languages

MDPI

Article

Speech Rate and Turn-Transition Pause Duration in Dutch and English Spontaneous Question-Answer Sequences

Damar Hoogland [1,*], Laurence White [1,*] and Sarah Knight [2]

[1] School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
[2] Department of Psychology, University of York, York YO10 5DD, UK; sarah.knight3@york.ac.uk
* Correspondence: d.hoogland2@newcastle.ac.uk (D.H.); laurence.white@newcastle.ac.uk (L.W.)

Abstract: The duration of inter-speaker pauses is a pragmatically salient aspect of conversation that is affected by linguistic and non-linguistic context. Theories of conversational turn-taking imply that, due to listener entrainment to the flow of syllables, a higher speech rate will be associated with shorter turn-transition times (TTT). Previous studies have found conflicting evidence, however, some of which may be due to methodological differences. In order to test the relationship between speech rate and TTT, and how this may be modulated by other dialogue factors, we used question-answer sequences from spontaneous conversational corpora in Dutch and English. As utterance-final lengthening is a local cue to turn endings, we also examined the impact of utterance-final syllable rhyme duration on TTT. Using mixed-effect linear regression models, we observed evidence for a positive relationship between speech rate and TTT: thus, a higher speech rate is associated with longer TTT, contrary to most theoretical predictions. Moreover, for answers following a pause ("gaps") there was a marginal interaction between speech rate and final rhyme duration, such that relatively long final rhymes are associated with shorter TTT when foregoing speech rate is high. We also found evidence that polar (yes/no) questions are responded to with shorter TTT than open questions, and that direct answers have shorter TTT than responses that do not directly answer the questions. Moreover, the effect of speech rate on TTT was modulated by question type. We found no predictors of the (negative) TTT for answers that overlap with the foregoing questions. Overall, these observations suggest that TTT is governed by multiple dialogue factors, potentially including the salience of utterance-final timing cues. Contrary to some theoretical accounts, there is no strong evidence that higher speech rates are consistently associated with shorter TTT.

Keywords: conversation; dialogue; turn-taking; turn-transition time; speech rate; final lengthening; turn timing

Citation: Hoogland, Damar, Laurence White, and Sarah Knight. 2023. Speech Rate and Turn-Transition Pause Duration in Dutch and English Spontaneous Question-Answer Sequences. *Languages* 8: 115. https://doi.org/10.3390/languages8020115

Academic Editors: Jürgen Trouvain and Bernd Möbius

Received: 21 November 2022
Revised: 31 March 2023
Accepted: 4 April 2023
Published: 22 April 2023

1. Introduction

The length of time it takes interlocutors to respond to each other in conversation is a salient and meaningful aspect of spoken interaction. Thus, listeners change their interpretation of an exchange or expectation of the answer depending on response latency (Bögels et al. 2020; Kendrick and Torreira 2015; Roberts et al. 2011; Roberts and Francis 2013). Furthermore, Templeton et al. (2022) reported that pauses between conversationalists' turns are shorter when they report feeling a closer 'click'.

The relationship between turn-transition time (TTT) and pragmatic meaning is not straightforward, however, as TTT has also been held to be affected by many other factors. These include linguistic and discourse-related features of the turn switch (Roberts et al. 2015), perceptual and cognitive processing constraints (Levinson and Torreira 2015) and—potentially—the language in which the conversation is spoken (Stivers et al. 2009).

Notwithstanding relatively small variations due to meaning or context, TTT is generally short. It is widely reported that most TTTs during conversations are around 200 ms

(e.g., Heldner and Edlund 2010). Such short response times are considered a challenge for theories of speech processing and comprehension, firstly because it is argued that 200 ms is shorter than typical verbal reaction times of 400 ms (Wesseling and van Son 2005; cited in Heldner and Edlund 2010) and, relatedly, because they require parallel production and comprehension processes (Garrod and Pickering 2015).

Our study examined the effects of speech rate and another timing factor—localised utterance-final lengthening—on the variation in turn-transition timing in two languages, Dutch and English. Specifically, we tested for relationships between TTT and (a) speech rate and (b) turn-final-syllable rhyme duration, using question-answer sequences taken from corpora of spontaneous dyadic conversations in Dutch and English. As discussed below, there are theoretical and empirical reasons to predict that both speech rate and final lengthening will influence TTT.

1.1. Speech Rate Entrainment and the Perception–Action Link in Speech

The idea that speech rate might affect conversational timing was considered from an anthropological perspective by Lehtonen and Sajavaara (1985) and developed, from a neuroscience perspective, in theoretical proposals on the mechanisms underpinning turn-taking (Garrod and Pickering, 2015; Wilson and Wilson 2005).

Wilson and Wilson (2005)'s proposal synthesized several observations. First, they suggested that the typical TTT is an integer multiple of mean syllable duration in conversational speech (the latter being approximately 200 ms; Wilson and Zimmerman 1986). Second, short TTT is common (based on a reanalysis of Wilson and Zimmerman 1986), suggesting that speakers often do not wait for the end of their interlocutor's turn to initiate a response, even when they do not intend to interrupt. Third, speaking rates of interlocutors have been observed to converge during conversations (Street 1984; Manson et al. 2013). Fourth, oscillatory neural activity, as measured by electroencephalography (EEG), has been implicated both in temporal perception and in cognitive processing more generally (Burle and Bonnet 1999, and others, cited in Wilson and Wilson 2005), and has since been shown to entrain to regular auditory stimuli. Such neural entrainment is also found for the quasi-regular flow of speech at approximately the timescale of syllables, and this entrainment has, in turn, been linked to speech processing (Luo and Poeppel 2007; Ding and Simon 2014).

From these observations, Wilson and Wilson (2005) hypothesized that the auditory cortex tracks the rate of an interlocutor's speech through entrained activity. They further suggest that neural oscillations, thus entrained, promote a cyclical increase and decrease in the likelihood of initiating speech. A clear prediction arising from this proposal is that the speech rate of a conversational turn will tend to influence the subsequent TTT.

Similarly, Garrod and Pickering (2015) suggested that listeners integrate predictions of the content of interlocutor's speech with a prediction about its timing. Like Wilson and Wilson (2005), they invoked auditory neural entrainment to speech rate, suggesting that entrained oscillatory activity in the auditory cortex is itself linked to motor areas in the brain, through a functional perception–action link. As with Wilson and Wilson (2005), a key implication of Garrod and Pickering's proposal is that higher speech rate promotes shorter response times.

Garrod and Pickering (2015) focused on functional pathways of control in the brain, whilst Wilson and Wilson (2005) proposed that fluctuations in the EEG data reflect a likelihood function of speech initiation. In essence, however, these two claims have the same implication for TTT: through a perception–action link, entrained neural activity in the auditory cortex causes an inverse relationship between speech rate and response time (i.e., higher speech rate is associated with shorter TTT). As a point of clarification about the specific empirical prediction arising from both these theoretical proposals (Garrod and Pickering 2015; Wilson and Wilson 2005), it is important for the interpretation of this and previous studies to note that the direction of the association between speech rate and TTT depends on the units used. In particular, Wilson and Wilson discussed speech rate in terms of average syllable duration (unit time per syllable), and thus predicted a *positive* association

between rate and turn-transition time (TTT). Most researchers, including ourselves, follow the more conventional approach of reporting speech rate as syllables per second, and thus the same prediction would be evidenced by a *negative* association between rate and TTT.

Note, also, that this perception–action link is not the only way neural tracking of speech rate could play a role in turn timing. Thus, if listeners use speech rate to predict the time of the end of a turn, but that prediction does not directly control their speech initiation, then we would not expect a relationship between speech rate and TTT.

Studies that empirically test the behavioural implications of these theories have come to contradictory conclusions. First, Beňuš (2009) tested the prediction that TTT should correlate with the speech rate of the previous turn in a corpus of American English task-based dialogues. Using automatically extracted measurements of speech rate (lexically defined syllable count per second over the inter-pausal unit of speech activity before the turn switch) and TTT (operationalized as the automatically extracted time between offsets and onsets of successive stretches of speech activity by different speakers, without considering functional aspects of the turn transition), he found no evidence for the expected correlation, except for the specific case of interruptions initiated during a pause. However, when using the rate of pitch accents, Beňuš (2009) did find that a higher pitch accent rate is associated with shorter TTT (defined with respect to pitch accent location, rather than turn-offset/onset per se).

Stronger support for the entrainment-based proposal came from Corps et al. (2020), who found, in an experimental study, that English speakers had shorter response latencies to speeded "yes-no" questions than to non-speeded questions (in line with the predictions of Wilson and Wilson 2005, and Garrod and Pickering 2015). They also manipulated the duration of the final monosyllabic word independently from the rest of the question and found that a shorter final word was associated with shorter TTT. Response times in their study were measured from the start of the final word, however, and thus are confounded with final word duration. When measured from the end of the final word, participants responded with *longer TTT* after shorter final words (Corps et al. 2020), which could be reinterpreted as a longer final word duration promoting quicker responding. We return to the use of final lengthening as a turn-transition cue below.

Contrary to the primary finding of Corps et al. (2020), Roberts et al. (2015) reported that *faster* speech appears to result in *longer* TTT. They used a random forest analysis to identify and rank predictors of TTT during American English telephone conversations (not limited to questions). Similarly to Beňuš (2009), they operationalized TTT as automatically extracted intervals between stretches of speech activity from different speakers, and they measured speech rate as the difference in duration of the stretch of speech relative to the expected duration given the sum of the average phone duration in the corpus. Roberts et al. (2015) found that speech rate was the fourth-most-important predictor of TTT, after speech act, first-turn duration, and second-turn duration. As stated above, they also (impressionistically) reported a positive relationship between speech rate and TTT; thus, the faster the first turn, the longer the TTT, based on the mean TTT for the fastest speech rate quartile being 100 ms longer than the slowest speech rate quartiles (they note that this is the case when backchannels and short turns are excluded). Levinson and Torreira (2015) suggested that this positive relationship may be due to processing constraints: listeners take longer to process relatively fast speech, and thus respond later.

In line with the Wilson and Wilson (2005) prediction, Torreira and Bögels (2022) found that Dutch speaking participants, when asked to say *ja* ('yes') in response to a string of repetitions of the meaningless syllable /ma/, had a shorter response initiation time after a fast string (200 ms/syllable) than after a slow string (300 ms/syllable). However, they only did so when no prosodic turn-ending-type cues were present. When the strings contained a lengthened final syllable (1.5 times the duration of the other syllables) or a phrase-final pitch contour (as opposed to a flat pitch contour), a faster preceding syllable rate resulted in a longer response latency (cf. Roberts et al. 2015, for spontaneous conversation). Torreira and Bögels conclude that, whilst perception-motor entrainment may constrain turn-timing

behaviour, this effect is probably often overshadowed by other influences on turn timing, such as prosodic cues to turn endings.

In conclusion, these studies do not present a uniform picture of the relationship between speech rate and TTT. In the studies of spontaneous conversations, Beňuš (2009) reported either no relationship or a *negative* relationship, while Roberts et al. (2015) reported a *positive* relationship. The experimental study by Corps et al. (2020) was in line with theoretical predictions of a *negative* relationship between speech rate and TTT (Wilson and Wilson, 2005; Garrod and Pickering, 2015), while Torreira and Bögels (2022)'s data indicated that the predicted entrainment-based effect may be overridden by specific local cues to turn endings.

These conflicting results may relate to methodological differences. First, the contrasting findings of Corps et al. (2020) and Roberts et al. (2015) could reflect demand characteristics in the former's speeded response task, and Corps et al.'s controlled experimental design also potentially eliminates some factors—such as a diversity of turn-taking cues—that influence TTT in natural spontaneous speech. Second, the somewhat contrasting findings of Beňuš (2009) and Roberts et al. (2015)'s corpus-based studies may relate, in part, to the different types of turn transitions in their datasets, along with exploratory statistical procedures that did not examine, in detail, how speech rate effects may be modulated by other factors that influence TTT.

1.2. This Study

We examined factors that predict TTT in question-answer sequences for two corpora (one Dutch, one English) of spontaneous conversations. Our main aim was to clarify the apparently contradictory relationship between speech rate and TTT found by Beňuš (2009) and Roberts et al. (2015), and to test whether some other key dialogue factors may modulate that relationship.

Specifically, we aimed to test, for a constrained set of cases, whether faster speech rates are associated with shorter or longer TTT. We also tested the influence of final rhyme duration on TTT. Phrase-final lengthening (particularly of the final syllable rhyme) is widely observed across languages (e.g., Oller 1973; Berkovits 1994) and has been long established as a cue to an upcoming boundary, including the end of an utterance (e.g., Price et al. 1991; for a review see White 2014). In particular, we tested whether the question-final rhyme duration (cf. comparable to Corps et al.'s compressed vs. unaltered final words) was inversely related to TTT (in line with Torreira and Bögels 2022, and one interpretation of the experimental observations of Corps et al. 2020). We also tested whether there is an interaction between these two speech timing variables.

Roberts et al. (2015) reported several dialogue factors as affecting TTT. To control some sources of variance, we used more functionally constrained corpora than Beňuš (2009), specifically natural question-answer sequences (comparable to the constructed stimuli of Corps et al. 2020). To test the contribution of naturally varying dialogue factors that remain within our corpora, we assessed the influences on TTT using multiple linear regression rather than bivariate correlational analyses.

Note that the perception-motor entrainment theories by Wilson and Wilson (2005) and Garrod and Pickering (2015) imply that the relationship between speech rate and turn-transition times is cyclical, not linear. Using linear methods, as previous studies did, therefore relies on the assumption either that most turns are initiated at the first cycle of the entrained perception-motor system and that the statistical model is robust to noise produced by turns that are not, or that the underlying mechanism is not cyclical. We adopted this assumption, as the previous studies discussed above have done, in part because Beňuš (2009) found no evidence for a cyclical pattern in turn-timing data through the visual inspection of histograms. Finally, we chose to use a relatively small, hand-aligned dataset, to complement previous studies that used large, automatically extracted datasets (Beňuš 2009; Roberts et al. 2015). Thus, we aimed to ensure that our speech timing variables (speech rate, final rhyme duration, TTT) were measured as consistently and accurately as

possible. We note that automatically extracted turn-timing data based on broad acoustic definitions of turns and turn transitions often do not reflect the functional nature of the interaction as effectively as manually labelled data (Corps et al. 2022).

We also explored the influence, on TTT, of question length in terms of syllables for two reasons. Firstly, Roberts et al. (2015) found question length to be a relatively important predictor of TTT: thus, longer turns (and very short turns) have longer TTT relative to intermediate-length turns. Secondly, question length may interact with the influence of speech rate on TTT, since longer utterances allow listeners more time to entrain to the local speech rate.

We further considered how two functional variables affect TTT: firstly, whether the question is polar (i.e., can be answered with "yes"/"no"); secondly, whether the answer directly addresses the question or not. Question type was added to make our analysis more directly comparable to previous studies that focused on polar questions (Corps et al. 2020; Stivers et al. 2009; and to some extent Torreira and Bögels 2022). The relevance of the answer was added because it has been shown to affect TTT in question-answer sequences: specifically, non-direct answer responses are slower than responses judged to answer the question (Stivers et al. 2009). This also allowed us to compare the size of any contribution of speech rate to TTT (in question-answer sequences) relative to other TTT predictors already established.

Finally, we used two corpora from different languages, Dutch and English, and considered language as a factor in our exploratory analyses, noting that Stivers et al. (2009) found a small difference in the mean TTT of samples from these two languages (~127 ms). We are primarily interested here in whether any speech rate influence on TTT is modulated by language, with the caveats that a. Dutch and English are typologically related and prosodically/syntactically similar; b. any observed differences may relate to corpus-specific factors as much as to language per se.

1.3. Research Questions

Speech timing factors: hypothesis-testing analyses

1. Is there an inverse relationship between Speech Rate and TTT?
2. Is there an inverse relationship between Final Rhyme Duration and TTT?

Additional factors: exploratory analyses

3. Are any observed effects of speech timing factors modulated by dialogue factors already shown to influence TTT (Question Length, Question Type, Answer Type)?
4. Do we observe any reliable differences, in the above factors between languages (Dutch and English)?

2. Materials and Methods

2.1. Corpora

Question-answer sequences were extracted from corpora of spontaneous conversations in Dutch (Van Son et al. 2008) and English (White et al. 2012). As described below, the corpora were both comprised of free conversations between people who were either friends or colleagues in university environments.

The Dutch dataset was sampled from nine (out of 20) conversations of the *Instituut voor Fonetiek Amsterdam Dialogue Video Corpus* (IFADV, Van Son et al. 2008). This corpus consists of dyadic lab-based conversations, and each published conversation is a 15 min continuous sample (stereo WAV format file) from a longer conversation. The published corpus includes Praat TextGrids (Boersma and Weenink 2021) annotated for dialogue act, gaze direction, orthographic transcriptions, and automatic word-level and phoneme-level alignments. Only the WAV files and the TextGrid tiers for orthographic transcription, dialogue act and phoneme alignment were used in this study. The parallel video recordings were not used.

The English dataset was sampled from a corpus of eight dyadic lab-based conversations (White et al. 2012, which also includes read sentences and map task speech, not analysed here). Each conversational recording was between 16 and 22 min long. The corpus includes Praat TextGrids with one tier per speaker containing orthographic transcriptions, along with codes for overlapped speech, silent and filled pauses, and turn-exchange information based on speech activity. Only the orthographic transcriptions were used in this study.

The 18 Dutch speakers (14 female) ranged in age from 18 to 65 years (M = 34 years) and spoke standard Netherlands Dutch with various regional accents. All pairs were described as friends or colleagues.

The 16 English speakers (10 female) of Standard Southern British English (SSBE) were students at the University of Bristol aged between 18 and 30 years old. They were (at minimum) acquainted with each other through participating in the corpus recordings, although some appeared to know each other prior to the recording.

All pairs of speakers were recorded while facing each other across a table and wearing headset microphones. For the Dutch speakers, video cameras were positioned behind each speaker, directed at the opposite speaker. Both sets of conversations featured some use of suggested topics for discussion.

2.2. Sampling of Question-Answer Sequences

For each corpus, question-answer sequences were identified by the first author (a native Dutch speaker also fluent in English). In Dutch, the speech act annotations and orthographic transcriptions were also consulted to identify questions, but the author's judgement was the criterion for inclusion. In the English corpus, the point at which questions and answers were delimited was matched to the boundaries already marked in the corpus transcription.

Questions were identified and delimited by their syntactic and/or prosodic structure, and by their apparent response-evoking purpose. We included utterances with the syntactic/prosodic form of a question, but pragmatically appearing to be a request for a statement of agreement, rather than asking for information. We excluded utterances with the form of a question that did not appear to be intended to prompt a response: these utterances included questioning backchannels (e.g., "oh really?") that were followed by a continuation by the speaker rather than a response, and questions that were obviously reported speech, e.g., [schematic example]: "And then she asked: 'Did you see her again?' To which I said . . . ".

2.3. Speech and Language Measurements

For each question-answer sequence, we measured and extracted a number of continuous and categorical variables, as outlined here.

Turn-transition time (TTT): TTT was operationalized as the duration of the interval between the end of the question and the start of the answer (in milliseconds). TTT is thus positive when there is a pause between the question and the answer (silent inter-speaker pauses are usually referred to as gaps, and we will follow this convention), and negative when the answer begins before the question turn ends (these cases are referred to here as overlaps). Where the first speaker asks two questions in a row and the second speaker answers the first question, the speech signals may be overlapped, but TTT can nonetheless be positive (being taken from the end of the first question). In the latter regard, our annotation of turns differs from some of the automatically labelled corpora discussed earlier (e.g., Roberts et al. 2015).

Speech rate: Speech rate (syllables per second) was derived from the number of acoustically realized vowels from the start of the question to the start of the final-syllable rhyme, divided by the duration of that stretch of speech. We use this operationalization of speech rate, rather than, e.g., counting phonologically defined syllables, to ensure that it was consistently applicable to speech data from two different languages by a single

annotator (in this case, a native speaker of Dutch, fluent but not native in English). We note that this is a more conventional and direct measure of speech rate than that used in some previous tests of the relationship between speech rate and TTT (e.g., Roberts et al. 2015, see above).

Notwithstanding pre-existing automated phone-level segmentation (for the Dutch corpus), final decisions about vowel identification, along with duration measurements, were carried out manually using Praat's default spectrogram settings (a view range of 0–5 kHz, window length of 5 ms, dynamic range of 70 dB with a visual inspection window between 450 and 600 ms). Vowels were required to have clear formant structure and glottal striations (Lin and Wang 2007), but not necessarily voicing (although the great majority of vowels were, of course, voiced). Consecutive (heterosyllabic) vowels were separated corresponding to the change in formant structure, a dip in the amplitude envelope, and/or glottalization. Nasalized vowels or syllabic nasals/approximants were counted as a syllabic unit in our speech rate measurement. One English speaker often whispered: here we used changes in the formant structure, supported by auditory impressions, to decide on the presence of vowels.

Unless the question contained a pause of over 1000 ms (see Section 2.4), pauses were included in the question duration used to define speech rate. We included short pauses because the speech rate appears to predict patterns of neural entrainment better than articulation rate, the latter being syllable rate with pauses removed (e.g., see Kayser et al. 2015, with regard to neural entrainment to theta-band fluctuations in the amplitude envelope of speech; see also Bosker 2017, for non-speech stimuli). In overlapped question-answer sequences, the final vowel that started before the overlap was considered the nucleus of the final syllable (see next paragraph), and thus the speech rate was based on the number of syllables up to and including the preceding syllable.

Final rhyme duration: When there was a gap between question offset and answer onset (i.e., TTT was positive), we measured final syllable rhyme duration from the final vowel onset to the offset of speech (which often equates to the end of periodicity) using guidelines adapted from White and Mattys (2007). Initial and final boundaries were placed at zero-crossing of the waveform. Vowel onsets were taken from the beginning of the first pitch period consistent with the shape of the following vowel and at the onset of the second formant: the two criteria typically aligned, with priority given to formant structure where they conflicted. After nasals, laterals and glides, the vowel onset was placed at a point of maximum change in format structure and/or a minimum in the amplitude envelope. If it was not possible to differentiate the prevocalic nasal or glide from the subsequent vowel, it was included in the vocalic interval. As in our speech rate measurement, nasalized vowels or syllabic nasals/approximants were counted as a syllabic unit.

Question length: Question length was measured as the number of syllables in the question, up to but not including, for gaps, the final syllable and for overlaps, the final syllable whose vowel started before the overlap. Thus, question length in syllables was defined over the same domain as speech rate (see above).

Functional factors: Each question-answer sequence was also categorised (by the first author) regarding two functional properties. Firstly, questions were classified for Question Type as polar or not, where polar questions are those that can be appropriately responded to with 'yes' or 'no' (*ja* or *nee* in Dutch). Secondly, questions were classified for their Answer Type as "Answered" or "Not answered", according to the first author's judgement about the relevance of the response to the preceding question. Examples of questions classified thus, from the English dataset, are shown below (1–2). This variable does not correspond precisely to the Conversation Analytic notion of "preference" (Bilmes 1988), since responses that are coded as answers can nevertheless still indicate dis-preferred answers.

1. Answered

   Q: Well how about yourself then all set for Christmas then?
   A: Not very set.

2. Not answered

   Q: Well [wu]d wha's your what are you doing your M.A. in?
   A: Eh MSc thank you very much.

Language: We included the language of the interaction (Dutch vs. English) in our exploratory analyses. We did not, however, make any strong interpretation about whether any observed differences are due to language per se. The two languages are closely related, with strong similarities in terms of syntax and prosody, and it is possible that observed differences could be due to other factors, such as the nature of the corpus and the relationships between speakers (noting, for example, that interlocutors in the Dutch corpus had pre-existing familiarity with each other, but that this was inconsistent in the English corpus).

2.4. Statistical Analysis

Question-answer sequences with questions shorter than four syllables and longer than 25 syllables (prior to the final syllable in the case of positive TTT, and prior to the interrupted syllable in the case of negative TTT) were excluded from the analyses. We also removed sequences with questions containing a silent intra-speaker pause longer than 1000 ms. Both of these steps were taken, in part, to eliminate speech rate anomalies (typically, excessively low rate values) arising due to very short utterances or from the presence of extended pauses. In addition, listener entrainment to heard speech takes time to build up (Luo and Poeppel 2007) and is likely to be disrupted or reset after extended gaps (e.g., Coffey et al. 2021; Van Bree et al. 2021; Xu and Ye 2015).

After excluding these question-answer sequences, we further excluded all sequences with TTT above or below 2.5 standard deviations from the mean (calculated across the full remaining dataset).

We used mixed-effect linear regression models to determine the predictors of TTT in our question-answer sequences. We firstly analysed the full dataset together, with both overlapped turns and gaps included (i.e., negative and positive TTT) and, additionally, carried out separate analyses for the Gaps and Overlaps datasets (where overlaps were defined as TTT < 0 ms and gaps as TTT >= 0 ms). A practical reason for separating these was to include final rhyme duration as a predictor for Gaps; this interval was not reliably measurable for Overlaps, given the arbitrariness of the overlap point with respect to the ongoing speech (see Heldner and Edlund 2010).

For each dataset (Full, Gaps, Overlaps), we divided our analysis procedure into two phases. In Phase 1, we tested our primary research questions concerning the relationship between TTT and Speech Rate. For Gaps, we also tested the influence of Final Rhyme Duration and the interaction of Speech Rate × Final Rhyme Duration. In Phase 2, taking our final regression models from Phase 1 as the base models, we explored the additional contributions of the functional language factors Question Type and Answer Type, along with Question Length and Language.

The contributions of all predictors and interactions were assessed using model comparisons, i.e., comparing models with and without the relevant factor using likelihood ratio tests (LRTs). We included a random intercept for Speaker in all models. We used a backward stepwise approach for the hypothesis-testing Phase 1 analyses. The initial model contained all possible fixed effects and their interactions (for the Phase 1 predictors Speech Rate and Final Rhyme Duration). These fixed effects were then tested for their contribution to model fit using LRTs, progressing from the most complex level (i.e., 2-way interactions) to the least complex. Effects which did not at least show a trend towards statistical significance (i.e., $p > 0.1$) were removed. The final resulting model was then used as the baseline for the exploratory analyses in Phase 2.

The Phase 2 analysis used a forward stepwise procedure. Each exploratory factor was added, in turn, to the baseline model from Phase 1. At each step, we retained the predictor that produced the largest reduction in AIC value (Akaike 1973) and then constructed further

models to test for, firstly, the additional contribution of the other exploratory predictors and, finally, the interactions between the predictors.

In a final step, LRTs were used to assess the contribution of each predictor to the completed model generated at the end of Phase 2 (i.e., each predictor was dropped in turn from the final model and this reduced model was compared to the full final model). The LRT (χ^2) statistics reported in the main body of the text reflect the contribution of the relevant factor at the time it was retained or dropped from the model. The LRTs reported in the tables reflect the contribution of each factor to the final models (i.e., after the iterative elimination of all non-significant factors).

All mixed-effects linear regression analyses were carried out using the *lme4* package (Bates et al. 2015) in the statistical program *R* (R Core Team 2022).

3. Results

3.1. Descriptive Statistics

We collected and measured 568 question-answer sequences (326 Dutch; 242 English). In both corpora, gaps were more frequent than overlaps (Table 1). The number of question-answer sequences per (question) speaker ranged from two to forty-two (mean: seventeen sequences per speaker). TTTs had a near-normal distribution, as expected (Figure 1; Table 2). After removing TTT outliers, very short questions, and questions containing pauses (see Section 2.4 for exclusion criteria), 413 of 568 observations remained. All subsequent data and analyses are based on the datasets after applying the exclusion criteria.

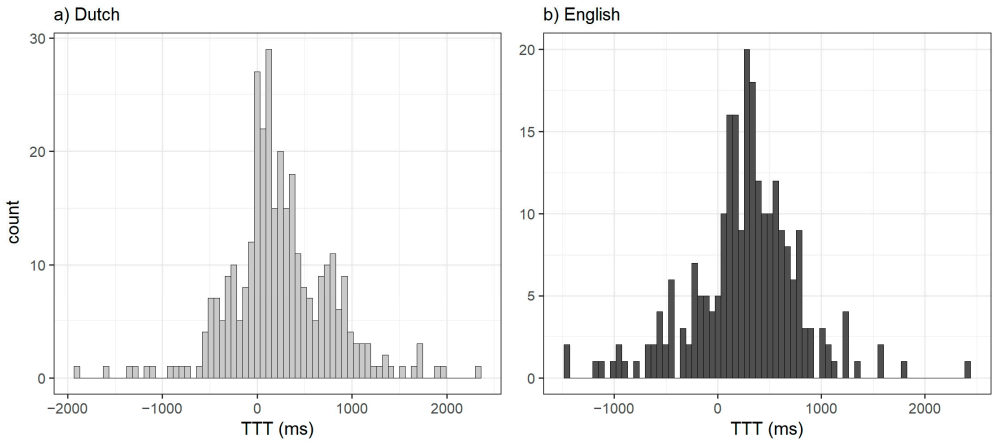

Figure 1. TTT distribution for the (**a**) Dutch and (**b**) English corpora (before applying exclusion criteria) in milliseconds based on 70 bins (bin size: Dutch 36 ms; English 35 ms). TTT distributions after applying exclusion criteria are in Appendix A, Figure A1.

Table 1. Number of question-answer sequences by transition type (gap vs. overlap) and language. Counts are after applying exclusion criteria (counts for the original dataset in brackets).

Transition Type	Both Languages	English	Dutch
Gaps	304 (428)	145 (189)	159 (239)
Overlaps	109 (140)	38 (53)	71 (87)
Total	**413 (568)**	**183 (242)**	**230 (326)**

Means for the speech timing variables are listed in Table 3, along with mean question length. The mean speech rate was higher in the Dutch corpus by ~0.7 syllables/s, $t(393.94) = 6.084$, $p < 0.001$. In line with this, mean final syllable rhyme duration was shorter in the Dutch corpus by ~28 ms, $t(323.47) = -2.148$, $p = 0.032$. The mean question length in the Dutch corpus was also higher by ~0.8 syllables: $t(405.99) = 1.8$, $p = 0.073$.

Table 2. Turn-transition time (TTT) central tendency measures (ms). Figures are for the datasets after applying exclusion criteria (figures for the original dataset in brackets). Highest density based on mode of 70 equal-duration bins, from lowest to highest TTT, and midpoint of observations within that bin. Bin size: Dutch 36 ms; English 35 ms; both languages 37 ms (original dataset bin size: Dutch 61 ms; English 56 ms; both languages 62 ms).

Corpus	Mean TTT in ms	Highest Density in ms
Both languages	218 (252)	74 (124)
English	242 (264)	279 (281)
Dutch	199 (244)	0 (123)

Table 3. Mean question length (in syllables, excluding final syllable), mean speech rate (syllables/s), mean final rhyme duration (ms). Figures are for the datasets after applying exclusion criteria over the full dataset, with the figures for the original dataset in brackets.

Corpus	Question Length (No. of Syllables)	Speech Rate (Syllables/s)	Final Rhyme Duration (ms)
Both languages	9.5 (8.3)	5.3 (5.0)	196 (200)
English	9.0 (8.1)	4.9 (4.7)	212 (212)
Dutch	9.8 (8.4)	5.6 (5.2)	183 (190)

3.2. Inferential Statistics

Analyses of predictors of TTT in these data are presented for the full dataset first, and then separately for Gaps and Overlaps.

3.2.1. Full Dataset

In the Phase 1 hypothesis testing, we compared a model predicting TTT from Speech Rate to a null (intercept-only) model, which showed that Speech Rate significantly improved the model fit, $\chi^2(1) = 3.91$, $p = 0.048$. Contrary to predictions based on most theoretical approaches (e.g., Wilson and Wilson 2005), Speech Rate was *positively* associated with TTT, i.e., faster questions were associated with longer TTT preceding the answers.

We took this model as the base for the Phase 2 exploratory analyses, which tested the additional contributions of Question Type (polar vs. non-polar) and Answer Type ("answered" vs. "not answered") as well as Language and Question Length. Following the iterative forwards procedure described above, we first added Answer Type (to the Speech Rate base model), $\chi^2(1) = 10.93$, $p < 0.001$, finding that responses which were judged not to have answered the question were slower than those which did.

The functional factor Question Type additionally contributed to predicting TTT, $\chi^2(1) = 9.47$, $p = 0.002$, with polar (yes-no) questions answered more quickly than non-polar questions.

Finally, Question Length was marginally significant, $\chi^2(1) = 3.20$, $p = 0.074$, indicating a statistical trend towards longer questions being answered more quickly than shorter questions.

LRTs indicated that Language did not contribute significantly to model fit at any stage, so it was not included.

We then tested all two-way interactions between the added exploratory variables and between the exploratory variables and Speech Rate. Only the interaction between Speech Rate and Question Type was significant, $\chi^2(1) = 7.14$, $p = 0.008$: faster polar questions were associated with longer TTT, while faster non-polar questions were associated with shorter TTT.

Thus, the final linear regression model for the full dataset included the following predictors of TTT: Speech Rate; Question Type; Answer Type; Question Length (marginal); and Question Type x Speech Rate interaction.

After adding the exploratory predictors, we retested all predictors by comparing the final model to a model with each of these predictors removed in turn. The effect of Speech

Rate no longer reached significance: Speech Rate, $\chi^2(1) = 2.38$, $p = 0.12$. Table 4 lists the LRT statistics and the AIC for each of the reduced models with one variable removed and compared to the final model, as a measure of relative effect size for each predictor. We note that Answer Type most reduced the AIC, closely followed by Question Type.

Table 4. LRTs comparing the final model for the full dataset and models with each of the predictors removed, as well as a summary of each predictor's effect (*marginal effects in italics*). N/A = non-significant effect.

Model	AIC	AIC Difference vs. Final Model	χ^2	*p*-Value	Effect on Turn-Transition Time (TTT)
Final	6204.0				
- Speech Rate	6204.3	+0.3	2.38	0.12	N/A
- Answer Type	6213.5	+9.5	11.49	<0.001	"Answer" responses <=> shorter TTT
- Question Type	6212.9	+8.9	10.96	<0.001	Polar questions <=> shorter TTT
- Question Length	6205.3	+1.3	3.36	*0.07*	*Longer questions <=> shorter TTT*
- Question Type × Speech Rate	6209.1	+5.1	7.14	<0.001	Polar questions: higher Speech Rate <=> longer TTT Non-polar questions: higher Speech Rate <=> shorter TTT

3.2.2. Gaps Only

For the Gaps dataset, we entered Speech Rate and Final Rhyme Duration as primary predictors in Phase 1, along with their interaction. The two speech timing variables were, unsurprisingly, somewhat correlated (Pearson's $r = -0.184$, $t(302) = -3.262$, $p = 0.001$) and so were Speech Rate and Question Length (Pearson's $r = -0.114$, $t(302) = -1.994$, $p = 0.047$), but the Variance Inflation Factors for a model with Speech Rate, Final Rhyme Duration and Question Length were below 1.04, so there was no adjustment.

In the Phase 1 backward regression model comparisons for Gaps, the Speech Rate x Final Rhyme Duration interaction made a marginally significant contribution to the model fit, $\chi^2(1) = 3.42$, $p = 0.064$ (i.e., it is a trend—see below). In line with our analysis protocol (see above), we therefore retained it in the model. The main effect of Speech Rate was significant: $\chi^2(1) = 3.97$, $p = 0.046$: overall, a higher Speech Rate was associated with longer TTT (as in the Phase 1 analysis for the full dataset). There was also a main effect of Final Rhyme Duration: $\chi^2(1) = 4.84$, $p = 0.028$, with longer Final Rhyme durations associated with longer TTT.

For illustration, Figure 2 explores the interaction of Speech Rate x Final Rhyme Duration via median splits for both timing predictors (although note that the interaction effects reported here were based on continuous data). From the perspective of a Speech Rate median split (Figure 2a), at low speech rates, longer final rhymes were associated with longer TTT (left panel), but at high rates, longer rhymes were associated with shorter TTT (right panel). Considering the same effect via a Final Rhyme Duration median split, as shown in Figure 2b, when final rhymes were long, a higher speech rate was associated with shorter TTT (left panel), but when final rhymes were short, a higher speech rate was associated with longer TTT (right panel). From both perspectives, relatively long final rhymes (long rhyme/fast preceding speech) seem to serve as effective local cues to the turn end, thus promoting faster responding. These interactions are discussed further below.

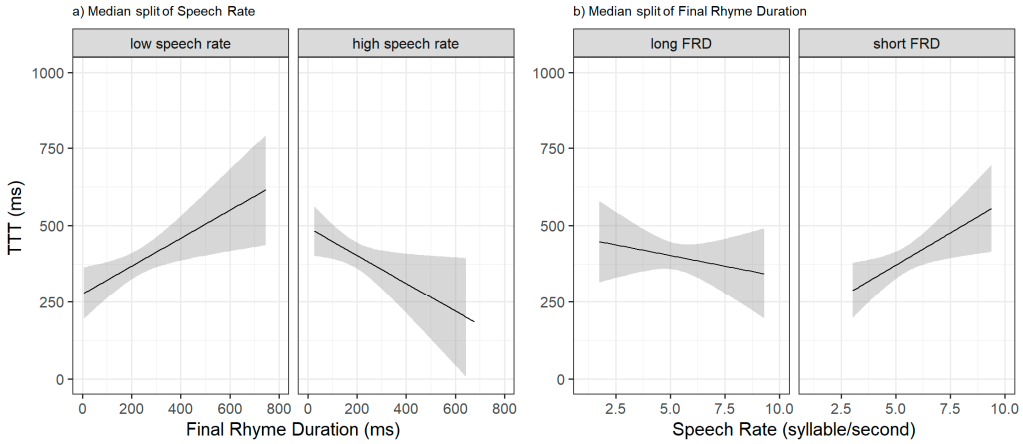

Figure 2. Interactions between Speech Rate and Final Rhyme Duration by median split of Speech Rate (**a**) and median split of Final Rhyme Duration (**b**). The statistical interaction derives from continuous data: the median splits are for illustration.

In Phase 2, we added the exploratory variables to the baseline model (Speech Rate + Final Rhyme Duration + Speech Rate x Final Rhyme Duration). Question Type was added first to the baseline model, $\chi^2(1) = 16.83$, $p < 0.001$; as in the full dataset, polar (yes-no) questions were answered more quickly than non-polar questions.

Adding Answer Type led to a singular fit, indicating overfitting; it was therefore not included. LRTs indicated that Language and Question Length did not contribute significantly to model fit at any stage, so they were also not included. Further LRTs indicated no significant interactions between Question Type and any of the other existing predictors.

Our final model for the Gaps dataset therefore included Speech Rate + Final Rhyme Duration + Speech Rate x Final Rhyme Duration + Question Type. We then retested each of the predictors in a backwards manner, by comparing the final model against a model with that predictor removed. As indicated by Table 5, the Phase 1 predictors no longer reached significance in this final model: Speech Rate, $\chi^2(1) = 2.60$, $p = 0.11$, Final Rhyme Duration: $\chi^2(1) = 2.88$, $p = 0.090$, Speech Rate x Final Rhyme Duration: $\chi^2(1) = 1.94$, $p = 0.164$.

Table 5. LRTs comparing the final model for the Gaps dataset and models with each of the predictors removed, as well as a summary of each predictor's effect (*marginal effects in italics*). N/A = non-significant effect.

Model	AIC	AIC Difference vs. Final Model	χ^2	*p*-Value	Effect on Turn-Transition Time (TTT)
Final	4306.82				
- Speech Rate	4307.42	+0.6	2.60	0.11	N/A
- Final Rhyme Duration	4307.70	+0.9	2.88	*0.09*	*Longer Final Rhyme<=>longer TTT*
- Speech Rate x Final Rhyme Duration	4306.76	−0.1	1.94	0.16	N/A
- Question Type	4321.65	+14.8	16.83	<0.001	Polar questions <=> shorter TTT

3.2.3. Overlaps Only

For the Overlaps dataset, there was no main effect of Speech Rate: $\chi^2(1) = 0.77$, $p = 0.38$. Adding each one of the functional language predictors to an intercept-only model did not improve the model fit: Question Type: $\chi^2(1) = 0.03$, $p = 0.87$; Answer Type: $\chi^2(1) = 0.67$, $p = 0.41$. There was no effect of Language: $\chi^2(1) = 0.60$, $p = 0.44$. There was a marginal effect of Question Length: $\chi^2(1) = 3.59$, $p = 0.058$ (which also slightly decreased the AIC: 1557.7 compared to 1559.2 for the null model). Thus, there was a trend towards longer questions being associated with shorter (i.e., more negative) TTT, as in the full dataset.

Thus, we found no strong evidence for any predictors of overlapped TTT, with a marginal influence of Question Length. This pattern of results suggests a qualitative difference with the Gaps dataset, as discussed further below.

4. Discussion

The primary goal of this study was to test whether higher speech rate is associated with shorter turn-transition time (TTT), specifically in spontaneous question-answer sequences. This negative relationship is predicted by theoretical accounts of conversational turn-taking by Wilson and Wilson (2005) and Garrod and Pickering (2015). It is supported by the experimental results of Corps et al. (2020), using a question-answering task, but has only weak, circumscribed support from the corpus study of Beňuš (2009) and is contradicted by the corpus analyses of Roberts et al. (2015). Our primary findings were in line with those of Roberts et al. (2015); thus, in cases where any relationship was found, faster speech in questions was associated with longer TTT before answers.

We also tested the effect on TTT of another speech timing factor, final syllable rhyme duration, and how this factor interacts with speech rate, as discussed below. Additional exploratory analysis tested the contribution of two functional variables—Question Type (polar or not polar) and Answer Type (answered or not answered)—as well as the length of the question (in syllables) and the language spoken (Dutch or English). As we will review, the impact of inclusion of some of these additional dialogue factors suggests that any direct influence of speech rate variation on TTT is subsidiary, at best, to considerations of the functional nature of the question-answer turn.

4.1. Speech Rate and Final Rhyme Duration as Predictors of TTT

We found that speech rate had a positive relationship with TTT both in the full dataset (which includes both overlapping turns and gaps) and in the Gaps dataset; thus, in both cases, faster speech in questions was associated with longer TTT before answers. This is contrary to the predictions by Wilson and Wilson (2005) and Garrod and Pickering (2015), and conflicts with the experimental observations by Corps et al. (2020) and, to some extent, the corpus study by Beňuš (2009), although he found only sporadic support for the hypothesised relationship. Our findings are more in line with the descriptive observation by Roberts et al. (2015) on the effect of the automatically extracted phoneme rate in their corpus of spontaneous conversation. This negative influence of rate on TTT is a clear indication that listener entrainment to the flow of speech in the attended turn is not the primary determiner of TTT as the listener becomes the speaker, but also raises the question of why faster questions should be associated with relatively delayed answers, rather than there being no observed relationship. One possibility is that faster questions, particularly those with relatively unpredictable content, may impose a higher cognitive load on the listener in understanding the questions and then formulating their response (e.g., Müller et al. 2019; Levinson and Torreira 2015). We would not commit to a strong interpretation of the finding, however, noting that any direct influence of speech rate on TTT disappears when our functional factors are also considered as predictors (see below).

We also found an effect of final rhyme duration in the Gaps dataset: thus, longer final rhymes were associated with longer TTT. This effect is contrary to the overall rate effect, where faster syllable rate, and hence shorter syllables, were associated with longer TTT. It also contradicts Corps et al. (2020): when response times were measured from final word offset, they found that a relatively short final *word* (not final rhyme, though all final words were monosyllabic in their first experiment) was associated with *longer* response times. It is also inconsistent with Torreira and Bögels (2022), who concluded that TTT should be shorter if turn-taking cues—such as final lengthening—can be recognized earlier.

It is possible that the phase of the hypothesized entrained motor-perception system is reset very locally, specifically in response to the final syllable. In that case, longer final rhymes being associated with longer TTT would be in line with entrainment theories (see Corps et al. 2019). However, it is claimed to be specifically the *onsets* of syllables

(or other entraining auditory stimuli), not their duration, that are responsible for phase resetting (Bosker 2017). If so, the lengthening of the final syllable rhyme occurs too late to reset the response phase. This local resetting of speech rate expectations is also somewhat contrary to longer-domain ("distal") speech rate effects on phonetic interpretation (e.g., Baese-Berk et al. 2014; Morrill et al. 2015). Finally, the local phase-resetting interpretation is not supported by all aspects of the marginal Gaps dataset interaction between speech rate and final rhyme duration.

4.2. Speech Rate × Final Rhyme Interaction and TTT

We found a marginal interaction between speech rate and final rhyme duration in the Gaps dataset. As shown in Figure 2, this was essentially a crossover interaction; thus, higher speech rate is associated with shorter TTT when the final rhyme is relatively long, and with longer TTT when the final rhyme is relatively short. Whilst we treat this marginal interaction with caution, it is nevertheless suggestive of a number of possible interpretations, potentially complementary rather than mutually contradictory.

Firstly, one aspect of the interaction offers at least superficial support for the entrainment theories (Wilson and Wilson 2005; Garrod and Pickering 2015), particularly when visualized in the median-split graphs for the speech timing predictors (Figure 2). Specifically, the relationship between speech rate and TTT depicted in Figure 2b (left panel) is negative, meaning that a faster speech rate is associated with a shorter TTT, as predicted. Note that the median split is for graphical illustration, and does not reflect a qualitative split in the data. However, the marginal interaction itself arises between two continuous predictors. Moreover, it not clear what might cause speech rate entrainment to prevail as the determiner of TTT specifically in this subset of the data. As this negative rate/TTT relationship is contrary to the overall positive influence of speech rate on TTT, it is more parsimonious to assume that something other than intermittent entrainment lies behind the observations.

In our view, a more plausible view of the interaction is that a greater contrast between final rhyme duration and the duration of preceding syllables promotes shorter TTT; specifically, faster foregoing speech and longer final rhymes conspire to make final lengthening more salient. It is well established that foregoing speech rate can impact the interpretation of local durational variation, such that segments are perceived as longer when the preceding rate is faster (e.g., Morrill et al. 2015; Reinisch et al. 2011). Such effects have been associated with predictive timing mechanisms, which may be underpinned by entrainment (at some neural/cognitive level) to the foregoing flow of syllables (e.g., Baese-Berk et al. 2014; White et al. 2022).

This interpretation accords with the view that turn timing arises from multiple speech processing mechanisms, including predictive and reactive elements, rather than as a direct result of perceptual-motor entrainment (cf. Torreira and Bögels 2022). It is also worth noting that where final rhymes are relatively short (Figure 2b, right panel), TTT increases with rate: thus, with less salient or absent utterance-final cues, listeners need longer to respond to faster questions, possibly due to processing demands (see above).

A qualification to all of these considerations is, of course, that the marginal crossover interaction, if robustly replicated, can been seen as being underpinned by a change in the relationship between rate and TTT, depending on final rhyme duration, and/or a change in the relationship between final rhyme duration and TTT, depending on rate. This modulation of influences on TTT, which is more evident when we consider other question and answer characteristics, accords with our primary conclusion, however: speech rate is not a consistent determiner of TTT in question-answer pairs.

4.3. Question Length and TTT

In the full dataset, longer questions (in terms of syllable number) were associated with shorter TTT and there was a similar trend ($p = 0.058$) in the Overlaps dataset, in line with question length effects found by Roberts et al. (2015) and Corps et al. (2019). There was

no question length effect in the Gaps dataset. Where found, such effects may indicate that longer questions afford earlier availability (with respect to turn ends) of transition-relevant information, such as prosodic and syntactic turn-ending cues (Torreira and Bögels, 2022). Thus, the listeners' ability both to predict turn endings and to plan their responses prior to the transition may be facilitated by lengthened questions.

4.4. Question Type, Answer Type and TTT

In the full dataset and the Gaps dataset, polar (yes-no) questions had shorter TTT than non-polar questions. This is unsurprising, given that TTT increases with answer complexity (Roberts et al. 2015) and polar questions typically solicit less complex answers than open questions. Additionally, in the full dataset, answers that were judged to respond directly to the question are associated with shorter TTT than non-answers, in line with findings of Stivers et al. (2009). More generally, questions that can be answered are likely to promote a quicker response (e.g., Kendrick and Torreira, 2015).

Additionally, in our data, these functional variables, where included in our regression models, explain more TTT variance than do speech rate or final rhyme duration (see AIC drops in Tables 4 and 5). Indeed, significant speech rate effects disappear when question type and answer type are added, along with question length, to the regression model for the full dataset; likewise, neither speech rate nor final rhyme duration are retained as significant TTT predictors in the Gaps dataset when question type is added (final rhyme duration $p = 0.09$).

This does not appear to be because question type or answer type covertly encode speech rate or final rhyme duration variation; polar questions and non-polar questions did not differ in speech rate, $t(204.75) = 0.830$, $p = 0.407$, nor in final rhyme duration, $t(214.24) = 0.887$, $p = 0.376$. Likewise, answered questions did not differ from non-answered questions in speech rate, $t(90.148) = -1.358$, $p = 0.178$, nor in final rhyme duration, $t(112.76) = -0.495$, $p = 0.621$. Thus, we conclude that our speech timing factors may have some utility in predicting TTT in some circumstances, but not only is the effect of speech rate contrary to the entrainment-based predictions (e.g., Wilson and Wilson, 2005), but any influence of rate or final lengthening is relatively trivial compared to dialogue factors such as question type.

4.5. Speech Rate × Question Type Interaction

In the exploratory analysis for the full dataset, we found an interaction between speech rate and question type. Thus, for polar questions, faster speech was associated with longer TTT, while for non-polar questions faster speech was associated with shorter TTT. This is not congruent with the experimental findings of Corps et al. (2020), who used only polar questions in their stimulus set and found that faster questions prompted shorter response latencies. We refrain from further interpretation of this unexpected result here, other than to speculate that the typical information structure of the two types of questions may interact differently with the effects of speech rate on cognitive load discussed above. We also note that, since polar questions constitute the larger proportion of the full dataset observations (299 polar vs. 114 non-polar), the overall positive relation between speech rate and TTT may be driven by the direction of the effect in polar questions in particular.

4.6. Language and TTT

Language (Dutch vs. English) did not predict TTT in any of our analyses. This is unsurprising given that the two languages are closely related and prosodically similar, although differences in TTT between Dutch and English, as well as between other languages, have previously been observed (Stivers et al. 2009). A variation in TTT could, however, arise from many factors, including dialogue task, conversational context and social relationships of interlocutors (also variation in measurement and annotation methods), and these factors were relatively consistent in our data. A study of similarly balanced corpora across a wider sample of languages would be required to establish whether our primary finding—a lack

of support for the entrainment-based negative relation between speech rate and TTT—is observed beyond question-answer pairs in these two Germanic languages.

4.7. Observational vs. Experimental Studies

The differences between Corps et al.'s (2020) findings and our results may be due to differences in task demands, since our study concerns spontaneous conversation, whereas Corps et al.'s data were derived from a speeded response task. Indeed, Corps et al. observe that their response latencies were longer than TTTs in spontaneous conversations, a comparison which Meyer et al. (2018) argued is relevant to judging whether experimental approaches adequately reflect conversational behaviour.

Additionally, there are not just many factors (Roberts et al. 2015) and different processing mechanisms (Torreira and Bögels 2022) affecting TTT, but there may also be different strategic approaches to turn taking, engaged by speakers at different times to different degrees (Heldner and Edlund 2010; Campbell 2008). Under experimental conditions, participants may orient their behaviour not towards responding at the "right time", but rather towards responding as quickly as possible. Such a strategy may lead, for example, to participants ignoring turn-ending prosody, as in Torreira and Bögels's (2022) no-prosody condition. This task-oriented strategy may promote overall speech rate entrainment more than spontaneous conversation.

4.8. Timing of Gaps vs. Overlaps

A practical reason for distinguishing gaps and overlaps in our analyses was that final rhyme duration cannot be equivalently defined for gaps and overlaps. Moreover, final lengthening is problematic to interpret as a cue in the case of overlaps. From a theoretical perspective, there are likely to be qualitative differences, at least in some cases, between gaps and overlaps: thus, it is not always simply the case that the listener's predictions are awry and thus lead to inadvertent interruption. Longer overlaps, in particular, may sometimes be purposive interruptions: in such cases, one would predict weaker or absent relationships between (negative) TTTs and the temporal properties of questions. Indeed, for the overlaps dataset, we found no effect of either speech rate or question length on TTT. Future quantitative studies could also aim to characterize qualitative differences more systematically in turn transitions, to attempt to distinguish those overlaps that are simply failed predictions about turn endings from those that are pragmatically judged interruptions (cf. Beňuš 2009).

5. Conclusions

The primary purpose of this study was to test for a relationship, in natural spontaneous conversations, between the speech rate of questions and the duration of the interval before the answers (turn-transition time: TTT). Contrary to predictions based on entrainment-based accounts of conversational dynamics (Wilson and Wilson 2005; Garrod and Pickering 2015), we found an overall positive association: thus, higher speech rate in questions was associated with longer TTT. There was no rate/TTT association in overlaps, i.e., where answers began before the end of questions.

For sequences with gaps between question offset and answer onset, there was also a marginal interaction with the duration of the final syllable rhyme: thus, when rhymes were relatively long compared to the foregoing syllables, TTTs were shorter. We tentatively suggest that this may reflect a variation in the salience of local timing cues to utterance boundaries. In the gaps dataset, we also found an interaction between speech rate and the type of question asked: the majority of the questions in our corpora were polar (yes-no questions), for which higher speech rate was associated with longer TTT. This may be due to the higher processing demands associated with answering relatively fast questions. In non-polar questions, we saw a negative relationship between speech rate and TTT as predicted by entertainment-based accounts.

Overall, these results point to multiple influences on turn-transition timing, with only a subsidiary role for speech rate. The factors—including discourse and interpersonal context, processing mechanisms and variation in behavioural strategies—that are most predictive of TTT are likely to be contingent on their salience and significance in specific interaction scenarios. We found scant evidence, however, that listeners' entrainment to speech rate directly governs their response time in individual question-answer pairs.

Author Contributions: Conceptualization, L.W. and D.H.; methodology, D.H and L.W.; software, D.H.; validation, D.H.; formal analysis, D.H., S.K. and L.W.; investigation, D.H. and L.W.; data curation, D.H.; writing—original draft preparation, D.H. and L.W.; writing—review and editing, D.H., L.W. and S.K.; visualization, D.H., L.W. and S.K.; supervision, L.W.; project administration, L.W.; funding acquisition, D.H. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Arts and Humanities Research Council grant number AH/R012415/1. The APC was funded by UKRI.

Institutional Review Board Statement: This study was approved by the Ethics Committee of Newcastle University (reference 9401/2020, 1st February 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the original studies.

Data Availability Statement: The data presented in this study are openly available at Newcastle University research repository, DOI: 10.25405/data.ncl.22657393.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

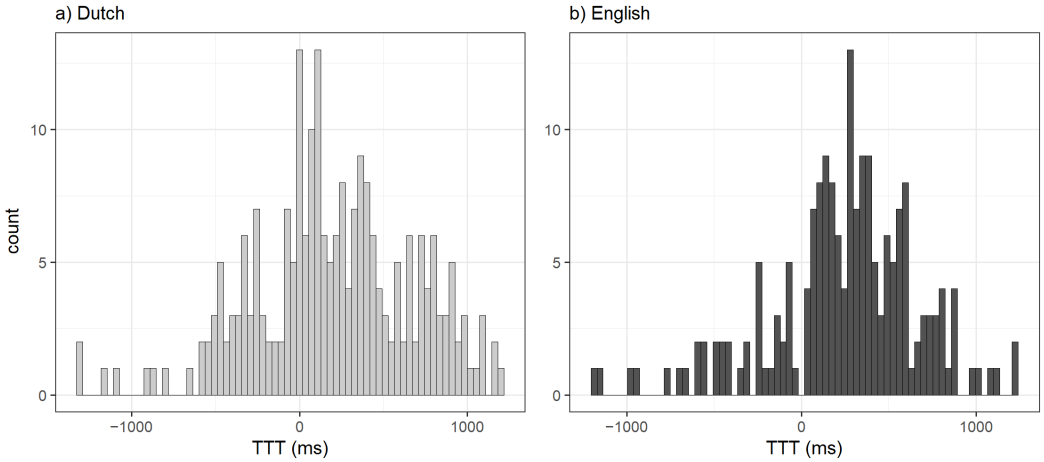

Figure A1. TTT distribution for the (**a**) Dutch and (**b**) English corpora (after applying exclusion criteria) in milliseconds based on 70 bins (bin width: Dutch: 61 ms, English: 56 ms).

References

Akaike, Hirotogu. 1973. Information Theory as an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 276–81.

Baese-Berk, Melissa M., Christopher C. Heffner, Laura C. Dilley, Mark A. Pitt, Tuuli H. Morrill, and J. Devin McAuley. 2014. Long-Term Temporal Tracking of Speech Rate Affects Spoken-Word Recognition. *Psychological Science* 25: 1546–53. [CrossRef] [PubMed]

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]

Beňuš, Štefan. 2009. Are We "in Sync": Turn-Taking in Collaborative Dialogues. Paper presented at Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, September 6–10.

Berkovits, Rochele. 1994. Durational Effects in Final Lengthening, Gapping, and Contrastive Stress. *Language and Speech* 37: 237–50. [CrossRef] [PubMed]

Bilmes, Jack. 1988. The Concept of Preference in Conversation Analysis. *Language in Society* 17: 161–81. [CrossRef]

Boersma, Paul, and David Weenink. 2021. Praat: Doing Phonetics by Computer. Available online: http://www.praat.org/ (accessed on 21 March 2022).
Bosker, Hans Rutger. 2017. Accounting for Rate-Dependent Category Boundary Shifts in Speech Perception. *Attention, Perception, & Psychophysics* 79: 333–43. [CrossRef]
Bögels, Sara, Kobin H. Kendrick, and Stephen C. Levinson. 2020. Conversational Expectations Get Revised as Response Latencies Unfold. *Language, Cognition and Neuroscience* 35: 766–79. [CrossRef]
Burle, Borís, and Michel Bonnet. 1999. What's an Internal Clock for?: From Temporal Information Processing to Temporal Processing of Information. *Behavioural Processes* 45: 59–72. [CrossRef]
Campbell, Nick. 2008. Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Berlin and Heidelberg: Springer, Volume 5042 LNAI, pp. 107–20. [CrossRef]
Coffey, Emily B. J., Isabelle Arseneau-Bruneau, Xiaochen Zhang, Sylvain Baillet, and Robert J. Zatorre. 2021. Oscillatory Entrainment of the Frequency-Following Response in Auditory Cortical and Subcortical Structures. *Journal of Neuroscience* 41: 4073–87. [CrossRef]
Corps, Ruth E., Birgit Knudsen, and Antje S. Meyer. 2022. Overrated Gaps: Inter-Speaker Gaps Provide Limited Information about the Timing of Turns in Conversation. *Cognition* 223: 105037. [CrossRef] [PubMed]
Corps, Ruth E., Chiara Gambi, and Martin J. Pickering. 2020. How Do Listeners Time Response Articulation When Answering Questions? The Role of Speech Rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46: 781. [CrossRef] [PubMed]
Corps, Ruth E., Martin J. Pickering, and Chiara Gambi. 2019. Predicting Turn-Ends in Discourse Context. *Language, Cognition and Neuroscience* 34: 615–27. [CrossRef]
Ding, Nai, and Jonathan Z. Simon. 2014. Cortical Entrainment to Continuous Speech: Functional Roles and Interpretations. *Frontiers in Human Neuroscience* 8: 311. [CrossRef] [PubMed]
Garrod, Simon, and Martin J. Pickering. 2015. The Use of Content and Timing to Predict Turn Transitions. *Frontiers in Psychology* 6: 1–12. [CrossRef] [PubMed]
Heldner, Mattias, and Jens Edlund. 2010. Pauses, Gaps and Overlaps in Conversations. *Journal of Phonetics* 38: 555–68. [CrossRef]
Kayser, Stephanie J., Robin A. A. Ince, Joachim Gross, and Christoph Kayser. 2015. Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. *The Journal of Neuroscience* 35: 14691–701. [CrossRef] [PubMed]
Kendrick, Kobin H., and Francisco Torreira. 2015. The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes* 52: 255–89. [CrossRef]
Lehtonen, Jaakko, and Kari Sajavaara. 1985. The Silent Finn. In *Perspectives on Silence*. Norwood: Ablex Publishing Corporation.
Levinson, Stephen C., and Francisco Torreira. 2015. Timing in Turn-Taking and Its Implications for Processing Models of Language. *Frontiers in Psychology* 6: 731. [CrossRef]
Lin, Hua, and Qing Wang. 2007. Mandarin Rhythm: An Acoustic Study. *Journal of Chinese Language and Computing* 17: 127–40.
Luo, Huan, and David Poeppel. 2007. Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron* 54: 1001–10. [CrossRef]
Manson, Joseph H., Gregory A. Bryant, Matthew M. Gervais, and Michelle A. Kline. 2013. Convergence of Speech Rate in Conversation Predicts Cooperation. *Evolution and Human Behavior* 34: 419–26. [CrossRef]
Meyer, Antje S., Phillip M. Alday, Caitlin Decuyper, and Birgit Knudsen. 2018. Working Together: Contributions of Corpus Analyses and Experimental Psycholinguistics to Understanding Conversation. *Frontiers in Psychology* 9: 525. [CrossRef]
Morrill, Tuuli, Melissa Baese-Berk, Christopher Heffner, and Laura Dilley. 2015. Interactions between Distal Speech Rate, Linguistic Knowledge, and Speech Environment. *Psychonomic Bulletin & Review* 22: 1451–57.
Müller, Jana A., Dorothea Wendt, Birger Kollmeier, Stefan Debener, and Thomas Brand. 2019. Effect of Speech Rate on Neural Tracking of Speech. *Frontiers in Psychology* 10: 449. [CrossRef] [PubMed]
Oller, D. Kimbrough. 1973. The Effect of Position in Utterance on Speech Segment Duration in English. *The Journal of the Acoustical Society of America* 54: 1235–47. [CrossRef] [PubMed]
Price, Patti J., Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. 1991. The Use of Prosody in Syntactic Disambiguation. *The Journal of the Acoustical Society of America* 90: 2956–70. [CrossRef]
Reinisch, Eva, Alexandra Jesse, and James M. McQueen. 2011. Speaking Rate Affects the Perception of Duration as a Suprasegmental Lexical-Stress Cue. *Language and Speech* 54: 147–65. [CrossRef]
Roberts, Felicia, and Alexander L. Francis. 2013. Identifying a Temporal Threshold of Tolerance for Silent Gaps after Requests. *Journal of the Acoustical Society of America* 133: EL471–77. [CrossRef]
Roberts, Felicia, Piera Margutti, and Shoji Takano. 2011. Judgments Concerning the Valence of Inter-Turn Silence across Speakers of American English, Italian, and Japanese. *Discourse Processes* 48: 331–54. [CrossRef]
Roberts, Sean, Francisco Torreira, and Stephen C. Levinson. 2015. The Effects of Processing and Sequence Organization on the Timing of Turn Taking: A Corpus Study. *Frontiers in Psychology* 6: 509. [CrossRef]
Stivers, Tanya, Nicholas J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and et al. 2009. Universals and Cultural Variation in Turn-Taking in Conversation. *Proceedings of the National Academy of Sciences* 106: 10587–92. [CrossRef] [PubMed]

Street, Richard L. 1984. Speech Convergence and Speech Evaluation in Fact-Finding Interviews. *Human Communication Research* 11: 139–69. [CrossRef]
Templeton, Emma M., Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont, and Thalia Wheatley. 2022. Fast Response Times Signal Social Connection in Conversation. *Proceedings of the National Academy of Sciences of the United States of America* 119: e2116915119. [CrossRef] [PubMed]
Torreira, Francisco, and Sara Bögels. 2022. Vocal Reaction Times to Speech Offsets: Implications for Processing Models of Conversational Turn-Taking. *Journal of Phonetics* 94: 101175. [CrossRef]
Van Bree, Sander, Ediz Sohoglu, Matthew H. Davis, and Benedikt Zoefel. 2021. Sustained Neural Rhythms Reveal Endogenous Oscillations Supporting Speech Perception. *PLoS Biology* 19: e3001142. [CrossRef] [PubMed]
Van Son, Rob, Wieneke Wesseling, Eric Sanders, and Henk van den Heuvel. 2008. The IFADV Corpus: A Free Dialog Video Corpus. Paper presented at Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May 28–30; pp. 501–8.
Wesseling, Wieneke, and Rob J. J. H. van Son. 2005. Early Preparation of Experimentally Elicited Minimal Responses. Paper presented at 6th SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal, September 2–3.
White, Laurence. 2014. Communicative Function and Prosodic Form in Speech Timing. *Speech Communication* 63: 38–54. [CrossRef]
White, Laurence, and Sven L. Mattys. 2007. Calibrating Rhythm: First Language and Second Language Studies. *Journal of Phonetics* 35: 501–22. [CrossRef]
White, Laurence, Sven Mattys, and Lukas Wiget. 2012. Segmentation Cues in Conversational Speech: Robust Semantics and Fragile Phonotactics. *Frontiers in Psychology* 3: 375. [CrossRef]
White, Laurence, Sven Mattys, Sarah Knight, Tess Saunders, and Laura Macbeath. 2022. Temporal Expectations and the Interpretation of Timing Cues to Word Boundaries. *Proceedings Speech Prosody* 2022: 322–26.
Wilson, Margaret, and Thomas P. Wilson. 2005. An Oscillator Model of the Timing of Turn-Taking. *Psychonomic Bulletin & Review* 12: 957–68. [CrossRef]
Wilson, Thomas P., and Don H. Zimmerman. 1986. The Structure of Silence between Turns in Two-Party Conversation. *Discourse Processes* 9: 375–90. [CrossRef]
Xu, Qin, and Datian Ye. 2015. Temporal Integration Reflected by Frequency Following Response in Auditory Brainstem. *Bio-Medical Materials and Engineering* 26: S767–78. [CrossRef] [PubMed]