution of some languages. Thus, our work points to a universal internal embedding of mentalistic representations in language.

The fact that mentalistic representations are embedded in demonstrative meaning suggests that listener attention—and joint attention in particular—is being computed and accessed even in the simplest linguistic events. This is in line with theories of language that posit a fundamental connection between semantics and pragmatics (6). It is also consistent with research showing that tracking of social gaze enjoys a dedicated neural circuitry in the prefrontal-anterior cingulate cortex (50, 51), and tracking others’ attention might be shared with non-human primates (51–53). This opens the possibility that embedding representations of attention into language might have come at little cognitive cost, as these representations were readily accessible. This is also consistent with evidence that attention sensitivity in communications is already at work in infants as young as six months old (54). In addition, the finding that demonstratives—one of the building blocks of human languages—signal joint attention is consistent with the view that representations of joint attention are central to the human ability to share intentions and communicate (55) and points to the possibility that shared intentionality is reflected in grammar itself.

Our study contributes cross-linguistic data from ten different languages to an extensive typological literature on demonstrative systems (56–59). However, those studies employed very different methodologies, making it difficult to generalize their findings across languages. For instance, fieldwork studies have shown that Turkish and Yucatec Mayan have specific demonstrative forms to signal referents not yet in joint attention (21, 60, 61). Our cross-linguistic data extend those findings by showing, with the same experimental paradigm, how speakers of all languages use their demonstrative systems to reorient their listener’s attention, even if their language lacks a specific demonstrative with that function.

Our work also opens two key questions. First, would mentalistic representations become even more evident in more naturalistic...
interactions where interlocutors have additional tools at their disposal, such as eye contact, gaze direction, and pointing gestures (62, 63)? Second, what other types of mentalistic representations might be internal to language? Beyond joint attention, tracking mutual knowledge is also central to communication (64, 65). As noted above, past work has argued that these are likely embedded at the grammatical level as well (e.g., articles in English; compare “I bought a car!” vs. “I bought the car!” (66). Indeed, even demonstratives can be used to introduce new referents in a conversation (e.g., “I met this guy at the movies”) or to maintain the same discourse referent (e.g., “I have a new roommate. I will ask that guy if he can help us”), suggesting that demonstratives can also be used to manage common ground (67). This suggests that language might not only encode visual attention (where to look in visual space) but also mental attention (where to search in conversational history).

In conclusion, our findings show that, in addition to encoding semantic universals of space (16, 19), demonstratives also encode mentalistic information about listener attention in their meaning. In all ten observed languages, demonstratives encode spatial, attention-correction and joint-attention meanings. This confirms that mastering some of the most basic words present in all languages already requires speakers to represent the listener’s cognitive states (namely, their attention). This capacity then works in orchestration with Gricean pragmatics to further generate and extract social meanings, showing how mentalistic representations that are both external and internal to language operate pervasively to make human communication possible.

Materials and Methods

Each experiment recruited an independent sample of participants. All research was done in compliance with the University of Oslo’s Faculty of Humanities guidelines, which is overseen by the Norwegian Agency for Shared Services in Education and Research (Sikt). Because the studies reported here consisted of nonmedical online anonymous surveys, General Data Protection Regulations did not apply and they were therefore deemed automatically exempt and not requiring Institutional Review Board approval. In experiment 2c, participants were recruited in person in The Gambia College, permission was obtained from the Dean to conduct the study on campus, and data collection was still anonymous since those participants performed the same online task as those recruited via Prolific. Informed consent was obtained from all participants at the start of the online task.

Experiment 1.

Participants. Three hundred native speakers of Mandarin Chinese, English, Hebrew, Hindi, Italian, or Polish (50 per language group) were recruited via Prolific (Mean age = 30; 51% Female; 48% Male; 1% Other). Materials and procedure. The task was implemented in the online survey builder Qualtrics. Participants completed 18 trials, each showing an interaction between a speaker and a listener over 4 objects placed equidistantly along a table (see Fig. 1 and OSF repository for stimuli). The speaker was always situated above the fourth (right-most) object and looking at the target, which could be the same discourse referent (e.g., “I have a new roommate. I will ask that guy if he can help us”), suggesting that demonstratives can also be used to manage common ground (67). This suggests that language might not only encode visual attention (where to look in visual space) but also mental attention (where to search in conversational history).

In conclusion, our findings show that, in addition to encoding semantic universals of space (16, 19), demonstratives also encode mentalistic information about listener attention in their meaning. In all ten observed languages, demonstratives encode spatial, attention-correction and joint-attention meanings. This confirms that mastering some of the most basic words present in all languages already requires speakers to represent the listener’s cognitive states (namely, their attention). This capacity then works in orchestration with Gricean pragmatics to further generate and extract social meanings, showing how mentalistic representations that are both external and internal to language operate pervasively to make human communication possible.

Materials and Methods

Each experiment recruited an independent sample of participants. All research was done in compliance with the University of Oslo’s Faculty of Humanities guidelines, which is overseen by the Norwegian Agency for Shared Services in Education and Research (Sikt). Because the studies reported here consisted of nonmedical online anonymous surveys, General Data Protection Regulations did not apply and they were therefore deemed automatically exempt and not requiring Institutional Review Board approval. In experiment 2c, participants were recruited in person in The Gambia College, permission was obtained from the Dean to conduct the study on campus, and data collection was still anonymous since those participants performed the same online task as those recruited via Prolific. Informed consent was obtained from all participants at the start of the online task.

Experiment 1.

Participants. Three hundred native speakers of Mandarin Chinese, English, Hebrew, Hindi, Italian, or Polish (50 per language group) were recruited via Prolific (Mean age = 30; 51% Female; 48% Male; 1% Other).

Materials and procedure. The task was implemented in the online survey builder Qualtrics. Participants completed 18 trials, each showing an interaction between a speaker and a listener over 4 objects placed equidistantly along a table (see Fig. 1 and OSF repository for stimuli). The speaker was always situated above the fourth (right-most) object and looking at the target, which could be any of the first three (left-most) objects. Both speaker and listener attention were shown by body orientation and line of gaze. The listener’s position and attention were varied across trials parametrically. The speaker and listener were looking at the same object in half of the trials (Aligned-attention condition; three trials for each target position) and at different objects in the other half of the trials (Misaligned-attention condition). The object closest to the speaker was never the target because only the proximal demonstrative (e.g., “This one” in English) could be used in that position, not allowing to test for the role of listener attention. The speaker always had a speech bubble saying “Now I need...” (in the language of test) and participants were asked to select the demonstrative that best applied to the event.

The cover story for the task explained that the participant was moving houses and asked a friend for help packing his things in boxes. Participants were asked to adopt the role of the speaker and complete the request in the speech bubble by clicking one of three radio buttons (e.g., “this one,” “that one” or “that one over there” in English). The instructions highlighted that participants should treat the scene as an interactive scenario and imagine what their natural choice would be in each trial.

Participants were also told that their friend would sometimes be looking at the wrong object. However, it was stressed that the target object was always the one the participant/speaker was looking at and that was the object they needed to request. The instructions therefore treated the listener’s focus of attention as irrelevant to the speaker’s request.

To prevent participants from developing strategies ahead of the task, the sample display shown in the instructions included 3 nouns to choose from (“the camera,” “the plane,” “that doorknob”), rather than the demonstrative pronouns they would be offered during the actual task. The task lasted approximately 3 min (18 trials), and the trial order was randomized across participants.

Analysis approach. Analyses consisted of three mixed-effects models designed to target each key hypothesis. In each case, we used the maximal model structure that converged (68). The final random effects structure for each analysis is available in our OSF repository and in SI Appendix.

The first analysis consisted of a logistic mixed-effects model predicting proximal demonstrative choice (coded as a binary variable), as a function of whether interlocutor perspectives were aligned or not (coded as a binary variable). The maximal model that converged included random intercepts and slopes as a function of perspective for both subjects and languages.

The second analysis focused exclusively on trials with misaligned attention. It consisted of a logistic mixed-effects model predicting proximal demonstrative choice (coded as a binary variable), as a function of whether the listener should look closer (coded as binary variable set to true whenever the listener’s attention was farther away than the referent, and false otherwise) and the referent’s position. The maximal model that converged included random intercepts and slopes as a function of position for each participant and random intercepts for each language.

The third analysis focused on trials with misaligned attention and consisted of a logistic mixed-effects model predicting marked distal choice (coded as a binary variable), as a function of whether the listener should look farther (coded as a binary variable set to true whenever the listener’s attention was closer to the speaker relative to the referent, and false otherwise) and the referent’s position. The maximal model that converged included random intercepts and slopes as a function of position for each participant, and random intercepts for each language.

Experiments 2a and 2b. Experiments 2a and 2b were identical, with the difference only that the languages in Experiment 2a had 2-way demonstrative systems and those in Experiment 2b had 3-way demonstrative systems, hence leading to slightly different data and analyses.

Participants. A new sample of 300 native speakers of Mandarin Chinese, English, and Polish (experiment 2a) and Japanese, Portuguese, and Spanish (experiment 2b) were recruited via Prolific (50 participants per language group; Mean age = 31; 49% Female; 50% Male; 1% Other).

Materials and procedure. The 18 trials employed in experiment 1 were recreated using photo realistic images of two interlocutors standing across a table with four objects. For greater ecological validity, the pictures were not taken from a birds-eye perspective, but from the speaker’s point of view (see Fig. 3 and OSF repository for stimuli). To ensure the most naturalistic results, the pictures were taken with a set of SMI eye-tracking glasses, placed on the speaker. The target object was indicated by the speaker’s pointing gesture, which was also in the picture frame. The speech bubble was moved to the lower half of the table, where the speaker was standing. The task procedure was the same as in experiment 1, only that the Chinese-, English- and Polish-speaking participants were only offered two radio buttons (e.g., “this one” vs. “that one” in English), after we removed the marked distal demonstrative. Japanese-, Portuguese- and Spanish-speaking participants were offered three radio buttons to choose from because they all have 3-way demonstrative systems (e.g., “este” “ese” “aquel” in Spanish).

Analysis approach. Analyses followed a parallel logic to the ones from experiment 1, using a separate analysis for experiments 2a and 2b. In experiment
2b, analyses focused on medial demonstrative choice because that indicates proximity to the listener (see SI Appendix, Text for prediction details). The maximal structure for each regression is available in SI Appendix.

Experiment 2c.
Participants. Fifty native speakers of Mandinka were recruited in person at The Gambia College in Brikama, Gambia (Mean age = 23; 41% Female; 59% Male; 0% Other). All participants received monetary compensation for their time and the cost of transport.

Materials and procedure. The English version of the task from experiment 2a was translated to Mandinka by two native speakers. As in previous experiments, the task was set up and administered via Qualtrics, so data collection was anonymous. Participants performed the task on a laptop that was connected to the Internet through a mobile hotspot. They were asked to click one of two radio buttons in each trial, one for the proximal demonstrative and the other for the distal demonstrative (equivalent to “this one” vs. “that one” in English). Other than for recruiting and testing participants in person, the materials, task, and procedure were identical to those in experiment 1a (including languages with 2-way demonstratives systems).

Analysis approach. Analysis approach was the same as experiments 2a and 2b. Maximal converging models for each regression are available in SI Appendix. Note that only one regression is presented for use of proximal and distal because participants in this task could only use the two demonstratives (SI Appendix).

Experiment 3.
Participants. A total of 201 native speakers of Spanish were recruited via Prolific (Mean age = 31; 47% Female; 53% Male; 0% Other). Sample size was originally set to 100 per condition, but one participant timed out in the distractor condition, so an additional participant was automatically recruited and their data were retained for analysis.

Materials and procedure. The materials from experiment 1 were translated into Spanish and used as the no demonstrator condition. These materials were adapted for the distractor condition, such that in misaligned trials, a mobile phone appeared on the table, next to the object that the Listener was originally looking at in the no demonstrator condition. The Listener’s attention (indicated by her line of gaze) was modified so that she would always be looking at the phone in misaligned trials. The aligned trials in the distractor condition were identical to those in the no demonstrator condition. Since these two conditions were administered between participants, aligned trials served as a control to establish comparable baseline performance.

Importantly, the instructions in the distractor condition were also adjusted such that participants were told that their friend would sometimes be distracted looking at her phone. This was done to clarify that the phone was not a potential referent, but a source of distraction for the Listener.

At the end of the task, in both conditions, participants were asked whether they thought that when their friend was looking at her mobile phone (distractor condition) or at the wrong object (no demonstrator condition), it was because she thought that was the object she was going to be asked for next. Participants were asked to respond to this question using a 0 to 100 slider, ranging from “No, I don’t think so” to “Yes, I think so.”

Experiment 4.
Participants. Fifty native speakers of Spanish were recruited via Prolific (Mean age = 29; 54% Female; 46% Male; 0% Other).

Materials and procedure. The last experiment included 3 tasks, presented to all participants in the same order. In task 1, participants were first shown a schematic of the scene in experiment 2b, with a silhouette representing the speaker in position 4 and a silhouette representing the listener in position 3 (Fig. 2A). Participants were then asked when they would use “este,” “ese,” or “aquel” (i.e., the proximal, medial, and distal demonstratives) in that situation. Participants were asked to write their responses in an open box, in 1 to 3 sentences per demonstrative.

In task 2, participants were presented with four trials from the Misaligned-perspective condition in experiment 2b: the target referent was in position 2 in two of the trials and in position 3 in the other two trials, with the Listener attention being on either position 1 or position 4 (balanced). These trials were selected to elicit attention-correction (given our experimental design) while not presenting participants with any trial from the Aligned-perspective condition for comparison. As in experiment 2b, participants were asked to select one of the three Spanish demonstratives in each trial.

In task 3, which was always administered last, participants were given a multiple-choice task where they were asked to select those situations in which they had used “este,” “ese,” or “aquel” in the previous task. See SI Appendix for exact questions.

Analysis approach. Open-box responses in task 1 were coded by two independent coders. Details are available in SI Appendix. Coders agreed on 99.33% of cases (n = 1,192) and disagreed on 0.067% of cases (n = 8). Trials where coders disagreed were resolved through discussion (with no involvement of the authors).

Attention-sensitivity in task 2 was coded as an integer between 0 and 2, which indicated the number of times that participants switched demonstrative choice when location was held constant, but listener attention had shifted. That is, attention sensitivity was determined by comparing the two answers for the pair of trials where referent was in position 2 (and attention varied between positions 1 and 4) and the pair of trials where referent was in position 3 (and attention varied between positions 1 and 4). Task 3 was analyzed by testing whether participants selected listener attention as relevant (binary variable) as a function of the degree of attention correction that they exhibited in task 2. This was done through a mixed-effects logistic regression using the maximal random structure that converged to control for demonstrative being queried and participant (SI Appendix).

Data, Materials, and Software Availability. Anonymized participant judgments, analysis scripts, experiment stimuli, and computational modeling code data have been deposited in Demonstratives track attention: Evidence from 10 languages (69).

ACKNOWLEDGMENTS. We thank Damian Blasi, Juergen Bohnemeyer, Sammy Floyd, Natalia Levshina, Frank Mollica, Alexander Paunov, and Amanda Royka for helpful comments. Thanks to Vrinda Bhatia and Anwesha Mahapatra for help with creating the visual stimuli for experiments 2 and 4 and to Fatoumata Jallow for help with participant recruitment and data collection at The Gambia College. Thanks also to Shlomit Ben-Ami, Vrinda Bhatia, Anna Filip, Fatoumata Jallow, Anwesha Mahapatra, Karla Sophia, Satomi Sugaya, Quianru Tao, and Francesco Veneroni for help with translating the instructions and task materials from English to their native languages. This research was supported by funding from the Research Council of Norway (275505) awarded to P.R.F. and NSF (BCS-2045778) awarded to J.-J.E.
