Cognition 222 (2022) 105015

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Timing in conversation is dynamically adjusted turn by turn in dyadic telephone conversations

Wim Pouw [a,b,*], Judith Holler [a,b]

[a] Donders Institute for Cognition, Brain, and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands
[b] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Negative Lag-1 autocorrelation
Timing
Turn taking
Floor transfer
Rhythm
Conversation

ABSTRACT

Conversational turn taking in humans involves incredibly rapid responding. The timing mechanisms underpinning such responses have been heavily debated, including questions such as who is doing the timing. Similar to findings on rhythmic tapping to a metronome, we show that floor transfer offsets (FTOs) in telephone conversations are serially dependent, such that FTOs are lag-1 negatively autocorrelated. Finding this serial dependence on a turn-by-turn basis (lag-1) rather than on the basis of two or more turns, suggests a counter-adjustment mechanism operating at the level of the dyad in FTOs during telephone conversations, rather than a more individualistic self-adjustment within speakers. This finding, if replicated, has major implications for models describing turn taking, and confirms the joint, dyadic nature of human conversational dynamics. Future research is needed to see how pervasive serial dependencies in FTOs are, such as for example in richer communicative face-to-face contexts where visual signals affect conversational timing.

1. Introduction

A well-known phenomenon in research on human communication is that two speakers in conversation are able to take turns rapidly with only about 200 millisecond silent gaps in between on average (Ten Bosch, Oostdijk and Boves, 2005; de Vos, Torreira, & Levinson, 2015; Stivers et al., 2009). This exquisite feat of timing is deemed to be too quick to be a purely reactive process, since producing a turn in response to a turn end would take considerably longer than this (more on the order of 600 ms, as evidenced by psycholinguistic experiments on word production times, Indefrey & Levelt, 2004; Levinson, 2016). Thus, a more anticipatory mechanism must be at play. By now, we know that this process involves prediction of the content (or at least the gist) of the unfolding turn, allowing next speakers to begin planning their turn while still listening to an on-going turn (Barthel & Levinson, 2020; Barthel, Sauppe, Levinson, & Meyer, 2016; Bögels, Magyari, & Levinson, 2015; Corps, Crossley, Gambi, & Pickering, 2018). At the same time, predicting turn content also allows upcoming speakers to project roughly when the current turn may end (de Ruiter, Mitterer, & Enfield, 2006; Sacks, Schegloff, & Jefferson, 1974; but see Corps, Gambi, & Pickering, 2018) based on a whole suit of possible informative sources, including semantics, syntax, and pragmatics. In addition, turn-final cues occurring close to turn end, such as phonetic (Local & Walker, 2012) and prosodic cues (Schaffer, 1983), help upcoming speakers to anticipate an imminent termination and to launch their turn on time (Barthel, Meyer, & Levinson, 2017; Levinson, 2016; Levinson & Torreira, 2015).

This rich suite of information that is predictive of turn ends could explain how fast response times in conversation can be achieved despite the psycholinguistic complexity involved. In addition, it has been proposed that the temporal structure of conversation may facilitate the timing of turns. Wilson and Wilson's (2005) coupled oscillator model presumes oscillatory cycles determined by a speaker's syllable rate rhythmically entrains interlocutors' putative endogenous oscillators. Such oscillators are in anti-phase relation with the rhythm dictated by the speaker's syllable rate, and they govern the listeners readiness to initiate a turn. Such a mechanism could account for the precise timing of turn transitions first noted by Sacks et al. (1974). The oscillatory cycles could also account for the timing of transitions that involve some overlap or gap which results from a whole host of utterance-related factors (such as speech rate, word frequency, turn duration and complexity, predictability, speech act, accompanying visual signals, and so forth e.g., Corps, Gambi, & Pickering, 2018; Holler, Kendrick, &

* Corresponding author at: Donders Institute for Cognition, Brain, and Behaviour, Radboud University Nijmegen, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands.
E-mail address: wim.pouw@donders.ru.nl (W. Pouw).

https://doi.org/10.1016/j.cognition.2022.105015
Received 6 May 2021; Received in revised form 4 January 2022; Accepted 5 January 2022
Available online 13 January 2022

Levinson, 2018; Roberts, Torreira, & Levinson, 2015). Also, at the end of turn units where transition becomes relevant, the right to the next turn cycles back and forth between interlocutors based on a set of turn-taking rules (Sacks et al., 1974), which can impact turn timing. Thus, an oscillatory timing mechanism entrained to quasi-rhythmic features of the conversation, in conjunction with turn-taking rules and the projective power of syntax, pragmatics and semantics as well as prosodic, visual and other cues announcing upcoming turn termination would allow next speakers to anticipate rather precisely when the next turn may be launched (Wilson & Wilson, 2005).

Human anticipatory timing mechanisms have been extensively studied in more basic, non-communicative paradigms. Most notably in sensorimotor synchronization paradigms, whereby subjects are asked to tap with the rhythm of an auditory or visual beat (Repp, 2005). Such research has revealed very stable patterns of human timing behaviour. For example, when tapping to an isochronous rhythm, humans tend to anticipate the upcoming signal by tapping earlier in time, and therefore tend to have a negative mean asynchrony in relative timing distributions which can go up to 100 milliseconds in average, though usually the negative mean asynchrony is about 10–40 milliseconds (Repp, 2005). This particular negative mean asynchrony phenomenon is unlikely to be applicable to FTOs in conversation, however, as this would mean that speakers tend to consistently overlap in their turns. Indeed, a negative mean asynchrony of around −100 ms does not match the statistical distribution of FTOs in spontaneous conversation, which tend to be characterized by a positive mean asynchrony of +200 ms (Stivers et al., 2009). Critically, there is a rather wide variation around this mean of several hundred milliseconds (with some FTOs resulting in considerable overlap and others long gaps, see Stivers et al., 2009), due to factors influencing cognitive processing and conversational pragmatics, as seen above.

A second major finding in basic non-communicative timing research is that there can be statistical serial dependencies over series of motor-auditory timing events (Comstock, Hove, & Balasubramaniam, 2018; Iversen, Patel, Nicodemus, & Emmorey, 2015; Repp, 2005). Namely, when tapping to an isochronous auditory rhythm (e.g., a beeping sound), subjects tend to adjust their timing based on their previous timing. For example, when a subject times their tap too early relative to the beep, then they time their next tap later relative to the next beep to an extent that linearly scales with the previous tap-to-beep asynchrony. This goes the other way too, such that a late tap is followed by an early tap with some consistent extent, i.e., with some opposite signed magnitude. Of course, such linear dependence between consecutive timing is not a perfect 1:1 correlation (which would lead to perfect oscillations), but rather a more noisy statistical scaling relation that is nevertheless persistent throughout the series. This statistical tendency in motor-auditory timings to become negatively correlated through time at a lag of 1 is referred to as a *lag-1 negative autocorrelation* (or negative AR1). Interestingly, it has been found that this serial dependence only applies to domains where humans show exquisite performance in timing (Hove, Fairhurst, Kotz, & Keller, 2013; Iversen et al., 2015), such as tapping to an auditory discrete stimulus (e.g. a beep) or to a visual continuously changing stimulus (e.g., a moving bar), but not when the stimulus does not match the respective modality's affinity (such as tapping to a continuous sound or a discrete visual signal; Chen, Penhune, & Zatorre, 2008).

The finger tapping paradigm has been extended to dyadic situations too (e.g., Konvalinka, Vuust, Roepstorff, & Frith, 2010; Pecenka & Keller, 2011). In a study by Konvalinka et al. (2010) subjects were to continue tapping with a simple auditory rhythm that ceased after auditory presentation. The rhythm continuation either was assisted by a computer-generated stable beat, a beat generated by another person who could not hear the subject (unidirectional), or a beat generated by another person who could also hear the person and thus interact (biderectional). Only for the biderectional condition it was found that subjects "corrected the duration of their [inter-tap intervals] in the opposite directions on a tap-to-tap basis in a mutual attempt to synchronize with one another" (p. 2227). Such anticipatory mutual adaptation (and synchronization performance) is more pronounced in paired individuals with high versus low individual timing abilities; abilities that relate to musical training (Pacenka & Keller, 2011). When participants are sequentially (rather than simultaneously) tapping with an isochronous rhythm that is continuously presented, a mimicking tendency is found where timing sequences are positively correlated at lag-1 (Nowicki, Prinz, Grosjean, Repp, & Keller, 2013). Thus in basic joint tapping research we also see a serially dependent dance of mutual adjustments that lead to counteraction or mimicking. But do humans enter into such a dance in conversation too? And would we see counteraction or mimicking?

It is surprising that the serial dependence in conversational timing has not yet been tested in conversational dynamics. Even more so because timing researchers have hypothesized for example that *musical* turn taking is a kind of ability that depends on basic entertainment mechanisms next to more top-down planning (Phillips-Silver & Keller, 2012). Conversational turn-taking is of course more complex, but a more comparable phenomenon to musical turn taking. Furthermore, issues of serial dependence can be directly relevant to questions in conversation research. For example, finding a lag-1 autocorrelation (in any direction) is a key statistical signature of *active* timing adjustments (Semjen, Schulze, & Vorberg, 2000). It would further mean that the joint timing system is using its own performance (previous turn) as feedback for its future performance (the next turn), in the most basic and cognitively cheap way possible, by only adjusting its timing based on the *previous* cycle's timing.

Although there is no reason that negative AR1 cannot be instantiated by currently popular models in turn taking, such serial dependencies in FTOs have neither been assessed nor predicted (Levinson, 2016; Wilson & Wilson, 2005). This is interesting because negative AR1 could be informative about whether FTOs between speakers are regulated by the individual speakers separately, or that speakers act as some kind of coupled system regulating the conversation *as a dyad* (Wilson & Wilson, 2005; Wilson & Zimmerman, 1986). As Wilson & Zimmerman (1986, p. 377) maintain: "between-turn silence [is] interactionally generated, involving both the current and the next speaker". Here we additionally suggest that the timing of the preceding turn is used projectively in the conversation to time the next turn. This would be very much in line with the notion of language as a bilateral, joint activity (Clark, 1996) and the deeply social nature of human cognition (De Jaegher, Di Paolo, & Gallagher, 2010; Sebanz & Knoblich, 2008).

Specifically, if we find that FTOs are timed as a response to the previous turn alone (lag-1 negative autocorrelation), this would suggest that turn timing operates on the level of the dyad as both speakers are timing *the conversation's* turn transitions based on the respective previous turn – which, typically, is produced by the other speaker. However, if negative autocorrelations occur *within* speakers only, then lag-2 autocorrelations are to be expected, as turns generally return at a lag of two (i.e., speaker A [lag 0], speaker B [lag 1], speaker A [lag2]). Lag-2 autocorrelations would signal that speakers are timing their own transitions with reference to their *own previous* deviation from average transition time, where the other speaker's FTO is ignored. This is an alternative intuitive assumption at face value, and one that presumes operations based on a mechanism much more grounded in individual cognition and behaviour regulation than the notion of turn transition timing emerging from two (or more) closely coupled systems.

In the current study we test whether the phenomenon of timing adjustments as obtained from basic research on sensorimotor synchronization can be observed in more complex conversational timing as well. We perform autocorrelation analysis on an open dataset of conversational turn taking in telephone conversations (Switchboard corpus of English telephone conversation, Godfrey, Holliman, & McDaniel, 1992), to test the assumption that a timing mechanism based on lag-1 autocorrelations may be underpinning the temporal coordination of turn

transitions in talk. Note, however, that it is neither necessary, nor obvious, for the series to be autocorrelated at lag 1. In fact, it may be that there is no short-lagged autocorrelation, or that autocorrelations happen over lags of 2 or more.

2. Method

2.1. Dataset

The current dataset (retrieved here: https://osf.io/dve6h/) consists of the 'Switchboard corpus of English telephone conversation' (Godfrey et al., 1992). It was further enriched by Roberts et al. (2015) with information about FTOs (i.e. turn transition times defined by the end of turn A and the beginning of turn B). To identify turns and turn transitions Roberts et al., 2015 state that (p. 509) "*We approximated "turns" by "gluing" phonological words together if they were from the same speaker and had less than 180 ms gap between them. The floor transfer offset (FTO) or "gap" and "overlap" duration between turns from different speakers was calculated using the same method as in* Heldner and Edlund (2011)". Heldner & Edlund's method (2010) consists of treating vocal activity as a binary activity (speech vs. silence) where "gaps" consist of between-speaker silences and "overlaps" of between-speaker overlap. FTOs capturing gaps between turns result in positive numbers, with FTOs capturing overlap resulting in negative numbers.

For our analysis outlined below we treat the FTO series as a time series, as one would do for asynchronous timing series based on rhythmic tapping to a metronome. However, there is a particular peculiarity to conversational FTO series in that depending on your theoretical standing we would either include or exclude FTOs that precede or follow a backchannel contribution. Backchannels are short interjections that can, for example, signal continued engagement (e.g., "yeah", "uhu"), and which are very frequent in conversation (Knudsen, Creemers, & Meyer, 2020). Crucially, in terms of pragmatic function, backchannels—also termed continuers—pass up the opportunity to speak and thus do not constitute turns as such (Schegloff, 1982; Yngve, 1970). Backchannels might carry an inherent autocorrelation signature, where short backchannel-FTOs are interwoven with longer non-backchannel-FTOs. To understand whether any serial dependencies we may find are dependent on backchannel contributions we therefore also consider a FTO series excluding values that involve a backchannel contribution. This resulted in an exclusion of 53% of the full observed FTO series. Note that backchannel contributions which occur in complete overlap with on-going speech by the respective other speaker are excluded also in the full FTO time series, following Roberts et al. (2015).

In total, there were 349 conversations, with an average of 60.43 (SD = 21.75) conversational turns, and an average conversation time of 288.33 s (SD = 23.47). Following Roberts and colleagues, we excluded all FTOs that were negatively or positively removed from zero with 2200 milliseconds (0.36% of the data). The conversations had an average FTO of 177 milliseconds (SD = 435), indicating a tendency to initiate a turn after the previous turn ended (see Fig. 1 for distributions). When excluding FTOs associated with backchannel contributions, the average FTO was 177 (SD = 356). For further information on the current dataset see Roberts et al. (2015).

2.2. Overview analysis procedure

For each conversation the sequence of FTOs was submitted to autocorrelation analyses at lag 1 through 4. An autocorrelation for a FTO series (see Fig. 2) is a simple association of a FTO at turn t with that of a FTO at a turn t minus a turn lag of magnitude l (i.e., FTO_{t-l}). Thus the autocorrelation of lag $l = 1$ (AR1), is the correlation between FTO_t vs. FTO_{t-1}. A negative AR1 would indicate that a faster FTO is immediately (i.e., with a lag of 1 turn) followed by a slower FTO of a related magnitude and vice versa. A positive AR1 would indicate that adjacent FTOs tend to have similarly signed direction, e.g., relatively fast FTOs

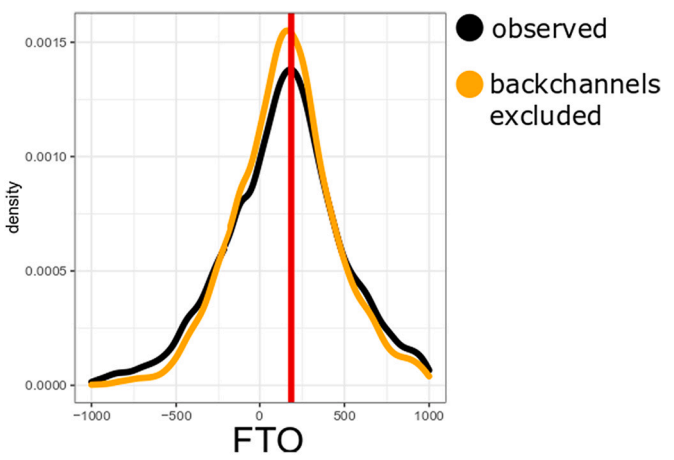

Fig. 1. Distribution of FTOs in the 'Switchboard corpus of English Telephone Conversation'.
Note. Smoothed density distributions of observed FTOs. Red-colored vertical line indicates the observed mean of the FTO distributions, with negative numbers indicating overlap, positive numbers indicating inter-turn gaps).

followed by relatively fast FTOs, and this would mean that there is a (slight) drift away from some initial average. To assess other possible serial dependencies with even longer distances, we will perform the same analyses for AR1-AR4 (e.g., Iversen et al., 2015). We will do so for a) the FTO series as observed, and b) for all *turn* transitions only, that is, excluding all backchannel responses (see Method). Critically, the autocorrelation analyses on the FTO series are performed on the FTO series (and the FTO series when excluding backchannels) as well as its order-shuffled complement, to make sure that autocorrelations are not determined by some chance ordering of the same series.

We further assess whether the degree of antipersistence at lag 1 of FTOs is related to particular pragmatic factors in an exploratory analysis. In contrast to autocorrelation values that are calculated over a series of values, the (anti-)persistence value of an FTO can be calculated point by point by assessing the change in FTO over the points (i.e., 1st derivative of FTO with respect to the turn). An individual point is persistent with a positive magnitude x, if it follows the direction of change in timing with magnitude x (e.g., if three consecutive FTO changes +20 milliseconds in timing we can conclude that the last two values are persistent). For a graphical example see Supplemental Figure 1. Conversely, an FTO value is antipersistent with a negative magnitude x, if that FTO value is opposite in the direction of change relative to the previous FTO. Average antipersistence should scale with a negative AR1 (which we will provide a sanity check for), and antipersistence values will be related in our further exploratory analyses to variables that dynamically change point by point; namely the pragmatic context, such as the dialogue act type of the previous turn, and the grouping of dialogue acts based on sequential position (initiating versus responding) and valence (negative versus positive), following the groupings used by Roberts et al. (2015).

3. Results

3.1. Autocorrelations for lag 1 (AR1) to 4 (AR4)

Fig. 3 provides an overview of the distributions of AR1 to AR4 observed for each conversation's FTO series, as observed and without backchannels, as well as a shuffled complement based on the observed series for comparison. There is a clear deviance as compared to the shuffled series at a lag of 1, indicating on average a negative AR1. We further assessed whether these differences were statistically reliable.

Firstly, we performed a simple two-sided t-test of the observed AR1 versus the AR1 of the shuffled series, yielding a highly reliable

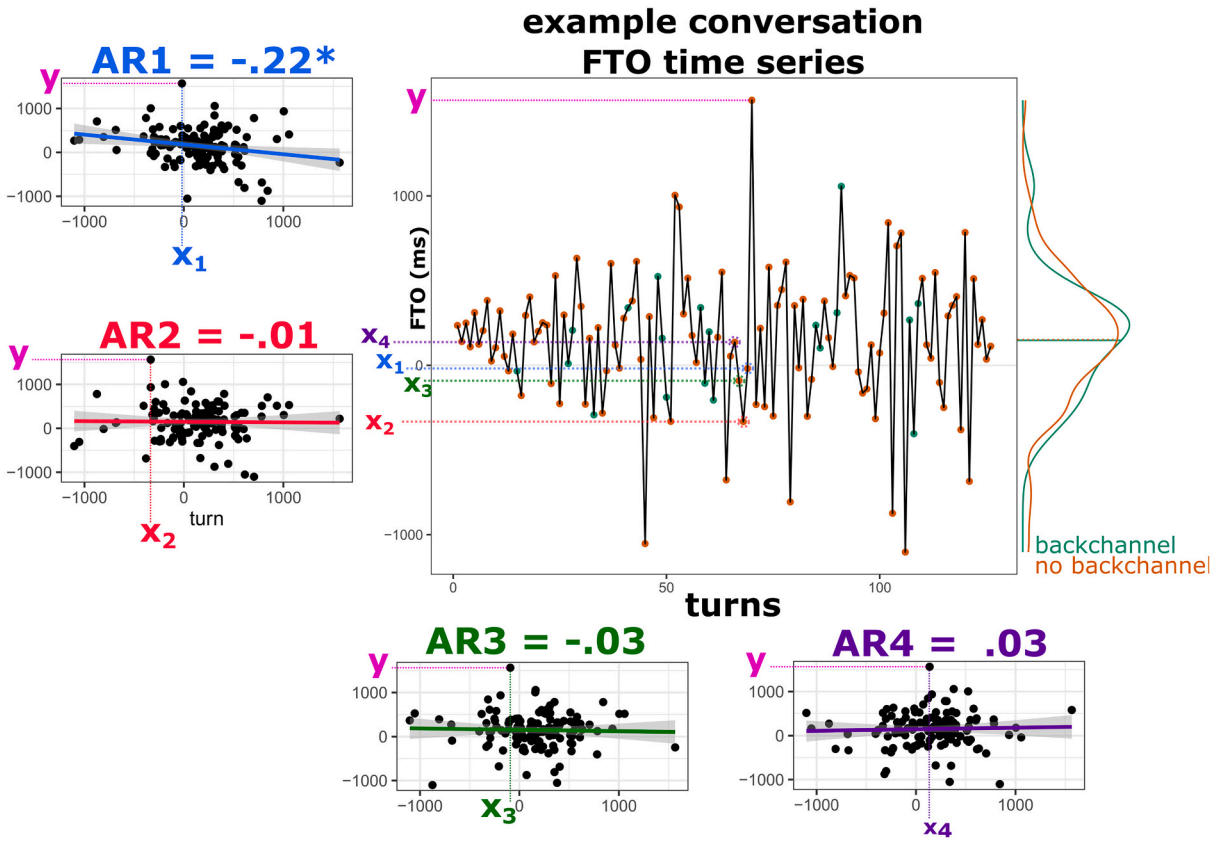

Fig. 2. Example of FTOs from a single Switchboard Corpus conversation.
Note. This is an example of the FTO data of one conversation from the Switchboard Corpus which had a particularly high magnitude negative AR1. For the time series (right upper panel), each point indicates a transition between speakers (i.e., no within-speaker transitions). All FTOs are shown in their original order in the conversation. Colored points (orange vs. green) indicate speaker FTOs involving backchannels (green), or FTOs not involving backchannels (orange). The lines orthogonal to the added density distributions of the time series FTOs indicate average FTO. Note that this time series has a negative lag-1 autocorrelation of $r = 0.22$ ($p = .012$) as indicated by the blue slope where FTO_i vs FTO_{i-1} is plotted. Please note that for autocorrelation plots, each FTO for a particular turn Y is plotted against the FTO of a previous turn X at lag 1–4 (i.e., x_1, x_2, x_3, x_4 in this example). There are no statistically reliable (p's > 0.05) AR2 (in red), AR3 (in green) or AR4 (in purple) observed for this series. The negative AR1 is reflected in the time series such that relatively fast timing of a turn tends to be immediately followed by a slower timing, and vice versa such that a slow transition is followed by a fast transition. However, the oscillations are not simply repeating or perfectly cyclical, but rather change in amplitude over the conversation, thereby losing dependencies over longer timescales (AR2, AR3, or AR4). For the final analysis we also performed an autocorrelation analysis on this series but excluding the backchannel-labelled FTOs.

difference, $t = -6.966$, $df = 348$, $p < .0001$. The series that did not include backchannels had a reduced AR1 relative to the observed series but this difference did not reach statistical significance, $t = -0.520$, $df = 348$, $p = .603$. Even when removing backchannels, the negative AR1 for that series was still reliably more extreme than the shuffled series, $t = -5.521$, $df = 348$, $p < .0001$.

Secondly, we determined the reliability of the effect of the negative AR1 irrespective of a shuffled baseline by performing mixed linear regression (using maximum likelihood estimation) of FTO onto FTO lag 1, with random intercept and slope for dyad. This also confirmed that there was a negative relation between FTO relative to FTO of the previous turn, b [95% CI: lower, upper] $= -0.064$ [-0.079, -0.049], $SE = 7.878$, t (20314) $= -8.350$, $p < .0001$, Cohen's $d = -0.120$, SD random effect per conversation $= 0.022$. The intercept in the model was estimated at $b = 202.70$, $SE = 7.88$, t (20314) $= 25.73$, $p < .001$, SD random effect intercept per conversation $= 131.99$.

When removing backchannels we also observed a reliable AR1, b [95% CI: lower, upper] $= -0.033$ [-0.054, -0.012], $SE = 0.011$, t (10450) $= -3.642$, $p < .002$, Cohen's $d = -0.06$, SD random effect per conversation $= 0.029$. The intercept in the model was estimated at $b = 201.25$, $SE = 9.83$, t (10450) $= 20.48$, $p < .001$, SD random effect intercept per conversation $= 131.99$.

4. Simulations

The current AR1 tells us that in a conversation a shorter FTO_t tends to be followed by a longer FTO_{t+1} and vice versa, but only on a turn by turn basis (i.e., not strictly periodically). There is a simple way to demonstrate that an exclusive negative AR1 does not show up simply because two different speakers take turns. Consider that in dyadic conversation two systems are in play, namely person A and person B, who both have intrinsic tendencies to speak (and interrupt) at a certain time (mean FTO A $\neq$ mean FTO B) with a certain variability (SD FTO A $\neq$ SD FTO B). Further person A and B alternate in taking turns. To exemplify what autocorrelation structure arises out of these simple facts, we computed for each conversation the observed *Mean* and *SD* for FTOs. Then we simulated FTOs based on Gaussian distributions with that same *Mean* and *SD*, alternating A and B turns, with a total amount of turn transitions equal to the observed amount of turn transitions for that conversation. We then recomputed the autocorrelations. Fig. 4 shows the simulated ARs next to the observed ARs. Table 1 also provides the AR coefficients for the simulated series.

The analyses on the simulated data reflect the periodic structure that we invoked by sampling from two Gaussian distributions that contribute FTOs to a conversation FTO series in alternating fashion: this resulted in positive correlations between FTOs produced by the same timing

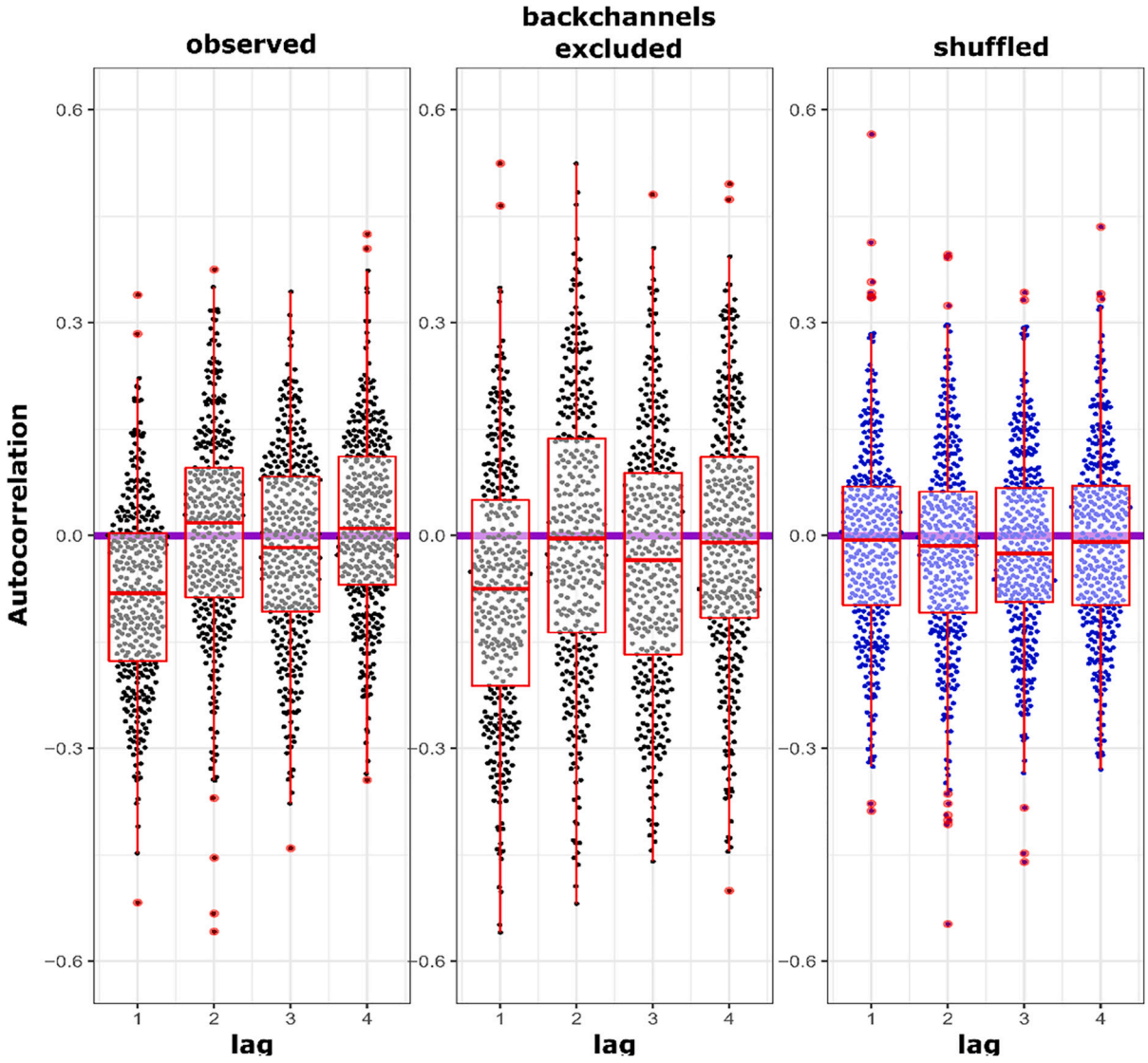

Fig. 3. Autocorrelations for observed, backchannel excluded, and shuffled series.
Note. Each point in a point cloud represents an autocorrelation value (for lag 1 through 4) of the respective FTO sequence for a particular conversation. The point clouds are organized according to their density distribution. The left panel shows the observed autocorrelations, and the middle panel the series with backchannels excluded. The outer right panel shows the autocorrelations performed on the randomly shuffled FTO sequences. It can be seen that for observed data there tends to be a negative lag-1 autocorrelation, which is clearly divergent from AR2–4 and the shuffled series. The serial dependency is destroyed when shuffling the series as shown in the right panel. This means that the dyads are timing FTOs with reference to the previous FTO, rather than with reference to *their own* previous FTO. Indeed, in case of the latter, negative lag-2 autocorrelations would be observed (since speaking turns tend to alternate between speakers).

distribution (AR2 and AR4) and negative correlations between FTOs contributed between different timing distributions (AR1and AR3), with no dissipation of correlation strength over lags (e.g., AR1 is statistically similar to the AR3; see Table 1), as the exact same timing distribution 1 and 2 is sampled from in alternating fashion. This is different from the observed data where exclusively a statistically reliable negative AR1 is found. What this means is that information is lost over the conversation in the observed data, where only lags of 1 are serially dependent, and that the dependency dissipates over longer lags due to dynamic variations that happen in FTOs during the conversation. Thus, we conclude that the observed data are not a product of two independent Gaussian processes that are sampled in alternation. The observed data show rather that there is a short-scale lag-by-lag serial dependency, suggesting a more dynamic and between-speaker dependency where the next speaker responds to the current speaker in a manner that relates to the extent of the temporal deviation from the mean that the previous speaker's response time caused. If the observed data were generated by two *independent* processes, then we should have found the persistent positive AR2 as indeed shown in the simulated data produced by two independent processes.

4.1. *Exploratory analysis: possible contributors to negative AR1 in conversation FTOs*

Each turn is characterized by a particular dialogue act, speech rate, duration, and other properties in the original coding of the Switchboard Corpus, and these factors are known to influence FTOs (e.g., Roberts et al., 2015). So how can we quantify which of those factors may be contributing to the negative AR1 we observe, which was calculated on the level of the whole conversation (other than excluding certain dialogue acts, as we did with backchannels)? We do this by relating the individual factors (including speech rate, turn duration, and dialogue act), with the (anti-)persistence value of that turn in the series, which should contribute to negative AR1. An antipersistent value (as shown in

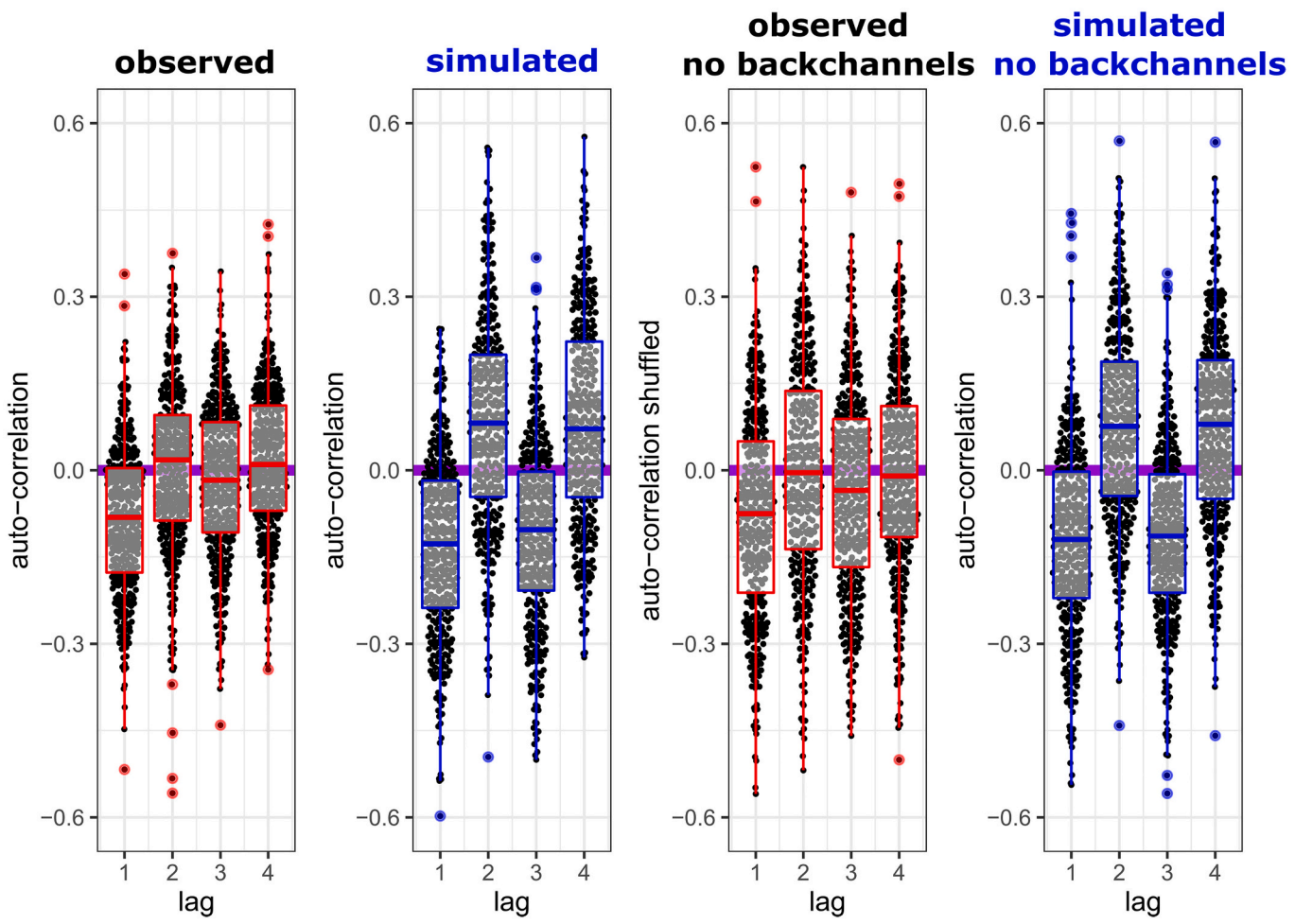

Fig. 4. Observed versus simulated series.
Note. Simulated data are shown in blue, where FTOs were simulated based on sampling from Gaussian distributions that were parameterized with the empirically observed *Mean* and *SD*s for the FTOs from person A (to B) and person B (to A). The observed data plots are the same as reported in Fig. 3. It can be seen that the simulated data are markedly different from the observed data, where for the observed data there is only a persistent dependency at lags of 1, while in the simulated data the dependency is more static, with persistent positive between-turn correlations.

Table 1
Autocorrelation (lag 1 through 4), Means and (SDs) for observed FTOs, observed data with backchannels removed ('no backchannels'), shuffled series, and simulated series.

	AR1 (*r*)	AR2 (*r*)	AR3 (*r*)	AR4 (*r*)
Observed *M* (*SD*)	−0.083 (0.136)	0.005 (0.143)	−0.020 (0.133)	0.016 (0.133)
95%CI[lower, upper]	[−0.096, −0.069]	[−0.010, 0.0.021]	[−0.034, −0.005]	[0.002, 0.029]
Observed No backchannels *M* (*SD*)	−0.078 (0.183)	−0.005 (0.198)	−0.040 (0.181)	−0.007 (0.177)
95%CI[lower, upper]	[−0.097, −0.059]	[−0.026, 0.016]	[−0.060, −0.020]	[−0.026, 0.011]
Shuffled *M* (*SD*)	−0.012 (0.137)	−0.020 (0.143)	−0.019 (0.131)	−0.008 (0.133)
95%CI[lower, upper]	[−0.027, 0.002]	[−0.036, −0.005]	[−0.33, −0.006]	[−0.022, −0.006]
Simulated *M* (*SD*)	−0.131 (0.162)	0.084(0.188)	−0.113 (0.168)	0.081 (0.182)
95%CI[lower, upper]	[−0.148, −0.114]	[0.064, 0.104]	[−0.131, −0.095]	[0.062, 0.101]
Simulated No backchannels *M* (*SD*)	−0.21 (0.168)	0.774 (0.174)	−0.117 (0.169)	0.072 (0.173)
95%CI[lower, upper]	[−0.139, −0.103]	[0.059, 0.095]	[−100, −0.133]	[0.054, 0.090]

supplemental Fig. A) is a FTO that changes in the opposite direction of the previous FTOs change. As a sanity check that the (anti-)persistence values and the AR1 autocorrelation of the whole series are correlated, Fig. 5 shows a summary of this dependence, where for each conversation the mean (anti-)persistence is given against the autocorrelation of that series, $r\,(347) = 0.487$, $p < .0001$). We can now use the antipersistence value to relate it to properties that are at the level of the turn rather than the conversation.

Since we have some information about each FTOs relative contribution to a negative AR1 of that series, it is now possible to see whether there are any clear contributors from turn-by-turn defined properties. We do not have a specific hypothesis at this stage and will report some exploratory analyses only. Firstly, speech rate, normalized last vowel and consonant duration of the turn did not seem to relate (*ps* = *ns.*) to the persistence value of the following FTO (see supplemental Fig. B). However, we did find a reliable but very weak correlation between antipersistence and previous turn duration, $r = -0.029$, $t\,(21011) = -4.211$, $p < .001$, where the longer a turn takes the more likely a floor

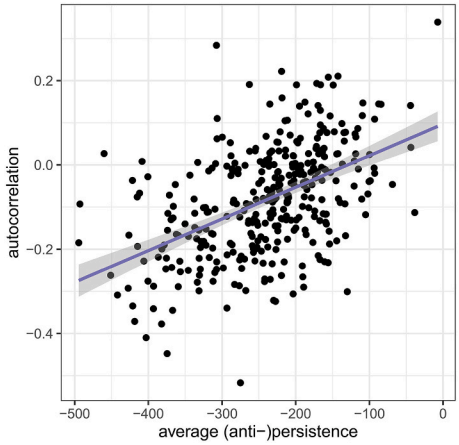

Fig. 5. Relation between mean FTO (anti)persistence values with the AR1 coefficient.
Note. Average persistence values (negative averages indicate general opposite change FTO of *M* milliseconds) plotted against the AR1 coefficient for each conversation. In general, more negative mean persistence values relate to more negative AR1.

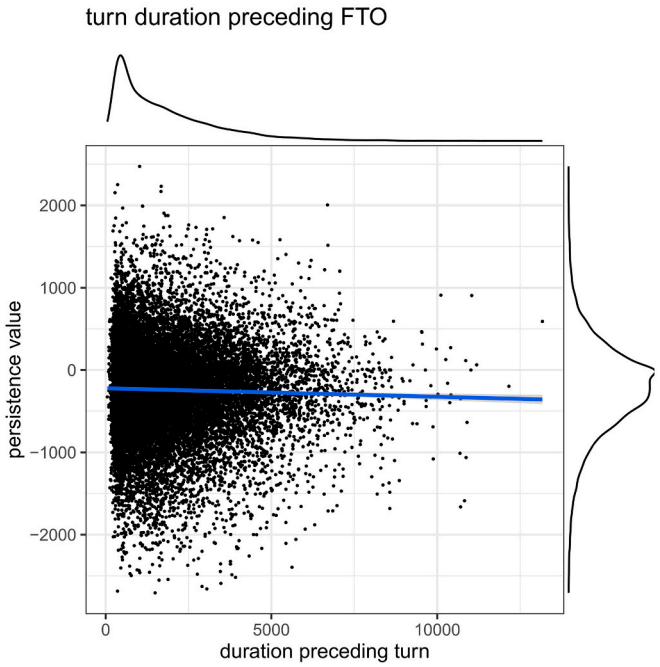

Fig. 6. Scatterplot of duration of previous turn with (anti)persistence value.

transfer offset follows that changes in the opposite direction from the previous FTO; thus when there is shorter (or longer) FTO, followed by a long turn, it is likely that the next FTO is longer (or shorter) (see Fig. 6).

Note. Duration of the previous turn in ms is given relative to the (anti)persistence value of the following FTO, together with their respective smoothed density distributions. It shows that there is indeed a weak tendency such that the longer the turn takes the more likely that the FTO goes in the opposite direction from the previous FTO.

We also globally assessed possible clear differences in FTO persistence values as related to the dialogue act type. For example, it is wholly possible that the timing pattern contributing to the negative AR1 (i.e., antipersistence) is related to only a few dialogue act types that occur a lot during the conversation. However, Fig. 7 shows that no particular dialogue act seems to be driving the negative AR1 as antipersistent FTOs are observed across the board for all dialogue act types. There might of course be differences in the degree of antipersistence of FTOs, but we will not test this post-hoc in the current sample given the many multiple comparisons and the concomitant type-1 error inflations, as well as the large variation in the number of datapoints per dialogue act category. To provide some indication for possible differences that would be statistically detectable, we have used a similar grouping of a subset of the dialogue act types as categorized by Roberts et al. (2015), where we distinguish between positively versus negatively valenced dialogue acts, as well as responses and initiations. Fig. 7 and Table 2 provide the main results for these comparisons. They show that these larger dialogue categories all tend to be antipersistent, and tend to have overlapping confidence intervals between categories. Judging from the non-overlapping confidence intervals there are no clear differences in antipersistence. We thus conclude that the current antipersistent turn-by-turn tendencies that contribute to the overall negative-lag 1 autocorrelation seems to be stable across the currently defined pragmatic contexts.

Note. Boxplots and density distributed jitter points for each turn's dialogue act type and its FTO. The original dialogue act types are shown in panel (A), and the broader categories used to group dialogue acts by Roberts and colleagues are shown in panels (B) & (C). FTOs can be persistent (to the right of the gold line) or antipersistent (to the left of the gold line). Antipersistence is contributing to the negative AR1 observed for the data. From the current graph it can clearly be seen that all dialogue acts seem to have in general antipersistent FTOs, suggesting that negative AR1 is a phenomenon that may not depend on a particular dialogue act type or group. There are possibly some specific dialogue act types in panel (A) that have more extreme antipersistent FTOs, but this is difficult to currently ascertain due to issues of multiple comparisons and unequal sample sizes. Future targeted hypothesis testing should be employed to see whether AR1s are more likely to be observed for particular types of dialogue acts and turn transitions (e.g., question-answer sequences).

5. Discussion

The current findings provide preliminary evidence that faster turn transitions are likely to be followed by slower turn transitions (FTO) at a turn lag of 1 and vice versa. This basic phenomenon applied to FTO series, referred to as lag-1 negative autocorrelation (negative AR1), is a well-established phenomenon in much more basic timing capabilities of humans (e.g., Iversen et al., 2015). Namely, when tapping to an isochronous auditory metronome, timers tend to show a negative AR1 in the relative timing of their tap, relative to the metronome's beat. This is a timing mechanism based on the short-term information of tapping too soon, which helps the next tap to be timed a little later. In conversation, two agents are in play, and the astonishing finding is that agents are sensitive to each other's FTOs as they adjust their own FTOs based on the previous FTO (determined by the previous speaker's response time), rather than only their own which typically occurs at a lag of 2. Indeed, we did not observe a reliable AR2 (or AR3 or AR4), but only a negative AR1 in the FTOs. Further simulating FTOs based on two Gaussian independent processes that were parametrized by the observed data, we confirmed that such independent processes cannot generate an exclusively negative AR1, instead showing a similar-valued positive AR2 and AR4, indicating autocorrelations of two *independent* alternating, periodic processes. We thus suggest that the current findings indicate that negative AR1 is a dynamic process adjusted turn by turn through two speakers timing to take the floor.

We reason that the negative AR1 phenomenon in conversation FTOs emerges from a process of joint action, based on interlocutors forming a coupled system. This is compatible with the idea that turn-taking is coordinated at the level of the dyad (or group) rather than resulting from individual action (e.g., Wilson & Zimmerman, 1986). Thus, the current findings extend a basic principle of the oscillator model of turn-taking

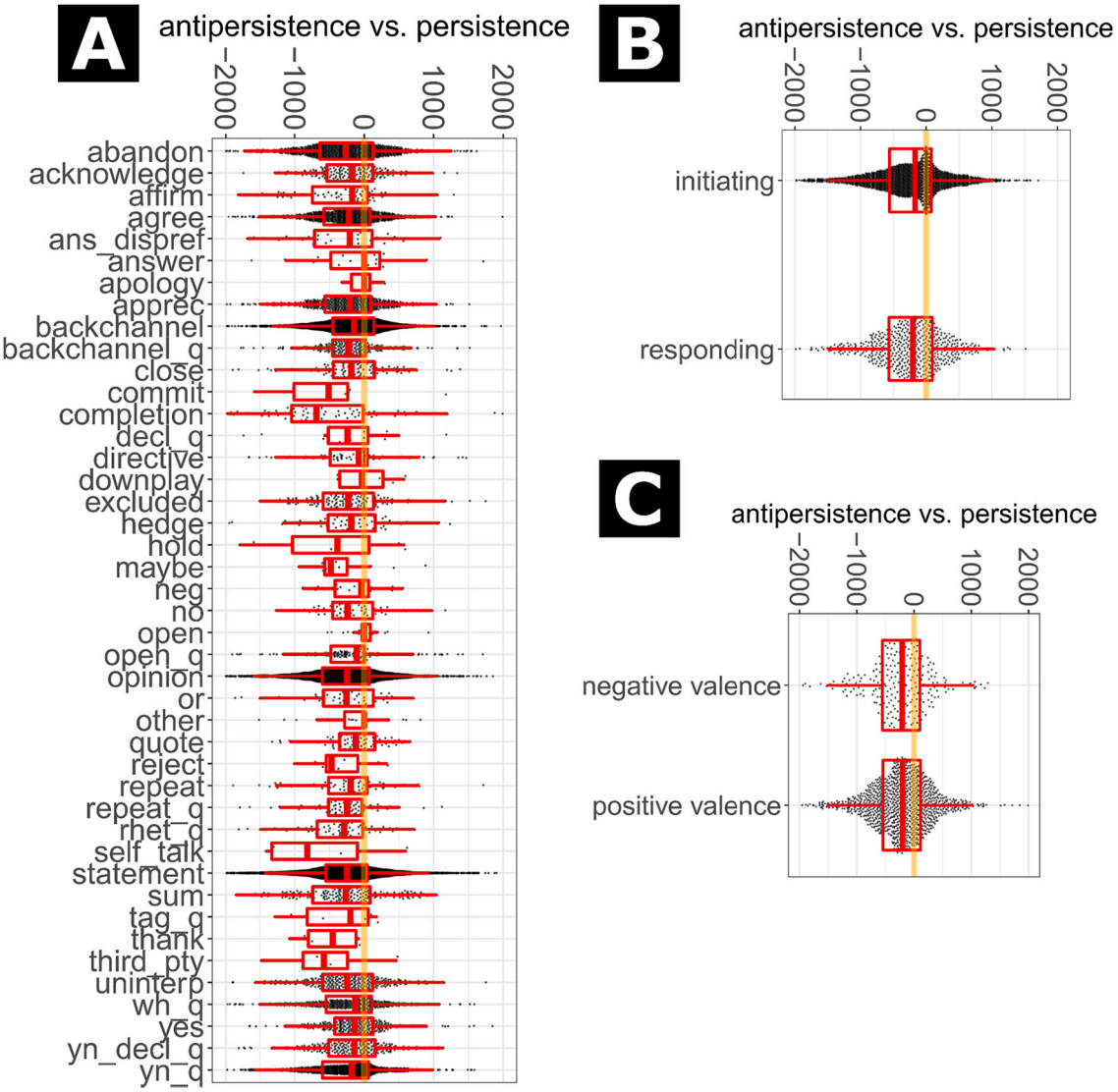

Fig. 7. Dialogue act type and (anti-)persistence.

Table 2
Descriptives for (anti-)persistence values per dialogue act grouping.

	Persistence *M* (SD) 95%CI[lower, upper]
Initiating	−241.56 (589.12) [−265.83, −217.28]
Responding	−256.22 (567.19) [−303.72, −208.71]
Negative valence	−257.08 (492.74) [−182.37, −331.79]
Positive valence	−227.60 (555.69) [−194.12, 261.07]

(Wilson & Wilson, 2005) to an account where interlocutors entrain not only to the syllable-rhythm cycling, as originally proposed (facilitating timing in terms of when to launch a turn), but also the relative timing of turn transitions (that is, the duration of turn transitions, which, in turn, influence the next). Investigating the particular ways turn transitions change the oscillators' cycling goes beyond the present research, but seems to be implementable in a number of ways, for example by shifting the phase or adjusting the period of the oscillator based on the previous lag-1 asynchrony (see Repp, 2005 for an overview of how timing adjustments can be implemented in such models).

While the negative AR1 confirms that the calibration of FTO is regulated at the level of the dyad, rather than at the level of the individual speaker, it is important to stress that the underlying dynamics for the current statistical signature for turn by turn serial dependence is a matter of speculation. To speculate, we imagine that there are varying levels of cognitive complexity that one can invoke to simulate the

current negative AR1. For example, since basic rhythmic timing as well as conversational timing show similar statistical dependencies, and both domains have been approached from coupled oscillator perspectives, it is possible that the negative AR1 is a basic phase-adjustment mechanism instantiated by a domain-general coupled-oscillator timing system.

However, it is also possible that there is another level of explanation, in the sense that timing a turn in relation to the previous turn's timing is an embodied mechanism that allows one to enact rules that seem to govern conversation. Namely, Sacks et al. (1974) state that interlocutors aim to initiate turns without much of a silent gap or overlap. We also know that turn transition times are highly variable and a turn coming in a little earlier or later is actually pragmatically meaningful. Thus precision attained on the level of the conversation (small mean turn transition times) is heavily contextualized by high variability generated by overlaps and silences of varying lengths. In this line of reasoning then the negative AR1 can be seen as a way to regulate precision under a highly variable system (see Keller, Novembre, & Hove, 2014). Interestingly, Repp and Keller (2008) showed that when participants are asked to time a tap with a computer-generated beat which has a phase adjustment in the opposite direction of the participant's timing tendency, thus hampering synchronization, musically trained participants will adjust their timing to counteract the artificial partner. Perhaps in a similar fashion negative AR1 FTO timing in conversations is a way to regulate precision under dynamic variability: To adjust one's timing in the opposite direction of the other's variable timing may help to enact Sacks et al.'s rule of aiming for consistent timing under the pragmatically functional perturbations deviating from the norm (minimal-overlap/minimal silence).

Above we have argued that the negative lag-1 phenomenon can be seen as a dynamic constraint on the natural variability. What mechanism enacts this constraint remains a matter of speculation going beyond the current results, as stated above. However, we can further make two broad distinctions in possible approaches (Stepp & Turvey, 2010): Namely those approaches that suggest that some anticipatory (timing) behaviour is based on an internal model that discretely computes timings based on modality-specific stored information of the past and current states. Alternatively, there are approaches that suggest that the future can be anticipated based on ongoing (time-delayed) coupling with relevant information in the environment.

Internal model approaches need not be complex, and a very lean account to explain negative lag-1 correlations in inter-tapping intervals is provided by Wing & Kristofferson (1973; for a broader perspective see, Wing, 2002). They show that one only needs to posit that a central timer dictates a timing interval (with some variability) from which tapping commands are issued at each interval boundary that however suffer from a variably delayed motor implementation (e.g., due to nerve-conduction constraints). If we then also assume that the variability of the central timer and the motor system are independent, then if a tap is a bit earlier implemented due to variable motor delay, by necessity, the tapping interval is shortened, and the next one inevitably lengthened. While it is not immediately clear how such a mechanism would extend into more complex conversational between-person timing, possibly compatible auxiliary assumptions can be made whereby feed forward models of the motor system are implemented that also issue motor predictions for the other person (Fisher, Hadley, Corps, Pickering, 2021). Such assumptions are invoked to explain for example why dual motor task can interfere with predicting turn ends in music and speech (Fisher, Hadley, Corps, Pickering, 2021).

Another set of approaches often referred to as dynamic or emergent timing approaches (Schöner, 2002) emphasize the ongoing continuous coupling with the environment, where often a system of (coupled) oscillators is invoked that respond to some temporal structure in the environment in a particular way as determined by the system-internal coupling relations between the systems' components (see also Abney, Paxton, Dale, & Kello, 2021; Large, Herrera, & Velasco, 2015; Tognoli, Zhang, Fuchs, Beetle, & Kelso, 2020). Negative lag-1 autocorrelations can be modeled as a continuously coupled oscillator timing system too (Schöner, 1994), where random variations that deviate from some typical continuous oscillatory cycle are attracted to enter back into the typical oscillatory regime. Recently it has even been shown that dynamic timing models that have an about 200 ms delayed self-coupling loop give rise to the characteristic negative mean asynchrony discussed in the introduction, which was also extended for joint timing, where a delayed coupling was established *between* agents that were tapping in alternation (Roman, Washburn, Large, Chafe, & Fujioka, 2019). Thus it seems that dynamical systems models can be in principle extended to account for serial dependencies in conversational timing, and indeed Wilson & Wilson (2005) propose such an extension by suggesting that persons entrain to one another continuously in an anti-phase fashion so as to time the next turn.

It is wholly possible that there are explanations of the current phenomenon that may arise out of the unique constraints of conversing, rather than arising out of basic timing mechanisms. A reviewer of the current research interestingly suggested that the negative lag-1 autocorrelation in FTO series may be accounted for in the following way: short FTOs are likely to follow longer turn durations because turn ends can be more easily anticipated for longer turns. Further, shorter gaps tend to precede turns that require less planning time, and such turns are often of shorter duration (Roberts et al., 2015). The reverse arguments would hold for longer gaps and longer turn durations. However, we do not think that this rationale can explain our findings, for the following reasons.

Firstly, it should be noted that we must explain a serial dependence that occurs throughout the FTO series. Thus suggesting that FTOs are driven by turn duration does not necessarily explain the serial dependency. After all, it is possible that indeed long turns tend to be followed by short gaps, and short turns tend to followed by long gaps, but if the turn durations themselves are not serially dependent but randomly structured (or oscillating between two random processes as in our simulations), we would not have a FTO series that is negatively lag-1 autocorrelated (but that can nevertheless be perfectly correlated with turn durations). Thus by letting FTOs be driven by turn duration one does not explain the current serial dependence of FTO series, one must also then predict (on some basis) a negative lag-1 autocorrelation in the turn duration. To check this we assessed turn duration autocorrelations lag-1 to 4, and it showed a more oscillatory pattern, with negative lag-1, positive lag-2, and negative lag-3 autocorrelations, much more akin to our simulation of two random independent processes generating an FTO in alternation (see supplemental materials). Thus since FTO autocorrelation structure does not match turn duration autocorrelation structure, an explanation where turn duration drives FTO serial dependence seems unlikely.

Secondly, in our exploratory analysis we find that longer turn durations were reliably correlated with the antipersistence of the FTO. This means that longer turn duration tends to follow with an FTO that goes in the opposite direction to the previous FTO. But importantly this means that longer turn durations can result in both shorter as well as longer FTOs (as compared to the previous FTO), and thus an exclusive relation where longer turn durations tends to follow a shorter FTO that contributes to the negative lag-1 autocorrelation does not seem to hold. Indeed, if we split out antipersistence values for longer or shorter antipersistent FTOs, we clearly see that longer and shorter antipersistent FTOs scale similarly with turn duration (see supplemental figure S3).

Of course, we think it is very much possible that there is an explanation - very much like the one above - that is solely rooted in conversational dynamics rather than the basic timing phenomenon in individual and joint tapping tasks. However, we think at present the most parsimonious explanation is one that accepts that timing in conversation is rooted in more basic timing abilities such as (jointly) tapping to a more predictable metronome. This timing mechanism would interact with the variability in timing related to pragmatic factors. One would expect, for example, that dispreferred responses, which tend to be

characterized by longer than average turn transitions (see Kendrick & Torreira, 2015 for an overview) would be associated with stronger antipersistence patterns, based on the basic timing mechanism we propose here, which would lead to counter adjustments in the following FTO (i.e., shorter than average FTOs, equalling weaker antipersistence). This trend was not observable in the present data, but due to the small sample size in this category, the data for this dialogue act by itself may not yield reliable information, and the heterogeneity of the dialogue act categories applied (Jurafsky et al., 1997) might further complicate drawing firm conclusions, emphasizing the need for future research into this matter. Also, we ought to bear in mind that the participants in the present dataset were strangers, being asked to talk with various interlocutors one after another, while conversation analytic investigations typically focus on more naturalistic settings. This may mean that the conversations analyzed here are characterized by different pragmatic constraints, potentially making certain actions more (e.g., agreements) and others less likely (e.g., disagreements), which may influence the frequency of their occurrence as well as their interactional significance.

One other interesting aspect of dialogue act type that may warrant further investigation is the timing of backchannel responses. Although pragmatically they function differently to speaking turns (allowing potential next speakers to forego the opportunity to take the turn, in fact), interestingly, backchannels seem to pattern in the same way as turns (i.e., based on negative AR1), judging by the reduction in the strength of negative AR1 when backchannels are removed from the dataset.

A related issue that we need to be aware of is that the timing of FTOs are not fixed targets. Indeed, a person may aim to intervene and therefore time their turn to interrupt, or may want to leave a silence to show surprise, or delay their response when it is dispreferred (see above). Therefore, in conversations the FTO target is moving all the time, and according to pragmatic constraints. It is therefore not surprising that we find a small effect size in terms of the negative lag-1 autocorrelation; after all the variability in conversations due to pragmatically moving targets injects what we now treat as random noise relative to the estimate of our effect. It would indeed be surprising if effect sizes found in tapping tasks would be replicated in terms of the same magnitude in conversational timings too. Given this variability, it is very interesting that we still find what seems to be a highly statistically reliable serial dependence in FTOs. Nevertheless, we also need to emphasize that the negative lag-1 autocorrelations themselves were highly variable over the conversations (see Fig. 3, left panel), implying that the current phenomenon should not be seen as a hard and fast rule, but rather a statistical tendency that for some conversations are very pronounced, while for a small minority this tendency is even completely absent.

A further caveat is that our floor transfer offset durations are derived from simple manual, tool-based acoustic measurements (distinguishing between phonation and silence). Such measurements may differ from interlocutors' psychological perceptions in conversational context. One way to arrive at more certainty concerning the equivalence of the two would be to compare the simple manual acoustic measurements to measures from experimental studies (e.g., combining encephalographic and reaction times) targeting the detection of acoustic stimuli boundaries in a variety of contexts, ideally embedded in interactive, conversation-like settings. Future research is needed to advance our knowledge in this domain.

There are some further promising avenues of research further informed by basic timing research. Namely, it has been shown that perceptual systems (e.g., auditory vs. visual) have modality-specific affinities for temporal coupling, with synchronization being most optimal for auditory-discrete and for visual-continuous stimuli (Comstock et al., 2018; Hove et al., 2013). Furthermore, in the study on sequential joint timing discussed in the introduction (Nowicki et al., 2013, see experiment 2), it was also found that auditory rather than visual information was important to induce serial dependencies in joint tapping. It might thus be possible that when *multimodal* perceptual systems are used in conversation that different temporal characteristics emerge in FTO series as currently observed in auditory-only coupling during the telephone conversations. It is also possible that especially in multimodal interactions the complexity of the temporal structure of social interaction increases (Pouw et al., 2021). Emergent timing perspectives have proposed several methods for gauging temporal complexity, such as complexity matching, which probes between-person temporal dependencies that happen on much longer time scales, spanning the whole conversation than the still short time scales (max. 4 turn transitions) have studied here (Abney et al., 2021).

Floor transfer timing may not only be different because there are different channels for communication however, they may also be different because the very information about turn transfer (silences, and overlaps) used by speakers to time a next turn can be perceived differently depending on the linguistic or wider (multimodal) context. Indeed, while our turn durations are derived from simple acoustic measurements, listeners attune to such acoustic energy in relation to the contexts, and thus we should also remind ourselves that our FTO measure in the current paper does not need to be a simple reflection of the psychological process of turn transition perception and action.

A further identification of the prevalence of the current joint timing phenomenon (or lack thereof) across animal species (Takahashi, Narayanan, & Ghazanfar, 2013; Okobi, Banerjee, Matheson, Phelps & Long, 2019) and different human languages and contexts (e.g., Stivers et al., 2009) can have important theoretical implications for elucidating domain-general mechanisms that may underly timing in communicative turn taking, next to possible domain-specific mechanisms (Castellucci et al., 2022). The negative AR1 phenomenon exhibited could further be accounted for by oscillator models of turn taking, possibly improving their predictive power (Takahashi, Narayanan, & Ghazanfar, 2013; Wilson & Wilson, 2005). Any of these models should take into account that the timing mechanism seems to operate despite variation due to dialogue act types. To conclude, the currently observed pervasiveness of the AR1, and the current findings that FTO is adjusted at the level of the dyad rather than the individual, provides a basis for the claim that timing to take the turn emerges from two speakers in interaction as originally emphasized by dynamical system models and their precursors (Wilson & Wilson, 2005; Wilson & Zimmerman, 1986).

Open data and analysis statement

The analyses scripts and data are available on the Open Science framework: https://osf.io/6tqhp/.

Credit author statement

WP generated the idea for the manuscript. WP analyzed the data. WP and JH wrote the manuscript.

Acknowledgement

This work is supported by a Donders Fellowship awarded to Wim Pouw and is financially supported by the Language in Interaction consortium project 'Communicative Alignment in Brain & Behaviour' (CABB), as well as by European Research Council consolidator grant (#773079) awarded to Judith Holler. We would like to thank Sean Roberts, as well as participants of the Interaction Meetings Nijmegen (especially Sara Bögels, Marisa Casillas, Mark Dingemanse, Natalia Levshina, James Trujillo, Marlou Rasenburg, and Marieke Woensdregt) for contributions to discussion of this research. We would like to thank the three anonymous reviewers and a non-anonymous reviewer Dr. Mathias Barthel for their extremely useful and constructive feedback on the current work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105015.

References

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2021). Cooperation in sound and motion: Complexity matching in collaborative interaction. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001018

Barthel, M., & Levinson, S. C. (2020). Next speakers plan word forms in overlap with the incoming turn: Evidence from gaze-contingent switch task performance. *Language, Cognition and Neuroscience, 35*(9), 1183–1202. https://doi.org/10.1080/23273798.2020.1716030

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final "go-signals.". *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00393

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01858

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5*(1), 12881. https://doi.org/10.1038/srep12881

Castellucci, G. A., Kovach, C. K., Howard, M. A., Greenlee, J. D. W., & Long, M. A. (2022). A speech planning network for interactive language use. *Nature*, 1–6. https://doi.org/10.1038/s41586-021-04270-z

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Moving on time: Brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of Cognitive Neuroscience, 20*(2), 226–239. https://doi.org/10.1162/jocn.2008.20018

Clark, H. H. (1996). *Using language*. Cambridge university press.

Comstock, D. C., Hove, M. J., & Balasubramaniam, R. (2018). Sensorimotor synchronization with auditory and visual modalities: Behavioral and neural differences. *Frontiers in Computational Neuroscience, 12*. https://doi.org/10.3389/fncom.2018.00053

Corps, R. E., Crossley, A., Gambi, C., & Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition, 175*, 77–95. https://doi.org/10.1016/j.cognition.2018.01.015

Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes, 55*(2), 230–240. https://doi.org/10.1080/0163853X.2017.1330031

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences, 14*(10), 441–447. https://doi.org/10.1016/j.tics.2010.06.009

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). *Switchboard: Telephone speech corpus for research and development, 1, 517–520. San Francisco, CA*. Retrieved from https://catalog.ldc.upenn.edu/LDC97S62.

Fisher, N. K., Hadley, L. V., Corps, R. E., & Pickering, M. J. (1768). The effects of dual-task interference in predicting turn-ends in speech and music. *Brain Research, 2021*, 147571.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *J. Phon., 38*, 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review, 25*(5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z

Hove, M. J., Fairhurst, M. T., Kotz, S. A., & Keller, P. E. (2013). Synchronizing with auditory and visual rhythms: An fMRI assessment of modality differences and modality appropriateness. *NeuroImage, 67*, 313–321. https://doi.org/10.1016/j.neuroimage.2012.11.032

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Iversen, J. R., Patel, A. D., Nicodemus, B., & Emmorey, K. (2015). Synchronization to auditory and visual rhythms in hearing and deaf individuals. *Cognition, 134*, 232–244. https://doi.org/10.1016/j.cognition.2014.10.018

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). Switchboard SWBD-DAMSL Shallowdiscourse-Function Annotation Coders Manual. *Institute of Cognitive Science Technical Report. Boulder*, 97–102.

Keller, P. E., Novembre, G., & Hove, M. J. (2014). Rhythm in joint action: Psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1658), 20130394. https://doi.org/10.1098/rstb.2013.0394

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes, 52*(4), 255–289. https://doi.org/10.1080/0163853X.2014.955997

Knudsen, B., Creemers, A., & Meyer, A. S. (2020). Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.593671, 593671. Nov 6.

Konvalinka, I., Vuust, P., Roepstorff, A., & Frith, C. D. (2010). Follow you, follow me: Continuous mutual prediction and adaptation in joint tapping. *Quarterly Journal of Experimental Psychology, 63*(11), 2220–2230. https://doi.org/10.1080/17470218.2010.497843

Large, E. W., Herrera, J. A., & Velasco, M. J. (2015). Neural networks for beat perception in musical rhythm. *Frontiers in Systems Neuroscience, 9*, 159. https://doi.org/10.3389/fnsys.2015.00159

Levinson, S. C. (2016). Turn-taking in human communication—Origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00731

Local, J., & Walker, G. (2012). How phonetic features project more talk. *Journal of the International Phonetic Association, 42*(3), 255–280. https://doi.org/10.1017/S0025100312000187

Nowicki, L., Prinz, W., Grosjean, M., Repp, B. H., & Keller, P. E. (2013). Mutual adaptive timing in interpersonal action coordination. *Psychomusicology: Music, Mind, and Brain, 23*(1), 6–20. https://doi.org/10.1037/a0032039

Okobi, D. E., Banerjee, A., Matheson, A. M., Phelps, S. M., & Long, M. A. (2019). Motor cortical control of vocal interaction in neotropical singing mice. *Science, 363*(6430), 983–988.

Pecenka, N., & Keller, P. E. (2011). The role of temporal prediction abilities in interpersonal sensorimotor synchronization. *Experimental Brain Research, 211*(3), 505–515. https://doi.org/10.1007/s00221-011-2616-0

Phillips-Silver, J., & Keller, P. (2012). Searching for roots of entrainment and joint action in early musical interactions. *Frontiers in Human Neuroscience, 6*, 26. https://doi.org/10.3389/fnhum.2012.00026

Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J., Kello, C., … Wiggins, G. A. (2021). Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B, 376*(1835), 20200334. https://doi.org/10.1098/rstb.2020.0334

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review, 12*(6), 969–992. https://doi.org/10.3758/BF03206433

Repp, B. H., & Keller, P. E. (2008). Sensorimotor synchronization with adaptively timed sequences. *Human Movement Science, 27*(3), 423–456. https://doi.org/10.1016/j.humov.2008.02.016

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00509

Roman, I. R., Washburn, A., Large, E. W., Chafe, C., & Fujioka, T. (2019). Delayed feedback embedded in perception-action coordination cycles results in anticipation behavior during synchronized rhythmic action: A dynamical systems approach. *PLoS Computational Biology, 15*(10), Article e1007371. https://doi.org/10.1371/journal.pcbi.1007371

de Ruiter, J.-P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*(3), 515–535. https://doi.org/10.1353/lan.2006.0130

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics, 11*(3), 243–257. https://doi.org/10.1016/S0095-4470(19)30825-3

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk, 71*, 93.

Schöner, G. (1994). From interlimb coordination to trajectory formation: Common dynamical principles. In S. Swinnen, J. Massion, H. Heuer, & P. Caesar (Eds.), *Interlimb Coordination* (pp. 339–368). https://doi.org/10.1016/B978-0-12-679270-6.50022-X

Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition, 48*(1), 31–51. https://doi.org/10.1006/brcg.2001.1302

Knoblich, G., & Sebanz, N. (2008). Evolving intentions for social interaction: from entrainment to joint action. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1499), 2021–2031.

Semjen, A., Schulze, H. H., & Vorberg, D. (2000). Timing precision in continuation and synchronization tapping. *Psychological Research, 63*(2), 137–147. https://doi.org/10.1007/pl00008172

Stepp, N., & Turvey, M. T. (2010). On strong anticipation. *Cognitive Systems Research, 11*(2), 148–164. https://doi.org/10.1016/j.cogsys.2009.03.003

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences, 106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Takahashi, D. Y., Narayanan, D. Z., & Ghazanfar, A. A. (2013). Coupled oscillator dynamics of vocal turn-taking in monkeys. *Current Biology, 23*(21), 2162–2168. https://doi.org/10.1016/j.cub.2013.09.005

Ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication, 47*(1), 80–86. https://doi.org/10.1016/j.specom.2005.05.009

Tognoli, E., Zhang, M., Fuchs, A., Beetle, C., & Kelso, J. S. (2020). Coordination dynamics: A foundation for understanding social behavior. *Frontiers in Human Neuroscience, 14*. https://doi.org/10.3389/fnhum.2020.00317

de Vos, C., Torreira, F., & Levinson, S. C. (2015). Turn-timing in signed conversations: Coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00268

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*(6), 957–968. https://doi.org/10.3758/BF03206432

Wilson, T. P., & Zimmerman, D. H. (1986). The structure of silence between turns in two-party conversation. *Discourse Processes, 9*, 375–390. https://doi.org/10.1080/01638538609544649

Wing, A. M. (2002). Voluntary timing and brain function: An information processing approach. *Brain and Cognition, 48*(1), 7–30. https://doi.org/10.1006/brcg.2001.1301

Yngve, V. H. (1970). On getting a word in edgewise. In M. A. Campbell, et al. (Eds.), *Papers from the sixth regional meeting of the Chicago linguistic society* (pp. 567–578). Chicago, IL, USA: Chicago Linguistic Society.