PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Kendrick KH, Holler J, Levinson SC. 2023 Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Phil. Trans. R. Soc. B* **378**: 20210473. https://doi.org/10.1098/rstb.2021.0473

Received: 8 July 2022
Accepted: 27 January 2023

One contribution of 15 to a discussion meeting issue 'Face2face: advancing the science of social interaction'.

Subject Areas:
behaviour

Keywords:
turn-taking, gaze, manual gesture, multimodality, conversation analysis, transition-relevance places

Author for correspondence:
Kobin H. Kendrick
e-mail: kobin.kendrick@york.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6423930.

THE ROYAL SOCIETY PUBLISHING

Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions

Kobin H. Kendrick[1], Judith Holler[2,3] and Stephen C. Levinson[3]

[1]Department of Language and Linguistic Science, University of York, York YO10 5DD, UK
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
[3]Max Planck Institute for Psycholinguistics, Nijmegen, Gelderland, The Netherlands

KHK, 0000-0002-6656-1439; SCL, 0000-0001-8961-5316

Human communicative interaction is characterized by rapid and precise turn-taking. This is achieved by an intricate system that has been elucidated in the field of conversation analysis, based largely on the study of the auditory signal. This model suggests that transitions occur at points of possible completion identified in terms of linguistic units. Despite this, considerable evidence exists that visible bodily actions including gaze and gestures also play a role. To reconcile disparate models and observations in the literature, we combine qualitative and quantitative methods to analyse turn-taking in a corpus of multimodal interaction using eye-trackers and multiple cameras. We show that transitions seem to be inhibited when a speaker averts their gaze at a point of possible turn completion, or when a speaker produces gestures which are beginning or unfinished at such points. We further show that while the direction of a speaker's gaze does not affect the speed of transitions, the production of manual gestures does: turns with gestures have faster transitions. Our findings suggest that the coordination of transitions involves not only linguistic resources but also visual gestural ones and that the transition-relevance places in turns are multimodal in nature.

This article is part of a discussion meeting issue 'Face2face: advancing the science of social interaction'.

1. Introduction

A striking feature of human informal social interaction is that participants rapidly alternate speaker and addressee roles—in short, they take turns in sharing the communication channel. Turn-taking characterizes the primary context for language use, it is found in all cultures across the globe, it creates the interactional niche in which children acquire language and in which language presumably evolved, and it is even observed in the vocal and gestural communication of other species [1–4]. In the case of human conversational exchange, the timing and precision of turn-taking raises interesting questions about the underlying mechanisms—turns alternate with very short gaps (on average close to the minimum human response time of 200 ms) and with very few overlaps (typically less than 5% of the speech stream; [5]). A puzzle is that this rapid exchange of turns is so seamlessly accomplished despite the fact that individual turns are unpredictable in length and content, the number of participants can vary, and provision has to be made for longer turns to allow, e.g. the telling of a story.

The remarkable precision of turn-taking seems to be grounded in principles that were first outlined in in the field of conversation analysis (CA), since tested in 40 years of close observation and detailed description of turn-taking in naturally occurring conversations (e.g. [6–14]). Research in this tradition has outlined a set of principles which we will call 'the standard model', organizing turn transition in a normative way [6]. The standard model makes many specific

predictions—on the timing of gaps, the distribution of overlaps and the size of turns—which have been repeatedly confirmed in subsequent research, including quantitative studies (e.g. [1,5,15,16]). Moreover, in recent years, the model has been confirmed and expanded by experimental research showing that next speakers engage in considerable parallel processing by preparing a next turn while a current one is still underway [17–25], that they use lexical and syntactic information to project turn ends [26–29], and that turn final signals, including prosody, can mark and facilitate turn end detection ([17,26,30–32]; see also [3]).

The standard model of turn-taking consists of two components and a set of rules [6] which allocate turns to two or more parties. A turn-constructional component describes the units (e.g. words, phrases and sentences) out of which turns are built, so-called turn-constructional units (TCUs). A turn-allocation component describes methods for the allocation of turns to next speakers (e.g. by the speaker addressing or gazing at a particular interlocutor, or by allocating the next turn to the speaker who starts first). Also, a set of rules order the rights to speak, giving priority to a speaker specifically selected in the prior turn, or in the absence of that to any other participant speaking first, and finally as the last option allowing continuation by the last speaker. TCUs, then, are the crucial units over which the turn-taking rules operate, and they have been characterized in terms of linguistic units (e.g. words, phrases or clauses). Turns, however, may consist of more than one TCU, but where they are marked as complete, transition between speakers is relevant. Research on these transition-relevance places (TRPs) has found that they tend to occur where boundaries of syntactic units coincide with boundaries of prosodic units, for example, a complete clause produced as a complete intonational phase ([8,12,30,31,33–37]; but see [29]), though semantic and pragmatic completeness also plays a role [8,27,28,38].

The standard model of turn-taking is thus primarily a vocal-auditory one, working with units that have been defined in linguistic terms. However, face-to-face interaction is multimodal, involving diverse semiotic resources, including visible bodily actions [39–44]. Within the model, gaze and gesture have been shown to be major resources for turn-allocation (e.g. by addressing a turn with gaze or selecting a next speaker with a point; [10,11,45,46]), and some smaller scale studies have suggested that visual modalities may contribute to the recognition of turn completions [47–51]. The speed of transitions between turns has also been shown to be affected by visual bodily actions: questions with gaze directed to the addressee and questions with gestures get faster responses [1,52], and visual signals often precede the onset of verbal utterances or corresponding semantic elements within them (e.g. [53–57]), thus equipping them with 'predictive potential' [44,54]. Despite these observations, however, the standard model of turn-taking for conversation remains primarily a vocal-auditory one.

There is, however, an alternative model which is multimodal [58–60]. In this model, transitions are organized by multimodal turn-yielding and turn-holding cues. Turn-yielding cues include, for example, the completion of a syntactic clause, rising or falling intonation, the termination of a manual gesture, or a shift in head direction, each of which has an additive effect (i.e. the more turn-yielding cues are present, the more likely for a turn transition to occur). Turn-holding cues include manual gesturing which overrides the turn-yielding cues. Although the inclusion of visual cues is an attractive feature of the model, it has serious shortcomings [5]. It is an incomplete model of turn-taking behaviour and does not account, for example, for the construction of turns through units with points of possible completion that afford opportunities for transition. Further, it assumes that cues are context independent when in fact their meanings depend on their contexts of use (see [61] on rising intonation). Also, it has not been verified and elaborated by subsequent research, having been eclipsed by the standard model.

Another line of research has also investigated the relationship between gaze and turn-taking, beginning with an influential study by Kendon [62, p. 42]. Kendon observed that gazing away from the addressee signals an intention to hold the floor whereas gazing at the addressee yields the floor and signals an expectation of a response. These observations have been confirmed and expanded in subsequent research (e.g. [63–67]; but see Gambi *et al*. [68] for evidence to the contrary in experiments with virtual agents), but the claim that gaze regulates turn-taking has been disputed by conversation analysts. Rossano [69] has argued that gaze is not a resource for turn-taking *per se* but is rather a resource for formation of social actions (e.g. telling a story versus asking a question), which have different implications for turn-taking (see also [70,71]). In this account, gaze is related to what speakers do with their turn, not with turn-taking per se. This debate makes relevant further investigation, including of a systematic and quantitative nature.

(a) The present study

In this article, we promote an integrative approach to understanding turn-taking in face-to-face conversation by combining the fine-grained analysis of speaking turns in terms of TCUs and points of possible completion with analyses of visual signals (here, manual gestures and gaze). To assess the relative merits of verbal-only and multimodal approaches to turn-taking, we use a custom-made multimodal corpus and test to what extent the vocal-auditory channel alone can predict turn transitions, or to what extent and precisely how, multimodal signals may contribute to observable turn-taking. Specifically, we ask whether and how gaze direction and manual gestures at points of possible turn completion are involved in the coordination of transitions between speakers. Several predictions can be made on the basis of the previous observations about gaze and gesture just reviewed. First, at the possible completion of a turn, addressee-directed gaze, completed gestures and gestural retractions (i.e. the point where the meaningful part of a gesture is over and the hand moves back into rest position) should be associated with a higher frequency of transitions whereas averted gaze and the preparation or stroke (i.e. meaningful) phases of a gesture should inhibit transitions and thus be associated with a lower frequency of them. Second, in line with Duncan's model, such visible bodily actions may have an additive effect: the more visual cues, the stronger the signal. This should appear as an interaction between variables in statistical models. And third, the occurrence of addressee-directed gaze and manual gestures within TCUs may speed up transitions between speakers, resulting in shorter gaps between turns. This prediction emerges from independent evidence suggesting that address-directed gaze is associated with faster responses and that turns with gestures are responded

to faster than those without, but their combined effect has not been examined before.

To test these predictions, we combine conversation analytic and quantitative methods [72] to analyse turn-taking behaviour, including both gaze and gesture, in a rich corpus of multimodal interaction (see [73]). Our aim is not to assess the effects of visual signals on the cognitive processes involved in turn-taking [44,54,74] but is rather to consider the integration of visual signals into a model of turn-taking based primarily on verbal turns [5,6]. As such, we aim to answer the broader question of whether the coordination of turn transitions in face-to-face interaction is a multimodal process and thus whether the TRPs within turns should be defined as multimodal constructs.

2. Methods

(a) Corpus and participants

The present analyses are based on the Eye-tracking in Multimodal Interaction Corpus (EMIC) which consists of casual conversations between already acquainted native speakers of English in groups of two and three [73]. Each recording session lasted approximately 40 min with a 20 min triadic conversation followed by a 20 min dyadic one. The participants sat in chairs oriented towards each other while they wore eye-tracking glasses (SMI) and head-mounted microphones (Shure) and were recorded by three synched video cameras (Canon Legria). All video, sound and eye-tracking signals were synchronized. For the analyses reported here, 10 of the dyadic conversations (out of 27) were selected (this was a random selection based on those interactions with satisfactory quality of the eye-tracking data, to make the amount of coding feasible). Of these, four were all female, three were all male, and three were mixed female–male dyads (age range: 19–68 years). For further details on laboratory set-up and equipment, see Holler & Kendrick [73].

(b) Annotations and analysis

(i) Segments of conversation

For each dyadic conversation, we selected a segment of approximately 5 min. The segment began at the first initiation of a sequence [75] that did not relate to the laboratory set-up or equipment (e.g. the first new topic after the researchers had left the room) and ended at the completion of the last TCU (see below) in the 5 min interval.

(ii) Points of possible completion

The audio recordings, including the high-fidelity recordings from the head-mounted microphones, were used to transcribe the segments without reference to the videos, and according to standard CA conventions [76,77]. The written transcripts together with the audio recordings were then used to identify points of possible completion in the turns at talk [6], which were manually annotated in the transcripts, again without reference to the videos. These are points where speaker transition could normatively occur without a sense of interruption. The identification of points of possible completion was a holistic process that involved careful attention to grammar, lexicon and other aspects of the talk's production, including its prosodic design. It was based on observational and experimental research in CA that has shown them to occur where boundaries of syntactic, prosodic and pragmatic units coincide (see [8,16,30,31,34,78,79]). The holistic process of identifying points of possible completion is a standard procedure in CA; moreover, earlier research which identified points of possible completion in this same manner was able to corroborate the validity of this approach with quantitative eye-tracking data [73]. For example, a point of possible completion would normally occur at the end of a complete interrogative clause that is produced as a complete intonational phrase and performs a recognizable social action [80] such as requesting information. Extract 1 shows where points of possible completion, represented in the transcripts by vertical bars, were identified in a short segment of conversation.

Extract 1 [EMIC_02d_00:14:15]

Conventions: A, B denote participants; (.) *indicates a very short pause;* (*0.3*) *a gap of c. 300 ms; : a lengthened sound; =a rapid juncture;* [*the alignment of overlap;* | *a point of possible completion.*

```
1  B:  I was wondering y- e- Rebecca, has
2      she- (.) how many children has she
3      go[t?|
4  A:    [two.|
5      (0.3)
6  B:  two,|
7      (0.4)
8  A   yeah.|=two daughters.|
9      (0.4)
10 B:  oka:y.|=and how old are they now?|
11     =th[e:n.|
12 A:     [.hhhhhh
13     e:h well the-you see they're quite
14     you:ng.|
```

The identification of points of possible completion in effect segments the talk into units, namely potential turns. The units out of which turns are constructed are, as mentioned, TCUs [6]. While most turns in extract 1 contain only one TCU, some contain multiples (lines 7 and 9). A distinction is also made between a TCU that is potentially complete and a following turn component, an 'increment', that continues the grammatical structure of a TCU past a point of possible completion [81,82]. In extract 1, 'the:n.' (line 9) is an increment because it occurs after the possible completion of the prior TCU and continues its grammatical structure. Both TCUs and increments end in points of possible completion. To assess the reliability of the identification of points of possible completion by the first coder (K.H.K.), a second coder (J.H.) annotated a 20% sample of each conversation, also based on the audio without reference to the videos, resulting in 83% agreement.

The annotations of points of possible completion in the transcripts were then used to segment the talk of each participant in Praat (5.3.82; [83]). The Praat textgrids were then imported into ELAN (4.61; [84]), resulting in a tier beneath the transcript for each speaker in which the talk was segmented into TCUs and increments. The end of each annotation thus corresponded to a point of possible completion in the speaker's turn (figure 1).

(iii) Turn transitions

For each point of possible completion, we coded whether a transition between speakers occurred. First, transitions were identified automatically by a script if the next speaker began a TCU within a temporal window of −500 and +1250 ms of a point of possible completion in the current speaker's turn. This window was defined inductively through close examination of turn transitions in the conversations. A current speaker can continue their turn past the first point of possible completion (see line 7 in extract 1) even as a next speaker takes a turn, resulting in overlap between current and next speakers [6]. Transitions were identified irrespective of whether the current speaker continued their turn in overlap. Transitions from full TCUs to vocal continuers (e.g. *uh-uh*, *mh-mm*, [85,86]) were included ($n = 84$) because practices that elicit them (e.g. prosodic completion) can also indicate the completion of the TCU. Transitions from vocal continuers to full TCUs were not included because current speakers do not orient to the completion

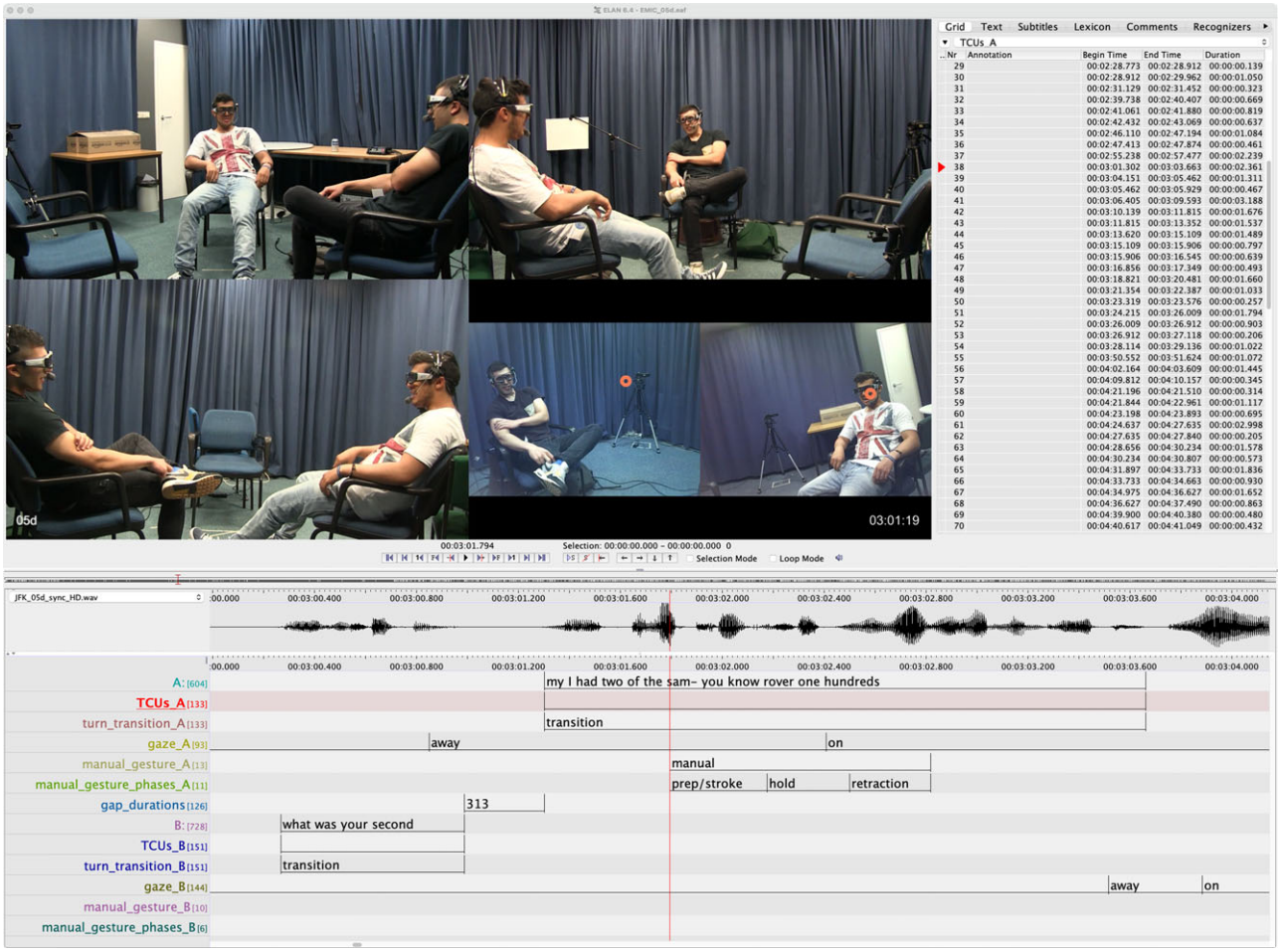

Figure 1. Multimodal annotations in the video analysis application ELAN. There are six tiers for each participant: transcription, TCUs, transitions, gaze, gesture and gesture phase. (Online version in colour.)

of a vocal continuer as relevant for the timing of their next TCU (e.g. they may freely overlap with them). The results of the automatic process were then used to generate annotations in ELAN. Second, because the automatic identification of transitions produced errors (i.e. invalid annotations of transitions between speakers), the annotations were manually corrected to align the measurement process with previous research on turn transitions [6,87–90]. Collaborative completions (i.e. when a next speaker produces a possible completion of the current speaker's in-progress TCU; [14,91]), interjacent overlaps (i.e. when a next speaker starts up in overlap in the middle of the current speaker's in-progress TCU without reference to its possible completion; [89,92]) and progressional overlaps (i.e. when a next speaker starts at a hitch or pause in the current speaker's in-progress TCU; [7]) were not coded as transitions for our purposes because such transitions occur before, and without reference to, points of possible completion and because their organization does not involve the practices by which turns come to recognizable completion.

(iv) Transition speed

To determine the speed of the transitions, we measured the duration of the interval in milliseconds between the end of the current speaker's TCU or increment and the first word or particle of the next speaker's TCU; inbreaths were included within the interval between turns (see offset2 as defined by [93, p. 12]). Negative values represent transitions that began before the point of possible completion (i.e. in overlap), whereas positive values represent transitions that began after such a point. While most positive durations correspond to intervals of silence between turns (i.e. gaps), some represent intervals in which both current and next speaker speak at the same time after the possible completion of the current turn. This occurred when a current turn reached a possible completion, at which point a next speaker began a turn, but at the same time the current speaker continued their turn with an additional unit, resulting in overlap. In such cases, we measured the interval from the possible completion of the first turn to the first word or particle of the next turn despite the overlap.

(v) Gaze direction

In order to annotate the participants' eye movements, the eye-movement fixations mapped onto a recording of the individual participant's visual field were synchronized to the ELAN files and the other high definition (HD) video recordings (see [73], for information on the output of the eye-trackers and synchronization procedure). The annotations were done manually, on a frame-by-frame basis. For each frame in the 5 min segments of conversation, the gaze fixation point generated by the SMI software for the participants was categorized as being: (i) on the other speaker, (ii) on self (e.g. when looking at their own hands), (iii) on the surroundings (e.g. the walls, the door, any equipment items in the room), or (iv) not identifiable from the eye-tracker data (i.e. the gaze fixation point was not visible in the respective video frames). This four-way distinction was then simplified to a two-way distinction between on or away from the other speaker, collapsing the distinction between (ii) and (iii). Frames categorized as (iv) were coded by visual inspection of the HD videos. If a participant's eyes were closed, their gaze was categorized as away. Inter-coder reliabilities were not established for the coding of gaze, since the large majority of the data was based on automatically generated indications of gaze direction through the eye-tracking software. For instances in which missing data were added based on information from the

HD videos, the values were added by one of the coders (L.D.) and checked by one of the lead researchers (K.H.K. or J.H.).

(vi) Gestures

For each 5 min segment, the occurrence of manual gestures was annotated. Manual gestures were defined as meaningful movements of the hands and torso that speakers produced as part of their social actions; this included iconic and metaphoric gestures [94] which imagistically depict the properties of concrete objects (e.g. 'ball' by holding an imaginary round object between the hands) or actions (e.g. 'a car driving past' with a hand moving fast from left to right) and abstract concepts (e.g. 'rising inflation' with a hand moving upwards); deictic gestures, such as pointing to absent or present entities [94], interactive gestures [95,96] which regulate the interaction between interlocutors, e.g. by handing over a turn with the palm facing up, hand open and fingertips towards the interlocutor, and pragmatic gestures [39] which add pragmatic meaning, e.g. by expressing negation or emphasizing parts of speech with a downward movement. An independent coder (G.C.), blind to hypotheses, identified all gestures meeting the above criteria (as one inclusive category labelled 'gesture'). Reliability was established with a second coder, also blind to hypotheses (Y.v.d.H.), based on 30% of the video data (randomly chosen) containing 36% ($n = 253$) of the manual gestures in our dataset ($n = 710$). This yielded a reliability of 78% for gesture identification indicating a high degree of agreement. Based on this, we coded whether each TCU and increment occurred with a manual gesture or not (meaning that any part of the gestural movement co-occurred with the TCU, with the minimum criterion being one frame overlap between gestural movement and TCU).

The temporal organization of the manual gestures was then examined at each point of possible completion and at each transition between speakers. Gestures often consist of phases: preparation (the hand moving into gesture space where the gesture is to be performed), stroke (the most meaning-bearing part of the movement), post-stroke hold (the hand is held still after the stroke has been performed), and retraction (the hand withdraws from the spatial location where the gesture was performed, often moving back into rest position) [39,97]. For TCUs and increments without a transition, we identified the movement phase at the point of possible completion (i.e. at the end of the unit). However, because transitions between speakers can occur well before (resulting in overlaps) or well after the end of the unit (resulting in gaps), the gesture phase produced at the point of possible completion may not in fact be the most relevant phase for the coordination of the transition. For this reason, for TCUs and increments that did have a transition, we measured the movement phase of the current speaker's gesture at the point at which the transition began (i.e. at the beginning of the next speaker's turn which may precede or follow the end of the current speaker's TCU). For each TCU and increment we thus categorized the temporal organization of the current speaker's gestures as being in (i) a preparation or stroke phase, (ii) a post-stroke hold phase, or (iii) a retraction phase, or as being (iv) already completed. We decided to collapse preparation and stroke phase into one category because the hand often already adopts a shape or posture during the preparation phase which can convey meaning [98], and also because, in any case, both preparation and stroke signal that the speaker is in the process of encoding meaning gesturally and not yet done. TCUs and increments that did not include a manual gesture were categorized as (v) no manual gesture. The inter-coder reliability for gesture phases was established between the main coder (K.H.K.) and an independent coder blind to the TCU boundaries (L.v.O.); for gesture phase categorization at TCUs, agreement was 87%, with a weighted Cohen's kappa value of 0.806, indicating substantial agreement [99]; gesture phase categorization at the point of transition was 85%, with a weighted Cohen's kappa of 0.931, indicating almost perfect agreement. In the examples presented in this article, gaze and gestures are transcribed using conventions developed by Mondada [100].

(vii) Statistical analyses

We fitted linear mixed-effects models to our data using the lme4 (1.1–28) package [101] in R (4.1.2; [102]). The main analysis of turn transitions is based on a binomial generalized linear mixed-effects model with turn transition as the dependent variable, gaze direction and gesture temporal organization as fixed predictors (reference levels = 'gaze on' and 'no gesture') along with the interaction between the two, and speaker as a random factor. The model did not include conversation as a random factor, nor random slopes, as these resulted in singular fits and were thus removed. The gesture temporal organization variable included five levels: no gesture, completed gesture (i.e. a gesture whose retraction was complete before the end of the TCU), preparation/stroke, post-stroke hold and retraction. We also tested for the effect of gesture presence on frequency of turn transition where we operationalized the predictor as a binary variable, with the levels gesture/no gesture. The main analysis of transition speed is based on a linear mixed-effects model with gap duration as the dependent variable, gaze direction and the presence or absence of a manual gesture during the TCU (reference levels = 'gaze on' and 'no gesture') as fixed predictors along with the interaction between the two, and speakers nested within conversations as random factors. The model did not include random slopes as these resulted in singular fits and were thus removed. However, since checks of the distribution of residuals did not meet the lmer assumption of normality, we ran the model using the rlmer function (robustlmm package, [103]).

3. Results

The sample included in total 2131 points of possible turn completion. On average in each 5 min segment of conversation a participant produced 107 TCUs (range: 57–149). Transitions between speakers occurred in a large minority of cases (42%, $n = 895$). The relatively low frequency of transitions at these points seems to reflect a relatively high frequency of short units (e.g. lexical TCUs such as 'yeah' and 'okay' in turn-initial position followed by other material, and single-word increments (after completion points) and in addition many multi-unit turns, that is, turns constructed so as to contain multiple TCUs (e.g. stories or explanations).

(a) Gaze direction and turn transitions

The speaker's gaze was directed to the addressee at most points of possible completion (61%, $n = 1298$). The proportion of actual transitions was significantly higher for units with gaze directed to the addressee compared to those with gaze directed away from the addressee (48%, $n = 620$ versus 33%, $n = 275$; $\beta = -0.65$, s.e. = 0.1, $z = -6.75$, $p < 0.001$). Because the proportion of transitions when a speaker's gaze was directed to the addressee was close to 50% (figure 2), the results suggest that addressee-directed gaze alone would not be a strong signal for transition relevance. Conversely, the relatively low proportion of transitions when a speaker's gaze was directed away from the addressee indicates that gaze aversion may project or predict turn continuation and thus inhibit speaker transition—this then may be a more reliable cue for turn-taking than gaze directed at the addressee.

The association between gaze aversion and turn continuation can be observed in extract 2. Here A tells B about her

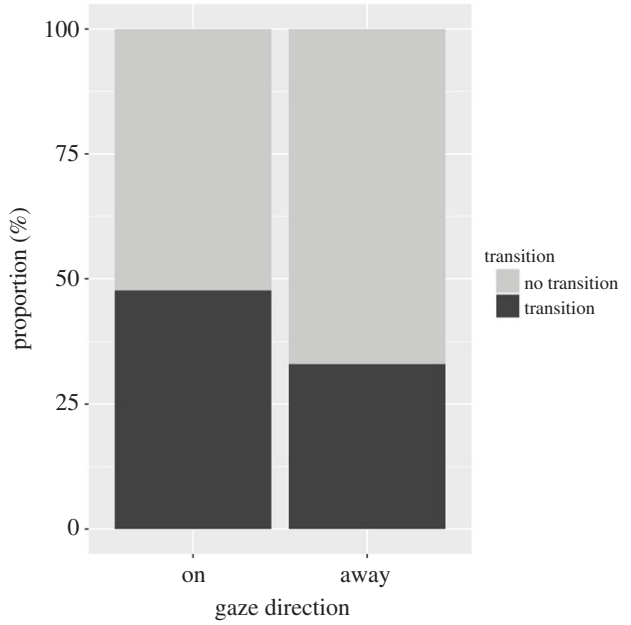

Figure 2. Proportion of turn transitions when the speaker's gaze is on or away from the addressee at points of possible completion.

difficulties in finding a suitable date to celebrate her daughter Lin's birthday.

Extract 2 [EMIC_02_01:03]

Conventions: A, B denote participants; underlining indicates prosodic prominence; .m a bilabial click; .hh an in-breath, (0.3) a pause of c. 300 ms; : a lengthened sound; [*the alignment of overlap;* | *a point of possible completion; * represents the beginning of an embodied action; –> that the action continues.*

```
1  A:   .m .hh anyway John is coming
        >>gazes away-->
2       nex*t week,| so *uh [when he comes=
3  B:                       [o:::::h.
   a    -->*to B--------*gazes away-->
4  A:   =uhm: (0.3) .hhh ehh but they go to
5       Berlin.| eh (0.2) Lin and Alex go
6       to Berlin,| so John: (0.2) won't
7       s:: will only see her on one day,|
8       so that can't be the birthda*y.| so
                                  -->*B-->
9       I don't know what we're going
10 A:   *to do. [hhhhhhh
11 B:           [aw::
   a    *away-->>
```

As the first TCU—an informing about a visit from A's son—comes to a possible completion, A directs her gaze to B, which elicits an prosodically marked change of state token [104] designed as both a receipt of the information and an appreciation of the good news. Before B's response, however, A has already extended her turn with an additional TCU and averted her gaze (lines 2–3). The gaze aversion coincides with the launch of a series of TCUs in which A details the difficulty; at the possible completion of each (lines 4–6), she gazes away from B and continues her turn. No transitions occur. When A then returns her gaze to B at the end of the penultimate TCU and across the bulk of the next as she formulates the trouble explicitly (lines 7–8), B responds with a display of sympathy ('aw::' at line 9), though at this point A's gaze is elsewhere.

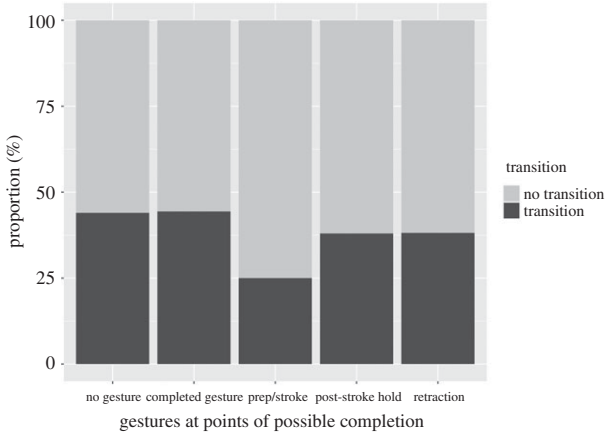

Figure 3. The proportion of turn transitions by the temporal organization of manual gestures.

Sequences such as this show that when a speaker's verbal turn comes to a point of possible completion, gaze aversion can project or predict a continuation of the turn and thereby suppress the relevance of transition.

(b) Manual gestures and turn transitions

The entire dataset contained 710 manual gestures. Out of the 2131 TCUs, 28% ($n = 588$) included at least one manual gesture, the presence of which had a significant effect on turn transitions: the proportion of transitions was significantly lower for TCUs with manual gestures (37%, $n = 217$) than for those without manual gestures (63%, $n = 371$; $\beta = -0.25$, s.e. = 0.1, $z = -2.31$, $p = 0.02$).

The majority of manual gestures spanned points of possible completion, i.e. there was overlap between a portion of the gesture and the TCU end (68%, $n = 397$). Figure 3 shows the proportion of turn transitions across five categories of the temporal organization of the gestures.

The temporal organization of the gestures was a significant predictor of turn transitions ($\beta = -0.81$, s.e. = 0.2, $z = -4.15$, $p < = 0.001$). As can be seen from figure 3, compared to turn transition when TCU ends co-occurred with no manual gesture (44%, $n = 679$), when a manual gesture was completed before the end of the TCU (44%, $n = 91$), when the transition occurred at a post-stroke hold (38%, $n = 49$) or with a retraction (38%, $n = 37$), the proportion of turn transitions was considerably lower when a speaker produced a preparation or stroke phase of a manual gesture (25%, $n = 39$) coinciding with the TCU end. The results suggest that the production of a preparation or stroke phase of a manual gesture at a possible completion point may suppress the relevance of transition between speakers. No evidence was found that the other movement phases, including retractions, were associated with a higher or lower proportion of transitions.

At the possible completion of a TCU, the production of a preparation or stroke phase of a gesture can visually project that more is to come or underway, that is, that a subsequent TCU is imminent and thus that the speaker's turn will continue. Consider, for example, the gestures produced by C in extract 3 as she talks about the researcher who ran the study (the 'she' on line 1) in which participants had to film themselves as they learned how to juggle.

Extract 3 [EMIC_03d_01:30:02]

Conventions: B, C denote participants; underlining indicates prosodic prominence; .hhh an in-breath; (0.3) a pause of c. 300 ms; - a cut off syllable; : a lengthened sound; ° quiet speech; [*the alignment of overlap;* | *a point of possible completion;* # *the point at which the figure occurs; * represents the beginning of an embodied action; –> that the action continues.*

```
1  C:   but I wonder if she em- yea:h (0.3)
2       ever checked ju[st *tuh (.)
3  B:                   [just a snippet.|
   c                        *click-->
4  C:   .hhh *just t'watch someone
          -->*click-->
5       film* themselves juggling for two
         -->*juggle-->
6       minutes,#| and then* just be like
                          -->*enactment-->
   fig           #fig4a
7       ↓yeah can't be bothered.#|°=no:w.|
8       =and sat* down.°|
              -->*
   fig                              #fig4b
9       (.)
10 B:   heh .hhhh or maybe she got
11      enthra::lled.
```

In the course of her first TCU, C produces a series of iconic gestures: she begins to visually depict a clicking action with the index finger of her right hand, as though turning on a camera (lines 2–3), but abandons and retracts this as B speaks in overlap, recycling part of her TCU ('just to'; [9]; cf. [105]). After the overlap has been resolved, she redoes the gesture in full (the second 'click' under line 2), immediately after which she produces an iconic gesture that depicts juggling balls in the air (line 4). As she repeats the stroke phase of this juggling gesture, her TCU comes to a point of possible completion after 'two minutes' (line 5). Figure 4*a* shows the last frame of the TCU; we can see the stroke of the gesture is still underway. C then continues her turn with an additional TCU that formulates (non-serious) direct reported speech by the researcher ('and then just be like ↓yeah can't be bothered'; lines 5–6), during which C performs a complex bodily enactment [106]: she turns her head to the side, sways her body and extends her right elbow as she grips the arms of the chair, movements that embody a lackadaisical stance that complements the reported speech. While the enactment is in progress, her turn reaches possible completion after 'can't be bothered' (line 6; figure 4*b*) and again after the increment 'no:w' (line 6). Only after the enactment is complete and yet another possible completion is reached (after 'sat down' in line 6) does a transition between speakers occur.

Sequences like this show that when a speaker's verbal turn comes to a point of possible completion, gestures that are visibly incomplete because they are in a preparation or stroke phase can interdict the relevance of transition.

(c) Gaze, gestures and turn transitions

In addition to establishing their individual effects, we set out to test the joint effect of the speaker's gaze direction together with manual gestures at points of possible completion: how do these affect the occurrence of transitions between speakers when they are both entered as predictors? The results suggest that their individual significance is retained (gaze aversion: $\beta = -0.73$, s.e. = 0.11, $z = -6.59$, $p < 0.001$; preparation/stroke: $\beta = -1.2$, s.e. = 0.26, $z = -4.46$, $p < 0.001$). The direction of both effects was the same: at points of possible turn completion, the aversion of gaze and the preparation or stroke phase of a gesture were associated with a decrease in the frequency of turn transitions. However, contrary to our predictions, there was no evidence of an additive effect, which would have appeared as an interaction between predictors (see the electronic supplementary material, figure S1).

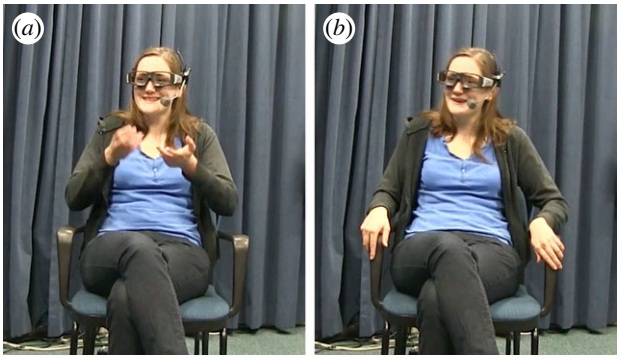

Figure 4. Gestural strokes at points of possible turn completion. (Online version in colour.)

(d) Gaze direction and transition speed

In most cases, the speaker's gaze was directed to the addressee at the possible completion of the turn (69%, $n = 620$). We predicted that transitions would be faster in such cases but found no evidence of this. Transitions for TCUs with addressee-directed gaze (mean: 246 ms, median: 167, mode: 85 ms) were not significantly faster than those for TCUs without addressee-directed gaze (mean: 245, median: 177, mode: 100 ms; $\beta = 41.92$, s.e. = 37.61, $t = -1.12$, $p = 0.27$; see the electronic supplementary material, figure S2). The results suggest that while gaze direction affects whether a transition occurs, it does not influence its speed.

(e) Manual gestures and transition speed

We also predicted that transitions should be faster when a TCU includes a manual gesture. This prediction was confirmed. Transitions for TCUs with manual gestures were faster (mean: 132 ms, median: 66 ms, mode: 25 ms) than those for TCUs without manual gestures (mean: 283 ms, median: 200 ms, mode 115 ms; $\beta = -124.88$, s.e. = 41.19, $t = -3.03$, $p < 0.003$; see the electronic supplementary material, table S1 for descriptive statistics for each gesture phase). Figure 5 shows the distribution of gap durations for TCUs with and without gestures. For those with a gesture, the peak of the distribution, which represents an estimate of the mode, occurs approximately 90 ms earlier and a smaller proportion occurs after long gaps (i.e. after approximately 400 ms).

The results suggest that while the direction of a speaker's gaze does not affect the speed of transitions between speakers, the production of manual gestures does: TCUs with gestures have faster transitions.

(f) Gaze, gesture and transition speed

As for the likelihood of transitions occurring, we were interested in the combined effects of gaze and gesture at points of possible completion on the speed of turn transitions. The results are the same as those of the models including them

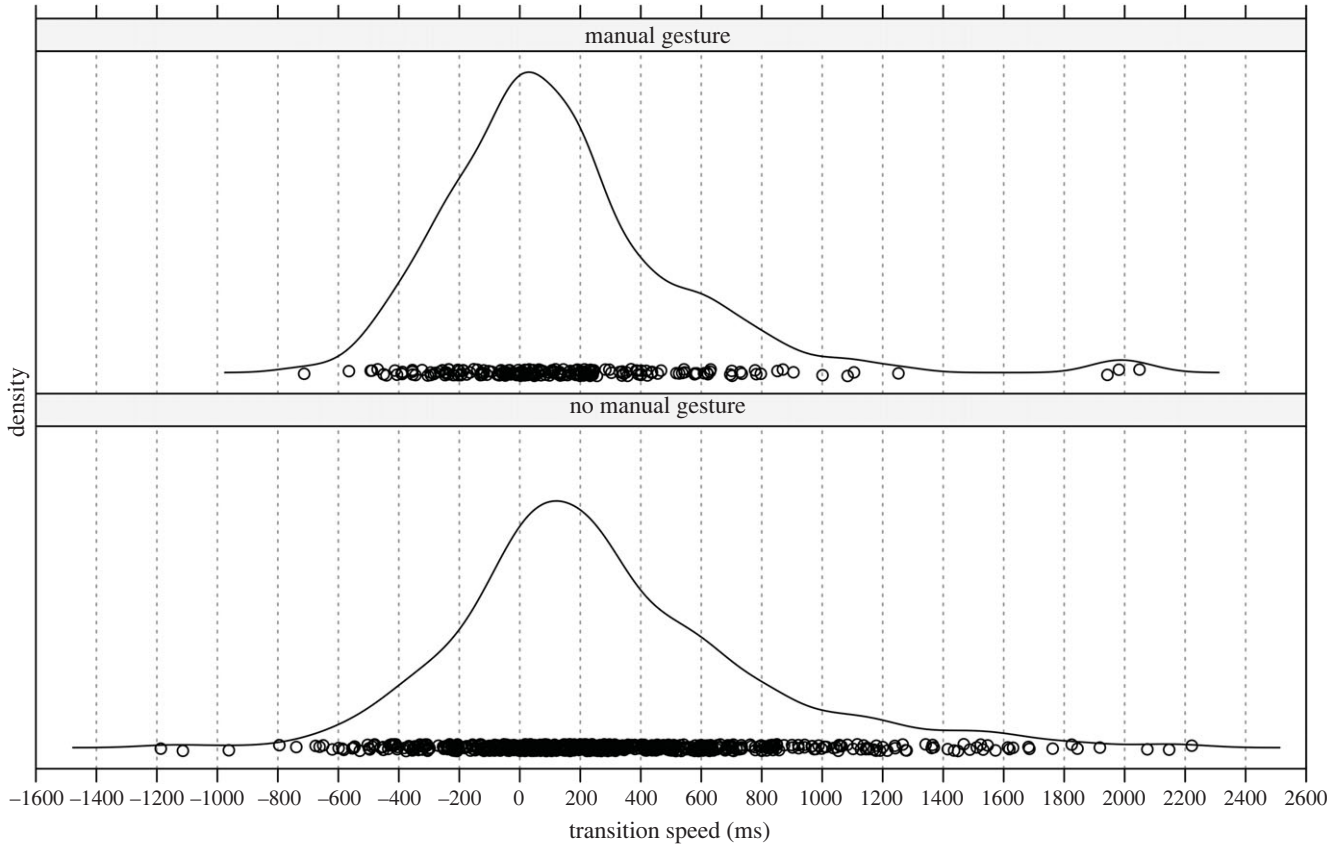

Figure 5. Gap durations for TUCs with and without manual gestures.

as individual predictors: while gaze direction did not have statistically significant effect ($\beta = 57.55$, s.e. = 35.08, $t = 1.64$) the presence of manual gesture did ($\beta = -86.88$, s.e.= 42.42, $t = -2.15$). There was no statistically significant interaction between the predictors (see the electronic supplementary material, figure S3).

4. Discussion

The standard model of turn-taking describes the close coordination of transitions between speakers at points of possible completion. The initial articulation of the model by Sacks *et al*. [6] did not, however, specify precisely which resources allowed for the anticipation and recognition of such points. Subsequent research focused predominantly on linguistic resources, especially the convergence of syntactic and prosodic units (e.g. [8,29–31,34,78]). As a consequence, despite decades of incremental development, the model has remained primarily an auditory-vocal one. Our results suggest, however, that the coordination of transitions between speakers involves not only linguistic resources but also visual gestural ones and thus that the standard model can and should be reconceptualized as multimodal.

We found a significant relationship between the direction of a speaker's gaze, their use of manual gestures and transitions between speakers. The results suggest that when a speaker's verbal turn comes to a point of possible completion, the aversion of a speaker's gaze or the production of a gesture can project a continuation of the turn and thereby suppress the relevance of transition, leading to fewer turn transitions in their presence. This confirms and amplifies previous observations in the literature and integrates them into the standard model. The findings are consistent with Kendon's [62] observation that gazing away from the addressee signals an intention to hold the floor. We did not, however, find strong evidence of an association between addressee-directed gaze and transitions [62,63,70]. Transitions occurred at points of possible completion with addressee-directed gaze around half the time. One explanation for this is that addressee-directed gaze is a less reliable cue, since it also has other functions, such as monitoring the addressee's state of understanding, attention and so forth (e.g. [62,63,107,108]).

With respect to the model of turn-taking proposed by Duncan *et al*. [58–60], the results are mixed. The low proportion of transitions when speakers gesture across points of possible completion is compatible with Duncan's observations, as well as those by Streeck & Hartge [47] and Zellers *et al*. [51]. However, we found no relationship between the completion or retraction of a gesture at a point of possible completion and turn transitions, as predicted by the model (but see [50] on Swedish and German conversations). The retraction or completion of a gesture thus do not appear to be turn-yielding cues, at least in the English conversations we analysed. We also found no evidence of an additive effect. The Duncan model predicts that gaze aversion together with the production of a gesture at a point of possible completion should have a greater effect than either practice on its own, but this prediction was not confirmed (see the electronic supplementary material, figure S1). Rather, even when speakers employ practices that suppress transition relevance, transitions still occur albeit at a low rate, which perhaps reflects a next speaker's right to self-select at a point of possible completion despite a current speaker's displayed intention to continue [6].

The design of the study does not yet allow us to disentangle the relative contributions of auditory and visual signals in

the coordination of turn transitions. The identification of points of possible completion involved close attention to the linguistic design of the turns at talk, including their prosodic production. What we have shown are robust associations between such points and visible bodily actions. It is possible that auditory and visual resources go hand in hand as multimodal gestalts that together signal turn completion [44]. Such multimodal gestalts may include a complementary relationship between the auditory and visual resources, with one modality being more salient, when the other is less so. Further conversation analytic and experimental research is needed to tease apart the relative contributions and detailed interplay of the different resources during the management of turn transitions.

The multimodal organization of the turn-taking is also evident in the timing of transitions between speakers. We found that TCUs which have gestural components are responded to faster than those that do not. This expands the scope of our previous observation, based on a different set of conversations in the same corpus, that questions with gestures get faster responses [52]. We now see the effect in a larger and more diverse sample, and one that is not limited to question-response sequences (895 TCU-TCU transitions versus 281 question-response transitions). The effect is, however, attenuated: while gestures in questions sped up turn transitions by approximately 200 ms, gestures across all TCU types in the present study did so by around 90 ms. This difference may be owing to questions in general being responded to faster than turns that do not require an answer, or owing to the increased competition in triadic versus dyadic conversations (see [109]). The exact mechanism that leads to TCUs with gestures getting faster turn transitions than turns without gestures is not yet known [52]. It is possible that gestures, which often precede associated turn components, may facilitate a next speaker's ability to predict in advance how a TCU will unfold, which in turn could enable a faster response [44,54]. It is also possible that gestures afford an advantage in the recognition of the TCU's action [44,110], or that the precise temporal organization of a gesture's production, such as the onset of its retraction, provides early cues for turn completion that influence when a transition occurs in English conversation [52]. The observation of a facilitation effect across a large and diverse sample of transitions further underscores the need for additional observational studies of the multimodal organization of turn transitions as well as experimental research on candidate psycholinguistic mechanisms.

The direction of a speaker's gaze did not, however, affect the speed of turn transitions. Our prediction here was based on a study of question-response sequences which showed that questions with addressee-directed gaze received faster responses [1]. Our results suggest that the facilitation effect of gaze may be unique to questions and does not generalize to all TCU types. Questions tend to be first pair-parts [75] which together with gaze direction or some other method of addressing a turn at talk constitute a next-speaker selection technique [6,10]. The faster transitions after questions with addressee-directed gaze may thus be a byproduct of the use of a next-speaker selection technique. If so, we would not expect the effect to generalize to all TCUs as not all TCUs employ such techniques. We would expect, however, to observe the effect in question-response sequences in the present sample, a prediction we plan to test in future research.

Taken together, the present findings suggest that adaptations to the standard model of turn-taking are required. We propose that such an adaptation should consist of an integrated model which has the fine-grained linguistic design of verbal utterances at its heart (TCUs), together with the visual signals that form part of them, including minimally gaze and manual gestures. The current observations could be accommodated by including multimodal signals in the percepts involved in the recognition of TRPs in face-to-face interaction; specifically gaze aversion and ongoing gestures seem to 'overwrite' linguistic signals of completion. The results thus suggest a further elaboration of the concept of a TRP, first introduced by Sacks *et al*. [6] and subsequently specified by Ford & Thompson [8]. On the basis of the present observations, we thus propose to extend the concept of a complex TRP to a multimodal TRP, or MTRP for short.

(a) Limitations and future avenues of research

The results allow us to conclude that the intricate mechanisms of turn-taking in face-to-face interaction are multimodal in nature. We have, however, examined just two kinds of visual signals, manual gestures, including their phases, and gaze direction. There are many other potentially potent signals, such as facial expressions and body position or torque [111], which should also be examined. Moreover, we have looked at TCUs as one broad group, but considering the various actions they perform [80] and the sequential positions in which they occur [75] may further refine the picture (see [69]). Also, the present analyses are based on dyadic conversations, which arguably capture one of the most common forms of interaction. However, since the dynamics of turn-taking can differ between dyadic and multiparty interactions [109], expanding the present analyses to conversations involving more than two participants is an important next step. Finally, although the present study was not designed to assess the influence of gaze and gesture on cognitive processes during turn-taking, it nonetheless provides an empirically grounded basis for generating experimental hypotheses to investigate such influences. Future studies may determine, for example, whether there is a causal relationship between the perception of specific gesture phases and the recognition of turn completions, or between the presence of gestures and the speed of turn transitions, and if so, what cognitive mechanisms underpin these effects.

5. Conclusion

The science of social interaction benefits from over 40 years of cumulative empirical research in the field of CA. Numerous basic phenomena—from TCUs and increments to first pair-parts and next-speaker selection techniques—have been identified and carefully described, and precise models of interactional systems have been developed and incrementally expanded. This wealth of naturalistic observational research, as we have shown here, can feed into controlled laboratory and experimental studies, which in turn further elaborate our models of interactional systems. The science of social interaction, in our view, thus involves a deliberate methodological pluralism that includes naturalistic observation and strategic interdisciplinary collaborations (see [72,112]).

Ethics. The study was approved by the Social Sciences Faculty Ethics Committee, Radboud University Nijmegen.

Data accessibility. The data are available as electronic supplementary material from the journal's website [113].

Authors' contributions. K.H.K.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, supervision, visualization, writing—original draft, writing—review and editing; J.H.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, supervision, validation, writing—review and editing; S.C.L.: conceptualization, funding acquisition, resources, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. We thank the European Research Council for financial support (advanced grant no. 269484 awarded to S.C.L.).

Acknowledgements. We would like to thank Georgia Carter, Yvonne van den Heuvel, Linda Drijvers and Leah van Oorschot for assistance with data coding/reliabilities. We would also like to thank Ludy Cilissen for help with synchronization of the audio and video datastreams.

References

1. Stivers T *et al.* 2009 Universals and cultural variation in turn-taking in conversation. *Proc. Natl Acad. Sci. USA* **106**, 10587. (doi:10.1073/pnas.0903616106)
2. Holler J, Kendrick KH, Casillas M, Levinson SC (eds) 2016 *Turn-taking in human communicative interaction*. Lausanne, Switzerland: Frontiers Media.
3. Levinson SC. 2016 Turn-taking in human communication – origins and implications for language processing. *Trends Cogn. Sci.* **20**, 6–14. (doi:10.1016/j.tics.2015.10.010)
4. Pika S, Wilkinson R, Kendrick KH, Vernes SC. 2018 Taking turns: bridging the gap between human and animal communication. *Proc. R. Soc. B* **285**, 20180598. (doi:10.1098/rspb.2018.0598)
5. Levinson SC, Torreira F. 2015 Timing in turn-taking and its implications for processing models of language. *Lang. Sci.* **6**, 731. (doi:10.3389/fpsyg.2015.00731)
6. Sacks H, Schegloff EA, Jefferson G. 1974 A simplest systematics for the organization of turn-taking for conversation. *Language* **50**, 696–735. (doi:10.2307/412243)
7. Jefferson G. 1984 Notes on some orderlinesses of overlap onset. In *Discourse analysis and natural rhetoric* (eds V D'Urso, P Leonardi), pp. 11–38. Padua, Italy: Cleup Editore.
8. Ford CE, Thompson SA. 1996 Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In *Interaction and grammar* (eds E Ochs, EA Schegloff, SA Thompson), pp. 134–184. Cambridge, UK: Cambridge University Press.
9. Schegloff EA. 2000 Overlapping talk and the organization of turn-taking for conversation. *Lang. Soc.* **29**, 1–63. (doi:10.1017/S0047404500001019)
10. Lerner GH. 2003 Selecting next speaker: the context-sensitive operation of a context-free organization. *Lang. Soc.* **32**, 177–201. (doi:10.1017/S004740450332202X)
11. Mondada L. 2007 Multimodal resources for turn-taking pointing and the emergence of possible next speakers. *Discourse Studies* **9**, 194–225. (doi:10.1177/1461445607075346)
12. Local J, Walker G. 2012 How phonetic features project more talk. *J. Int. Phonetic Assoc.* **42**, 255–280. (doi:10.1017/S0025100312000187)
13. Blythe J, Gardner R, Mushin I, Stirling L. 2018 Tools of engagement: selecting a next speaker in Australian aboriginal multiparty conversations. *Res. Lang. Soc. Interaction* **51**, 145–170. (doi:10.1080/08351813.2018.1449441)
14. Bolden GB, Hepburn A, Potter J. 2019 Subversive completions: turn-taking resources for commandeering the recipient's action in progress. *Res. Lang. Soc. Interaction* **52**, 144–158. (doi:10.1080/08351813.2019.1608096)
15. Wilson T, Zimmerman D. 1986 The structure of silence between turns in two-party conversation. *Discourse Process.* **9**, 375–390. (doi:10.1080/01638538609544649)
16. Robinson JD, Rühlemann C, Rodriguez DT. 2022 The bias toward single-unit turns in conversation. *Res. Lang. Soc. Interaction* **55**, 165–183. (doi:10.1080/08351813.2022.2067436)
17. Barthel M, Sauppe S, Levinson SC, Meyer AS. 2016 The timing of utterance planning in task-oriented dialogue: evidence from a novel list-completion paradigm. *Front. Psychol.* **7**, 1858. (doi:10.3389/fpsyg.2016.01858)
18. Barthel M, Meyer AS, Levinson SC. 2017 Next speakers plan their turn early and speak after turn-final 'go-signals'. *Front. Psychol.* **8**, 393. (doi:10.3389/fpsyg.2017.00393)
19. Barthel M, Levinson SC. 2020 Next speakers plan word forms in overlap with the incoming turn: evidence from gaze-contingent switch task performance. *Lang. Cogn. Neurosci.* **35**, 1183–1202. (doi:10.1080/23273798.2020.1716030)
20. Bögels S. 2020 Neural correlates of turn-taking in the wild: response planning starts early in free interviews. *Cognition* **203**, 104347. (doi:10.1016/j.cognition.2020.104347)
21. Bögels S, Magyari L, Levinson SC. 2015 Neural signatures of response planning occur midway through an incoming question in conversation. *Sci. Rep.* **5**, 12881. (doi:10.1038/srep12881)
22. Boiteau TW, Malone PS, Peters SA, Almor A. 2013 Interference between conversation and a concurrent visuomotor task. *J. Exp. Psychol.* **143**, 295. (doi:10.1037/a0031858)
23. Corps RE, Gambi C, Pickering MJ. 2018 Coordinating utterances during turn-taking: the role of prediction, response preparation, and articulation. *Discourse Process.* **55**, 230–240. (doi:10.1080/0163853X.2017.1330031)
24. Magyari L, De Ruiter JP, Levinson SC. 2017 Temporal preparation for speaking in question-answer sequences. *Front. Psychol.* **8**. (doi:10.3389/fpsyg.2017.00211)
25. Sjerps, Meyer 2015. (doi:10.1016/j.cognition.2014.10.008)
26. Lammertink I, Casillas M, Benders T, Post B, Fikkert P. 2015 Dutch and English toddlers' use of linguistic cues in predicting upcoming turn transitions. *Front. Psychol.* **6**, 274–291. (doi:10.3389/fpsyg.2015.00495)
27. Magyari L, de Ruiter JP. 2012 Prediction of turn-ends based on anticipation of upcoming words. *Front. Psychol.* **3**, 376. (doi:10.3389/fpsyg.2012.00376)
28. Riest C, Jorschick AB, de Ruiter JP. 2015 Anticipation in turn-taking: mechanisms and information sources. *Front. Psychol.* **6**, 89. (doi:10.3389/fpsyg.2015.00089)
29. De Ruiter JP, Mitterer H, Enfield NJ. 2006 Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language* **82**, 515–535. (doi:10.1353/lan.2006.0130)
30. Bögels S, Torreira F. 2015 Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *J. Phonetics* **52**, 46–57. (doi:10.1016/j.wocn.2015.04.004)
31. Bögels S, Torreira F. 2021 Turn-end estimation in conversational turn-taking: the roles of context and prosody. *Discourse Process.* **58**, 903–924. (doi:10.1080/0163853X.2021.1986664)
32. Gravano A, Hirschberg J. 2011 Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* **25**, 601–634. (doi:10.1016/j.csl.2010.10.003)
33. Local J, Kelly J, Wells WH. 1986 Towards a phonology of conversation: turn-taking in Tyneside English. *J. Ling.* **22**, 411–437. (doi:10.1017/S0022226700010859)
34. Schegloff EA. 1998 Reflections on studying prosody in talk-in-interaction. *Lang. Speech* **41**, 235–263. (doi:10.1177/002383099804100402)
35. Selting M. 1996 On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics* **6**, 371–388.
36. Caspers J. 2003 Local speech melody as a limiting factor in the turn-taking system (in Dutch). *J. Phonetics* **31**, 251–276. (doi:10.1016/S0095-4470(03)00007-X)
37. Walker G. 2013 Phonetics and prosody in conversation. In *The handbook of conversation*

analysis (eds J Sidnell, T Stivers), pp. 455–474. New York, NY: John Wiley & Sons, Ltd.

38. Selting M. 2000 The construction of units in conversational talk. *Lang. Soc.* **29**, 477–517. (doi:10.1017/S0047404500004012)
39. Kendon A. 2004 *Gesture: visible action as utterance*. Cambridge, UK: Cambridge University Press.
40. Streeck J, Goodwin C, LaBaron C (eds) 2011 *Embodied interaction: language and body in the material world*. Cambridge, UK: Cambridge University Press.
41. Deppermann A. 2013 Multimodal interaction from a conversation analytic perspective. *J. Pragmatics* **46**, 1–7. (doi:10.1016/j.pragma.2012.11.014)
42. Mondada L. 2016 Challenges of multimodality: language and the body in social interaction. *J. Socioling.* **20**, 336–366. (doi:10.1111/josl.1_12177)
43. Perniss P. 2018 Why we should study multimodal language. *Front. Psychol.* **9**, 1109. (doi:10.3389/fpsyg.2018.01109)
44. Holler J, Levinson SC. 2019 Multimodal language processing in human communication. *Trends Cogn. Sci.* **23**, 639–652. (doi:10.1016/j.tics.2019.05.006)
45. Heath C, Mondada L. 2019 Transparency and embodied action: turn organization and fairness in complex institutional environments. *Soc. Psychol. Q.* **82**, 274–302. (doi:10.1177/0190272519843303)
46. Auer P. 2021 Gaze selects the next speaker in answers to questions pronominally addressed to more than one co-participant. *Interactional Ling.* **1**, 154–182. (doi:10.1075/il.21002.aue)
47. Streeck J, Hartge U. 1992 Previews: gestures at the transition place. In *The contextualization of language* (eds P Auer, A Di Luzio), pp. 135–157. Amsterdam, The Netherlands: John Benjamins Publishing Company.
48. Ford CE, Fox BA, Thompson SA. 1996 Practices in the construction of turns: The 'TCU' revisited. *Pragmatics* **6**, 427–454.
49. Fox B. 1999 Directions in research: language and the body. *Res. Lang. Soc. Interaction* **32**, 51–59. (doi:10.1080/08351813.1999.9683607)
50. Zellers M, House D, Alexandersson S. 2016 *Prosody and hand gesture at turn boundaries (in Swedish)*. In *Proc. Speech Prosody, 31 May-3 June 2016, Boston, MA*, pp. 831–835. (doi:10.21437/SpeechProsody.2016-170)
51. Zellers M, Gorisch J, House D, Peters B. 2019 Hand gestures and pitch contours and their distribution at possible speaker change locations: a first investigation. In *Proc. 6th Gesture and Speech in Interaction – GESPIN 6, 11–13 September 2019*. Paderborn, Germany: Universitaetsbibliothek Paderborn.
52. Holler J, Kendrick KH, Levinson SC. 2018 Processing language in face-to-face conversation: questions with gestures get faster responses. *Psychon. Bull. Rev.* **25**, 1900–1908. (doi:10.3758/s13423-017-1363-z)
53. Schegloff EA. 1984 'On some gestures' relation to talk'. In *Structures of social action* (eds JM Atkinson, J Heritage), pp. 266–296. Cambridge, UK: Cambridge University Press.
54. Ter Bekke M, Drijvers L, Holler J. 2020 The predictive potential of hand gestures during conversation: an investigation of the timing of gestures in relation to speech. In *Proc. of the 7th GESPIN - Gesture and Speech in Interaction Conf., 7–9 September 2020.* Stockholm, Sweden: KTH.Royal Institute of Technology.
55. Nota N, Trujillo JP, Holler J. 2022 Specific facial signals associate with categories of social actions conveyed through questions. *PsyArXiv*. (doi:10.31234/osf.io/qrhdf)
56. Kaukomaa T, Peräkylä A, Ruusuvuori J. 2013 Turn-opening smiles: facial expression constructing emotional transition in conversation. *J. Pragmatics* **55**, 21–42. (doi:10.1016/j.pragma.2013.05.006)
57. Kaukomaa T, Peräkylä A, Ruusuvuori J. 2014 Foreshadowing a problem: turn-opening frowns in conversation. *J. Pragmatics* **71**, 132–147. (doi:10.1016/j.pragma.2014.08.002)
58. Duncan S. 1972 Some signals and rules for taking speaking turns in conversations. *J. Pers. Soc. Psychol.* **23**, 283–292. (doi:10.1037/h0033031)
59. Duncan S. 1974 On the structure of speaker–auditor interaction during speaking turns 1. *Lang. Soc.* **3**, 161–180. (doi:10.1017/S0047404500004322)
60. Duncan S, Fiske DW. 1977 *Face-to-face interaction: research, methods, and theory*. London, UK: Routledge.
61. Geluykens R. 1988 On the myth of rising intonation in polar questions. *J. Pragmatics* **12**, 467–485. (doi:10.1016/0378-2166(88)90006-9)
62. Kendon A. 1967 Some functions of gaze-direction in social interaction. *Acta Psychol.* **26**, 22–63. (doi:10.1016/0001-6918(67)90005-4)
63. Goodwin C. 1981 *Conversational organization: interaction between speakers and hearers*. New York, NY: Academic Press.
64. Bavelas JB, Coates L, Johnson T. 2002 Listener responses as a collaborative process: the role of gaze. *J. Commun.* **52**, 566–580. (doi:10.1111/j.1460-2466.2002.tb02562.x)
65. Cummins F. 2011 Gaze and blinking in dyadic conversation: a study in coordinated behaviour among individuals. *Lang. Cogn. Process.* **27**, 1525–1549. (doi:10.1080/01690965.2011.615220)
66. Ho S, Foulsham T, Kingstone A. 2015 Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PLoS ONE* **10**, e0136905. (doi:10.1371/journal.pone.0136905)
67. Degutyte Z, Astell A. 2021 The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. *Front. Psychol.* **12**, 616471. (doi:10.3389/fpsyg.2021.616471)
68. Gambi C, Jachmann TK, Staudte M. 2015 The role of prosody and gaze in turn-end anticipation. In *Proc. of the Annual Conf. of the Cognitive Science Society, 23–25 July 2015,* pp. 764–769. Pasadena, CA: Cognitive Science Society.
69. Rossano F. 2013 Gaze in conversation. In *The handbook of conversation analysis* (eds J Sidnell, T Stivers), pp. 308–329. Oxford, UK: Blackwell Publishing Ltd.
70. Stivers T, Rossano F. 2010 Mobilizing response. *Res. Lang. Soc. Interaction* **43**, 3–31. (doi:10.1080/08351810903471258)
71. Streeck J. 2014 Mutual gaze and recognition: revisiting Kendon's 'Gaze direction in two-person conversation'. In *From gesture in conversation to visible action in utterance* (eds M Seyfeddinipur, M Gullberg), pp. 35–54. Amsterdam, The Netherlands: John Benjamins Publishing Company.
72. Kendrick KH. 2017 Using conversation analysis in the lab. *Res. Lang. Soc. Interaction* **50**, 1–11. (doi:10.1080/08351813.2017.1267911)
73. Holler J, Kendrick KH. 2015 Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Front. Psychol.* **6**, 98. (doi:10.3389/fpsyg.2015.00098)
74. Drijvers L, Holler J. 2022 The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*. (doi:10.3758/s13423-022-02178-x)
75. Schegloff EA. 2007 *Sequence organization in interaction: a primer in conversation analysis*. Cambridge, UK: Cambridge University Press.
76. Jefferson G. 2004 Glossary of transcript symbols with an introduction. In *Conversation analysis: studies from the first generation* (ed. GH Lerner), pp. 13–31. Amsterdam, The Netherlands: John Benjamins Publishing Company.
77. Hepburn A, Bolden GB. 2017 *Transcribing for social research*. Beverley Hills, CA: SAGE Publications Ltd.
78. Wells B, Macfarlane S. 1998 Prosody as an interactional resource: turn-projection and overlap. *Lang. Speech* **41**, 265–294. (doi:10.1177/002383099804100403)
79. Rühlemann C, Gries ST. 2020 Speakers advance-project turn completion by slowing down: a multifactorial corpus analysis. *J. Phonetics* **80**, 100976. (doi:10.1016/j.wocn.2020.100976)
80. Levinson SC. 2013 Action formation and ascription. In *The handbook of conversation analysis* (eds J Sidnell, T Stivers), pp. 101–130. London, UK: Blackwell Publishing Ltd.
81. Couper-Kuhlen E, Ono T. 2007 'Incrementing' in conversation. A comparison of practices in English, German and Japanese. *Pragmatics* **17**, 513–552.
82. Schegloff EA. 2016 Increments. In *Accountability in social interaction* (ed. JD Robinson), pp. 238–263. Oxford, UK: Oxford University Press.
83. Boersma P, Weenink D. 2014 Praat: doing phonetics by computer (version 5.3.82). See http://www.praat.org/.
84. Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H. 2006 ELAN: a professional framework for multimodality research. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC 2006), 24–25 May 2006*, pp. 1556–1559. Genoa, Italy: European Language Resources Association. See http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:60436.

85. Schegloff EA. 1982 Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In *Analyzing discourse: text and talk* (ed. D Tannen), pp. 7l–93. Washington, D.C.: Georgetown University Press.
86. Zama A, Robinson JD. 2016 A relevance rule organizing responsive behavior during one type of institutional extended telling. *Res. Lang. Soc. Interaction* **49**, 220–237. (doi:10.1080/08351813.2016.1196551)
87. Jefferson G. 1973 A Case of precision timing in ordinary conversation: overlapped tag-positioned address terms in closing sequences. *Semiotica* **9**, 47–96. (doi:10.1515/semi.1973.9.1.47)
88. Jefferson G. 1988 Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In *Conversation: an interdisciplinary perspective* (eds D Roger, P Bull). Clevedon, UK: Multilingual Matters.
89. Jefferson G. 1986 Notes on 'latency' in overlap onset. *Hum. Stud.* **9**, 153–183. (doi:10.1007/BF00148125)
90. Clayman SE. 2013 Turn-constructional units and the transition-relevance place. In *The handbook of conversation analysis* (eds J Sidnell, T Stivers), pp. 150–166. London, UK: Blackwell Publishing Ltd.
91. Lerner GH. 1991 On the syntax of sentences-in-progress. *Lang. Soc.* **20**, 441–458. (doi:10.1017/S0047404500016572)
92. Drew P. 2009 In *Comparative Aspects of Conversation Analysis* (eds M Haakana, M Laakso, J Lindström), pp. 70–93. Helsinki, Finnish Literature Society.
93. Kendrick KH, Torreira F. 2015 The timing and construction of preference: a quantitative study. *Discourse Process.* **52**, 255–289. (doi:10.1080/0163853X.2014.955997)
94. McNeill D. 1992 *Hand and mind: what gestures reveal about thought*. Chicago, IL: University of Chicago Press.
95. Bavelas J, Chovil N, Lawrie D, Wade A. 1992 Interactive gestures. *Discourse Process.* **15**, 469–489. (doi:10.1080/01638539209544823)
96. Bavelas J, Chovil N, Coates L, Roe L. 1995 Gestures specialized for dialog. *Pers. Soc. Psychol. Bull.* **21**, 394–405. (doi:10.1177/0146167295214010)
97. Kita S, Gijn Iv, Hulst Hvd. 1998 Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and sign language in human-computer interaction* (eds E Efthimiou, G Kouroupetroglou, S-E Fotinea), pp. 23–35. Heidelberg: Springer.
98. Holler J, Bavelas J, Woods J, Geiger M, Simons L. 2022 Given-new effects on the duration of gestures and of words in face-to-face dialogue. *Discourse Process.* **59**, 619–645. (doi:10.1080/0163853X.2022.2107859)
99. Landis JR, Koch GG. 1977 The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174. (doi:10.2307/2529310)
100. Mondada L. 2018 Multiple temporalities of language and body in interaction: challenges for transcribing multimodality. *Res. Lang. Soc. Interaction* **51**, 85–106. (doi:10.1080/08351813.2018.1413878)
101. Bates D, Maechler M, Bolker B, Walker S. 2015 lme4: Linear mixed-effects models using Eigen and S4. See http://CRAN.R-project.org/package=lme4.
102. R Core Team. 2015 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See http://www.R-project.org/.
103. Koller M. 2016 robustlmm: an R package for robust estimation of linear mixed-effects models. *J. Stat. Softw.* **75**, 1–24. (doi:10.18637/jss.v075.i06)
104. Heritage J. 1985 A change-of-state token and aspects of its sequential placement. In *Structures of Social Action (Studies in Emotion and Social Interaction)* (ed. J Atkinson), pp. 299–345. Cambridge: Cambridge University Press.
105. Oloff F. 2013 Embodied withdrawal after overlap resolution. *J. Pragmatics* **46**, 139–156. (doi:10.1016/j.pragma.2012.07.005)
106. Cantarutti M. 2020 The multimodal and sequential design of co-animation as a practice for association in English interaction. PhD thesis, University of York, York, UK. See http://etheses.whiterose.ac.uk/27344/.
107. Argyle M, Cook M. 1976 *Gaze and mutual gaze*. Cambridge, UK: Cambridge University Press.
108. Clark HH, Krych MA. 2004 Speaking while monitoring addressees for understanding. *J. Memory Lang.* **50**, 62–81. (doi:10.1016/j.jml.2003.08.004)
109. Holler J, Alday PM, Decuyper C, Geiger M, Kendrick KH, Meyer AS. 2021 Competition reduces response times in multiparty conversation. *Front. Psychol.* **12**, 3720.
110. Lilja N, Piirainen-Marsh A. 2019 How hand gestures contribute to action ascription. *Res. Lang. Soc. Interaction* **52**, 343–364. (doi:10.1080/08351813.2019.1657275)
111. Schegloff EA. 1998 Body torque. *Soc. Res.* **65**, 535–596.
112. De Ruiter JP, Albert S. 2017 An appeal for a methodological fusion of conversation analysis and experimental psychology. *Res. Lang. Soc. Interaction* **50**, 90–107. (doi:10.1080/08351813.2017.1262050)
113. Kendrick KH, Holler J, Levinson SC. 2023 Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. Figshare. (doi:10.6084/m9.figshare.c.6423930)