

Conversational Turn Timing: The Effects of Prosody and Pragmatic Context in Production and Perception

Damar Hoogland

Submitted for the degree of Doctor of Philosophy

School of Education, Communication and Language Sciences

February 2025

Abstract

This thesis examines the production and perception of turn-transition timing in spontaneous conversation, focusing on the role of speech timing and pragmatic context.

Study 1 addressed the claim that turns are timed precisely, typically after pauses of ~200ms, through a systematic review of turn-timing distributions from 25 studies. All studies operationalised turn timing as the time between the end of one turn and the start of the next, but most did not provide acoustic-phonetic definitions of turn boundaries. Turn timing distributions had high kurtosis and were generally right-skewed, with central tendency measures clustered between 0-500ms. Means of task-oriented conversations (e.g., map tasks) were distributed more widely than free conversations.

Study 2 tested the perception-action entrainment hypothesis (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005) that higher speech rates should correlate with faster turn timing, by analysing local and global speech rate as predictors of turn timing in question-response sequences from 17 spontaneous conversations in Dutch and English using linear regression model selection. Contrary to the hypothesis, faster questions predicted slower turn timing, while shorter final syllables predicted faster turn timing. Polar questions were answered faster than open questions, and relevant responses were initiated earlier than non-relevant responses. Moreover, speech-timing variables did not significantly explain turn-timing variance when these pragmatic variables were included.

Not all turns are ‘on time’. Study 3 created a perceptual ‘timeliness’ variable from participant ratings of responses from Study 2. Speech rate predicted turn timing only for responses judged ‘timely’, which were earlier after slower questions. Timeliness judgements depended on both inter-turn pause duration and question-final syllable length, suggesting that the usual pause-based operationalisation of turn timing may not capture perceived turn timing.

These findings do not support entrainment hypotheses, emphasise the importance of pragmatic factors in conversational turn timing, and suggest that perceptually-validated operationalisations of turn timing are needed.

Acknowledgements

I would first like to thank my supervisors, Laurence White and Kai Alter, for their advice, time and patience the last few years, and my progress panel, Ghada Khattab and Adam Brandt, for their invaluable insights.

Several other people also helped me with specific aspects of this thesis. I would like to thank Sarah Knight her help with the statistical analysis in Chapter 4, Linda Errington for advice on how to conduct a systematic review, Evelyn Fischer for being the second reviewer, Lauren Ackerman for advice on so many aspects of statistics and design, Christoph Rühlemann for helping me with plotting, Mieke Bal, Anna Bruggeman, and Daan Hoogland for proofreading and pilot testing, and all study participants for their time.

This research was funded by the Northern Bridge Consortium, and I would like to thank them for making it possible. Particularly, thanks to Carole Palmer and Sarah Rylance for all their efforts!

I would also like to thank Senja Pollak and Matthew Purver for the opportunity to be part of their team at the JSI.

These are some more of the people who made all this possible:

- To all the members of the Phonetics and Phonology group: thank you for creating such a friendly and inspiring research environment.
- To Francis Nolan: thank you for your enthusiasm and encouragement over the years.
- To PhD friends and colleagues at the various offices, the linguistics writing group, the P&P statistics group and the PGR community, Andreas, Ayes, Bella, Boško, Bruce, Caitlin, Carroll-Ann, Fengting, Gustavo, Jaya, Madhava, Sergio, Yanyu: thank you for the fun and support.
- To Theo, Pete and Derrick: thank you for being great housemates in South Shields.
- To my meicogsci friends and colleagues, Aisha, Ana, Asura, Benni, Björn, Christina, Evelyn, Honza, Toby and Xenia: thank you for sharing your views on life and science.
- To my friends in various places during PhD times, especially Hania, Alifa, Ella, Robbie, and Louise, thank you for your support and friendship.

- Karolina, thank you for our calls lately, and for simply always being there since the start of this adventure.
- Timotej, thanks for bringing me snacks.
- Henk, the thesis is for the University, but the ten propositions will be for you.
- Mieke, no schools could ever beat our bicycle rides or French dinners.
- Daan, all roads led to linguistics, and you paved them. Thank you for our conversations.
- Trix, thank you for your love.

Table of Contents

Abstract.....	<i>i</i>
Acknowledgements.....	<i>iii</i>
Table of Contents	<i>v</i>
List of Figures	<i>x</i>
List of Tables	<i>xv</i>
Chapter 1. <i>Introduction</i>.....	<i>1</i>
1.1 General Context of Turn Timing Research	<i>1</i>
1.2 Thesis overview.....	<i>2</i>
1.2.1 Study 1: turn timing in conversation: a systematic review	<i>2</i>
1.2.2 Study 2: speech rate and turn timing in question-response sequences.....	<i>3</i>
1.2.3 Study 3: speech rate and turn-timing perception: a rating study of perceived response timeliness.....	<i>4</i>
Chapter 2. <i>Literature Review</i>	<i>6</i>
2.1 Introduction.....	<i>6</i>
2.2 What is Turn Timing?	<i>6</i>
2.2.1 Conversations, turn taking and turns.....	<i>7</i>
2.2.2 Turn timing	<i>9</i>
2.2.3 Why does turn timing need an explanation?	<i>12</i>
2.2.4 The scope of variation and precision in turn timing	<i>14</i>
2.2.5 Summary and research questions.....	<i>18</i>
2.3 Theories of Turn Timing Production	<i>19</i>
2.3.1 Coupled oscillator models	<i>20</i>
2.3.2 Indirect support for coupled oscillator models.....	<i>24</i>
2.3.3 Prediction-reaction models	<i>28</i>
2.3.4 Summary of predictions	<i>30</i>
2.3.5 The structure of TTT distributions	<i>30</i>
2.3.6 Speech rate and turn timing.....	<i>32</i>
2.3.7 Summary and research questions.....	<i>38</i>
2.4 Turn ‘Timeliness’ and Turn Timing Perception.....	<i>39</i>

2.4.1	Turn-timing fluency and naturalness	40
2.4.2	Interpretations of delay	43
2.4.3	Perception of turn timing and speech-rate normalization	44
2.4.4	Summary and research questions.....	47
2.5	Summary and Outlook on Empirical Chapters	48
Chapter 3.	<i>Turn Timing in Conversation: A Systematic Review</i>	50
3.1	Introduction.....	50
3.1.1	Motivation	50
3.1.2	Previous research	51
3.1.3	What is a systematic review?.....	52
3.1.4	Research questions and methodology	52
3.2	Methods	54
3.2.1	Information sources and search strategy.....	55
3.2.2	Inclusion and exclusion criteria	56
3.2.3	Screening process	60
3.2.4	Quality assessment	63
3.2.5	Further exclusion steps for the purpose of this chapter	64
3.2.6	Data extraction process	65
3.2.7	Synthesis methods	67
3.3	Results and Discussion	71
3.3.1	Selection results and report characteristics	71
3.3.2	Data sources and corpora	73
3.3.3	Summary of included languages	73
3.3.4	Operationalisations of turn timing and methodological variables.....	74
3.3.5	Conversation variables.....	78
3.3.6	Central tendency	80
3.3.7	Kurtosis and skewness.....	82
3.3.8	Subset 1	84
3.3.9	Subset 2	87
3.3.10	Entire selection.....	91
3.3.11	Languages.....	92
3.3.12	Repeated observations from the same corpora	93
3.4	Conclusion	96
Chapter 4.	<i>Speech Rate and Turn Timing in Spontaneous Question-Response Sequences.....</i>	100

4.1	Introduction	100
4.1.1	Motivation	100
4.1.2	Research questions and methodology.....	102
4.1.3	Question-response sequences	103
4.1.4	Question type, response type, question length and response length as modulating factors 104	
4.1.5	Linearity	105
4.1.6	Manual extraction of data	105
4.1.7	Languages and corpora	106
4.2	Methods	106
4.2.1	Corpora.....	106
4.2.2	TTT measurements in the full conversations	109
4.2.3	Sampling of question-response sequences	109
4.2.4	Speech- and language measurements	111
4.2.5	Statistical analyses	116
4.3	Results	120
4.3.1	Descriptive statistics of full corpus	120
4.3.2	Descriptive statistics of question-response sequences.....	122
4.3.3	TTT distributions	122
4.3.4	Differences between corpora.....	124
4.3.5	Correlations between predictor variables	125
4.3.6	Number of questions per speaker and conversation.....	126
4.3.7	Hypothesis testing – Full dataset.....	127
4.3.8	Exploratory analysis – Full dataset	127
4.3.9	Hypothesis testing – Gaps	130
4.3.10	Exploratory analysis - Gaps.....	132
4.3.11	Hypothesis testing and exploratory analysis – Overlaps.....	135
4.4	Discussion and Conclusion	135
4.4.1	Speech rate as predictor of TTT	136
4.4.2	Final syllable rhyme duration as predictor of TTT	137
4.4.3	Question Speech Rate × Final Rhyme Duration interaction and TTT	138
4.4.4	Question Length, Response Length and TTT	139
4.4.5	Question Type, Response Type and TTT	140
4.4.6	Question Speech Rate × Question Type interaction	141
4.4.7	Language and TTT	142
4.4.8	Observational vs. experimental studies	142
4.4.9	Timing of gaps vs. overlaps	143

4.4.10	Turn-timing distributions	143
4.4.11	Conclusion.....	145
Chapter 5. <i>Speech Rate and Turn-Timing Perception: A Rating Study of Perceived Response Timeliness</i>		147
5.1	Introduction.....	147
5.1.1	Motivation	147
5.1.2	Research questions and methodology	149
5.2	Methods	152
5.2.1	Participants.....	152
5.2.2	English participants	152
5.2.3	Dutch participants	153
5.2.4	Procedure	154
5.2.5	Stimuli	156
5.2.6	English stimuli	157
5.2.7	Dutch stimuli.....	158
5.2.8	Variables.....	160
5.2.9	Analysis	160
5.3	Results	164
5.3.1	Preliminary analysis of the English study	165
5.3.2	Preliminary analysis of the Dutch study	170
5.3.3	Mean Standardised Rating.....	174
5.3.4	Speech rate and turn timing in Early, Timely and Late sequences	176
5.3.5	Speech Rate Normalisation in English	180
5.3.6	Speech Rate Normalisation in Dutch.....	184
5.4	Discussion and Conclusion	186
5.4.1	Task salience and task demands	187
5.4.2	Speech rate and turn-timing in timely turn transitions	189
5.4.3	Speech rate normalisation	189
5.4.4	Turn-final segments as part of the turn transition	189
5.4.5	Conclusion.....	192
Chapter 6. <i>Discussion and Conclusion</i>		194
6.1	Introduction.....	194
6.2	Summary of Context.....	194
6.3	Summary of Findings.....	195

6.3.1	Study 1: Turn timing in conversation: a systematic review	195
6.3.2	Study 2: Speech rate and turn timing in spontaneous question-response sequences	196
6.3.3	Study 3: Speech rate and turn-timing perception: a rating study of perceived response timeliness.....	197
6.4	Implications.....	199
6.4.1	Models of turn timing	199
6.4.2	The role of turn-final segments in perception and production	201
6.4.3	Measuring turn timing.....	203
6.5	Conclusion	206
References		208
Appendix A1. PRISMA Checklists.....		228
Appendix A2. Examples of Records Excluded Based on Abstract		235
Appendix A3. Table of Included Reports.....		237
Appendix A4. Additional Grouping Plots.....		243
Appendix A5. Visualisations of Distributions		250
Appendix A6. Rating Study Participant Information Sheet		261
Appendix A7. Participant Instructions.....		263
Appendix A8. A Continuous Analysis of Means Standardised Ratings		265

List of Figures

Figure 2.1 Probability density function of a normal distribution with mean = 200ms and SD = 400ms.....	13
Figure 3.1 Adjusted PRISMA flow diagram (Page et al., 2021) indicating the number of studies included and excluded at each stage.	61
Figure 3.2 Frequency of reports published per year.	72
Figure 3.3 Frequency of reports that include conversational data from a particular corpus.	72
Figure 3.4 Frequency of reports that include conversational data from each language. NGT stands for ‘Nederlandse Gebarentaal’ (Sign Language of the Netherlands).....	74
Figure 3.5 Frequency of reports that include conversational data from a particular dialect of English, of the 20 reports that include data in English.	74
Figure 3.6 Heatmap depicting associations between the conversation- and methodological variables, measured by Cramer’s V. Statistically significant correlations at $\alpha=.05$ are marked with an asterisk (*) (The levels of the variable ‘TTT values included’ were grouped into the categories: ‘gaps’, ‘overlaps’ and ‘all’).	79
Figure 3.7 The proportion of the 41 samples in each level of each conversation variable.	79
Figure 3.8 Means, medians and modes of all included samples, arranged by sample size. Point size reflects the natural logarithm of the sample size.....	80
Figure 3.9 Measure of central tendency by sample size for the mean (a), median (b), mode (c) and SD (d) of each sample. The vertical line indicates the median of the means, medians, modes and sample sizes across samples.	82
Figure 3.10 The size and relative difference between the mean and median/mode as heuristic for assessing skewness. Point size reflects the logarithm (base 10) of the sample size.	84
Figure 3.11 The mean, median and mode of samples in subset 1, arranged by sample size. The sample estimates are colour-coded by Conversation Type. Point size reflects the natural logarithm of the sample size.	85
Figure 3.12 The mean, median and mode of distributions in subset 2, arranged by sample size. The sample estimates are colour-coded by Dialogue Act. Point size reflects the natural logarithm of the sample size.	88

Figure 3.13 The mean, median and mode of distributions in subset 2, arranged by sample size. The sample estimates are colour-coded by Conversation Type. Point size reflects the natural logarithm of the sample size.	89
Figure 3.14 The mean, median and mode of distributions in the entire selection of samples arranged by sample size. The sample estimates are colour-coded by TTT Values. Point size reflects the natural logarithm of the sample size.	90
Figure 3.15 The mean, median and mode of distributions in entire selection, arranged by sample size. The sample estimates are colour-coded by Conversation Type. Point size reflects the natural logarithm of the sample size.	91
Figure 3.16 Measures of central tendency per language for all samples in subset 2 that were from a language that was included at least twice in subset 2.	93
Figure 3.17 The mean, median and mode of distributions in the entire selection of samples, arranged by sample size. The sample estimates are colour-coded by Corpus. Point size reflects the natural logarithm of the sample size.	94
Figure 4.1 Histograms of the distributions of TTTs of all speech activity switches automatically extracted from a) the Dutch corpus, and b) the English corpus. Bin size = 100 observations.....	121
Figure 4.2 Histograms of TTT distributions in Dutch and English, before and after applying exclusion criteria. Bin size = 50 observations.....	123
Figure 4.3 Correlation heatmap displaying Pearson correlation between all variables, after applying exclusion criteria, for a) All, b) Gaps and c) Overlaps.....	125
Figure 4.4 Number of questions asked by each of the two speakers in each conversation, before applying exclusion criteria. Each row corresponds to a conversation. Numbers on the middle axis indicate the number of questions asked in that conversation in total, while the bars and numbers on each side indicate the number of questions asked by each speaker. The conversations are ordered by number of questions asked in total, from most (top) to least (bottom). Note that the speakers in each conversation were randomly assigned to be speaker 1 or speaker 2.....	126
Figure 4.5 Boxplots illustrating the difference in TTT between a) polar and non-polar questions and b) answer responses (answers) and non-answer responses (non-answers).....	128

Figure 4.6 Partial regression plots for each predictor in the final model for the Full dataset: a) Question Speech Rate, b) Question type, c) Response Type, d) Question Length, e) Question Type × Question Speech Rate.	130
Figure 4.7 Interaction between Speech Rate and Final Rhyme Duration by median split of Question Speech Rate (a) and median split of Final Rhyme Duration (b). The statistical interaction derives from continuous data and the median splits are depicted to illustrate	131
Figure 4.8 Boxplot illustrating the difference in TTT between polar and non-polar questions.	132
Figure 4.9 Partial regression plots for each predictor in the final model for the Gaps dataset: a) Question Speech Rate, b) Final Rhyme Duration, c) Question type, d) Question Speech Rate × Final Rhyme Duration.	134
Figure 5.1 Stimulus presentation screen in English (A) and Dutch (B).	154
Figure 5.2 Distribution of ratings by English-speaking participants.	165
Figure 5.3 Illustrations of the spread of ratings per stimulus for the English study for raw and scaled data.....	166
Figure 5.4 Illustration of the effect of presentation order in the English data as Trial Number against mean Rating by Trial Number	169
Figure 5.5 Distribution of ratings by English speaking participants in the first and last quarter of the study (i.e., first and last quantile of Trial Number).	170
Figure 5.6 Distribution of ratings by Dutch-speaking participants.	171
Figure 5.7 Illustrations of the spread of ratings per stimulus for the Dutch study.	172
Figure 5.8 Illustration of the effect of presentation order in the English data as Trial Number against mean Rating by Trial Number.	174
Figure 5.9 The distribution of Mean Standardised Ratings.....	175
Figure 5.10 Scatterplot of TTT against Mean Standardised Rating per stimulus.....	175
Figure 5.11 3D scatterplot of TTT against Mean Standardised Rating (MSR) against Question Speech Rate (QSR), from two different angles in a) and b).	177
Figure 5.12 Scatterplots of Question Speech Rate (QSR) and TTT and density distributions of TTT for a) Early, b) Timely, and c) Late subsets of the full dataset.	179
Figure 5.13 Scatterplots of Rating against TTT, Final Rhyme Duration (FRD) and Question Speech Rate (QSR), for English.	182

Figure 5.14 Scatterplots of Rating against TTT, Final Rhyme Duration (FRD) and Question Speech Rate (QSR), for Dutch.	185
Figure A 1 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Eye-contact. Point size reflects the natural logarithm of the sample size.	243
Figure A 2 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Naturalistic. Point size reflects the natural logarithm of the sample size.	244
Figure A 3 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Dyadic. Point size reflects the natural logarithm of the sample size.	245
Figure A 4 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Dialogue Acts. Point size reflects the natural logarithm of the sample size.	246
Figure A 5 The mean, median and mode of samples in subset 1, arranged by sample size. The sample estimates are colour-coded by the variable Eye-contact. Point size reflects the natural logarithm of the sample size.	247
Figure A 6 The mean, median and mode of samples in subset 1, arranged by sample size. The sample estimates are colour-coded by the variable Dyadic. Point size reflects the natural logarithm of the sample size.	247
Figure A 7 The mean, median and mode of samples in subset 2, arranged by sample size. The sample estimates are colour-coded by the variable Eye-contact. Point size reflects the natural logarithm of the sample size.	248
Figure A 8 The mean, median and mode of samples in subset 2, arranged by sample size. The sample estimates are colour-coded by the variable Naturalistic. Point size reflects the natural logarithm of the sample size.	248
Figure A 9 The mean, median and mode of samples in subset 2, arranged by sample size. The sample estimates are colour-coded by the variable Dyadic. Point size reflects the natural logarithm of the sample size.	249
Figure A 10 Histogram with normal distribution overlayed (Heldner & Edlund, 2010) Scottish English	250

Figure A 11 Histogram with normal distribution overlayed (Heldner & Edlund, 2010) Swedish.....	250
Figure A 12 Histogram (Stivers et al., 2009) Tzeltal	251
Figure A 13 Histogram (Stivers et al., 2009) Yélî Dnye	251
Figure A 14 Histogram (Stivers et al., 2009) Japanese	252
Figure A 15 Histogram (Stivers et al., 2009) Korean	252
Figure A 16 Density plot (De Vos, Torreira & Levinson, 2015) sign naive.....	253
Figure A 17 Histogram with normal distribution overlayed (Heldner & Edlund, 2010) Dutch	253
Figure A 18 Histogram (Levinson & Torreira, 2015) all	254
Figure A 19 Density plot (Pouw & Holler, 2022) ms	254
Figure A 20 Density plot (Roberts et al., 2015)	255
Figure A 21 Histogram (Stivers et al., 2009) English	256
Figure A 22 Histogram (Stivers et al., 2009) Lao.....	256
Figure A 23 Histogram (Stivers et al., 2009) ꠌAkhoe Hai om milliseconds, binsize = 100ms	257
Figure A 24 Histogram (Stivers et al., 2009) Dutch	257
Figure A 25 Histogram (Stivers et al., 2009) Italian.....	258
Figure A 26 Histogram (Stivers et al., 2009) Danish.....	258
Figure A 27 Histogram and density plot, seconds (Buanzur et al., 2018).....	259
Figure A 28 Histogram (Mori, 2009)	259
Figure A 29 Density plot, ms (Bögels, 2020)	260
Figure A 30 Density plot (De Vos, Torreira & Levinson, 2015) stroke-to-stroke	260
Figure A 31 Illustration of the interaction effects with categorically split data. Subsets consist of one-third of the Mean Standardised Rating scale. The Question Speech Rate × Final Rhyme Duration × Mean Standardised Rating is visualised as the product of Question Speech Rate and Final Rhyme Duration against TTT.	267

List of Tables

Table 3.1 Records used to develop the search strategy.	56
Table 3.2 (Part 1 of 2) Search strategy. For each database, this table lists the interface used to access it, the search queries used to search it and number of results for each search on 31-03-2023.....	57
Table 3.3 Reasons for excluding reports based on their operationalisation of turn timing, along with counts.	64
Table 3.4 Categories of extracted data.	68
Table 3.5 Variables coded from the narrative description of the conversations and methods in each report.....	69
Table 3.6 Type and number of reports and samples included in each synthesis subset, and the synthesis methods used for each subset.	70
Table 3.7 Methods by which the boundaries of turns were identified, with the number of reports that used each method.	75
Table 3.8 Number of observations of each level for each methodological variable.	78
Table 3.9 Summary of the distribution of five summary statistics across samples in the entire selection of samples.	81
Table 3.10 Skewness estimated by whether the median/mode is higher than the mean (left skew) or lower than the mean (right skew).	83
Table 3.11 Summary of the distribution of five summary statistics across samples in subset 1. The range, median, interquartile range and no. reported are listed for the sample means, medians, modes, standard deviations and sample sizes.....	85
Table 3.12 Summary of the distribution of five summary statistics across samples in subset 2. The range, median, interquartile range and no. reported are listed for the sample means, medians, modes, standard deviations and sample sizes.....	87
Table 4.1 Overview of interlocutor characteristics in the Dutch IFADV corpus.....	107
Table 4.2 Overview of conversation duration in the English Leverhulme corpus. Speaker demographic information is not available	108
Table 4.3 Descriptive statistics of turn switch distributions in the full corpora. The central point of the highest bin is listed, with bin size = 100 observations.	120
Table 4.4 Parameters of the best fitting mixture model in each language.	121

Table 4.5 Numbers of question-response sequences observed per language and per transition type (Gaps, Overlaps and both) before applying exclusion criteria (pre) and after (post).	122
Table 4.6 Measures of TTT central tendency per language before applying exclusion criteria (pre) and after (post). The central point of the highest bin is listed with bin size = 50.....	122
Table 4.7 Mean speech timing variables per language before applying exclusion criteria (pre) and after (post).	123
Table 4.8 Proportion of polar and non-polar questions per language before applying exclusion criteria (pre) and after (post).	124
Table 4.9 Proportion of answer and non-answer responses per language before applying exclusion criteria (pre) and after (post).	125
Table 4.10 Model estimates for the final model, full dataset. Categorical variables were sum coded such that polar = 1, non-polar = -1, answered = 1, not-answered = -1, so their coefficients represent main effects.	129
Table 4.11 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.	129
Table 4.12 Model estimates for the final model, Gaps dataset. Categorical variables were sum coded such that polar = 1, non-polar = -1, answered = 1, not-answered = -1, so their coefficients represent main effects.	133
Table 4.13 LRTs comparing the final model for the Gaps dataset and models with each of the predictors removed.....	133
Table 4.14 Model estimates for the final model, Overlaps dataset.....	135
Table 5.1 Variables associated with each question-response sequence.....	161
Table 5.2 Categories of responses from the English study to the debriefing question “To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?”.	167
Table 5.3 Categories of responses from the Dutch study to the debriefing question “To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?”.	173
Table 5.4 Overview of number of observations per subset.	178
Table 5.5 Coefficients for Question Speech Rate for models predicting TTT from Question Speech Rate in the Early, Timely and Late subsets.	178

Table 5.6 Coefficients predictors in final models predicting TTT for Gaps in the Early, Timely and Late subsets.	180
Table 5.7 Model coefficients, standard errors and LRT statistics for each predictor (English).	184
Table 5.8 Model coefficient and standard errors and LRT statistics for each predictor (Dutch).	186
Table A 1 PRISMA checklist (Page et al., 2021).	232
Table A 2 PRISMA-S checklist (Rethlefsen et al., 2021).	234
Table A 3 A list of all included reports (part 1 of 6).	237
Table A 4 A list of all included reports (part 2 of 6).	238
Table A 5 A list of all included reports (part 3 of 6).	239
Table A 6 A list of all included reports (part 4 of 6).	240
Table A 7 A list of all included reports (part 5 of 6).	241
Table A 8 A list of all included reports (part 6 of 6).	242
Table A 9 Model coefficients, standard errors and LRT statistics for the final model for the Full dataset with Mean Standardised Rating as continuous predictor.	266
Table A 10 Model coefficients, standard errors and LRT statistics for the final model for the Gaps dataset with Mean Standardised Rating as continuous predictor.	266

Chapter 1. Introduction

During conversations, interlocutors must time their turns appropriately, to avoid interrupting each other or disrupting the conversation with long silences. At the same time, variations in conversational turn timing can be pragmatically meaningful – interlocutors feel they click more when the gaps between their turns are shorter, for example (Templeton et al., 2022), while disruptions to turn timing due to delayed telephone transmissions lead them to think of each other as disengaged (Schoenenberg et al., 2014).

To explain how speakers manage to time their turns with such apparent precision, coupled oscillator models of turn timing (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005) propose that oscillatory neural activity in the auditory cortex of listeners tracks the speech rate of their interlocutor. Through a perception-action link, interlocutors are predicted to time their turns in time with the previous turn's speech rate: thus, faster turns should largely be associated with shorter *turn-transition time* (TTT; the time between the end of the first turn and the start of the second). Because variation in turn timing can be meaningful, we might furthermore expect that this relationship is modulated by the degree to which turns sound early, on time, or delayed.

In this thesis, I present three studies on conversational turn timing: The first study is designed to contribute to the discussion on the extent of variation and precision in turn timing, and the second and third studies test psycholinguistic hypotheses that turn timing behaviour and perception are affected by the speech rate of the preceding turn.

Section 1.1 discusses the relevance of studying turn timing in general and Section 1.2 then provides a broad outline of the rest of this thesis.

1.1 General Context of Turn Timing Research

One of the main contexts in which humans use language is spoken interaction, or conversation. Conversation exists in all cultures (Dingemanse & Floyd, 2014), has been called the 'ecological niche' of language (Levinson & Torreira, 2015), and has been hypothesised to be the context in which we developed language in the first place (cf. Tomasello, 2010; Hauser et al., 2002). As utterances in conversations are necessarily placed in time, and delayed turn timing conveys pragmatic meaning (F. Roberts & Francis,

2013), a theory of language as a system is therefore not complete without an account of turn timing.

Besides this theoretical relevance of turn timing as an aspect of language, understanding how speakers time their turns also has practical applications. For example, the way speakers time their turns can be affected by various conditions such as Alzheimer's Disease (Nasreen et al., 2021) or psychotic disorders (Howes & Lavelle, 2023). It has been used as a biomarker to detect some of those conditions in clinical assessments (Alpert et al., 1997) and research attempting to use it in automatic, machine-learning-driven diagnostic tools has become frequent in recent years (De Boer et al., 2020; Oakey-Frost et al., 2022). In the context of such research, understanding the variability of turn timing across populations and conditions is especially relevant to help ensure that diagnostic tools are not harmful.

Such practical applications could be refined by detailed psycholinguistic models of how turn timing is achieved. From a psycholinguistic perspective, turn timing is often presented as a surprisingly precise behaviour. A widely cited estimate is that speakers often initiate their turns within approximately 200ms from the end of the previous turn (Stivers et al., 2009), which is in the same order of magnitude as the blink of an eye (Caffier et al., 2003). Psycholinguists have argued that, because this fast timing requires speakers to process and produce speech at the same time, it forms a puzzle for theories of language processing, which have traditionally focused on production and perception in isolation (Bögels & Levinson, 2017). In this thesis, I am mostly concerned with turn timing as a psycholinguistic phenomenon: a special case of speech timing.

1.2 Thesis overview

1.2.1 Study 1: turn timing in conversation: a systematic review

In Section 2.2 of the literature review, I discuss the notion that turn timing is a 'precise' behaviour and argue that it is worth defining the statistical properties of turn timing distributions, and extent of variation between samples.

Previous studies that characterised turn timing statistically have sometimes used secondary data, but did not use systematic methods to collect it (Heldner & Edlund, 2010; Levinson & Torreira, 2015). To my knowledge, no systematic review of turn timing in adult human conversation exists (cf. Nguyen et al., 2021). To contribute to the discussion

on precision and variation in turn timing, Chapter 3 therefore presents a systematic review that addresses the following research questions:

Chapter 3. Research questions

Are assumptions about turn timing often cited in the literature supported by empirical evidence?

What is the existing scope of variation in turn timing?

Based on a dataset of 41 turn timing samples from 25 included studies, Chapter 3 describes the spread of measures of central tendency of turn timing distributions and the shape of those distributions. It also discusses how variation between the estimates does not appear to be related to the language spoken or to any methodological aspects of the studies, except perhaps to the type of conversation.

1.2.2 Study 2: speech rate and turn timing in question-response sequences

Psycholinguistic models of turn timing, particularly the coupled oscillator models (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005), are discussed in Section 2.3, along with previous experiments and corpus studies that have come to contradictory conclusions. These contradictory conclusions may be due to the differences between experimental conditions and spontaneous conversations, or to an interaction between speech rate and other factors that influence turn timing, or to other speech timing features of the turn. Chapter 4 presents a study of the timing of responses to questions in two corpora of spontaneous conversations in Dutch and English, to address the following research questions:

Chapter 4. Research questions

Is speech timing (particularly speech rate) predictive of turn-transition time?

If so, what is the nature of their relationship, and is it modulated by other factors such as the answer type, the question type, turn length, and the language?

Chapter 4 found that, contrary to the predictions by Wilson and Wilson (2005) and Garrod and Pickering (2015) and also contrary to previous experimental results, faster turns are associated with longer positive turn-transition time. This effect does not appear to be very important compared to other dialogue-related influences on turn timing.

1.2.3 Study 3: speech rate and turn-timing perception: a rating study of perceived response timeliness

Turn timing can convey meaning. To get an intuitive idea of this phenomenon, imagine talking to a friend at a bus stop and asking if they would like to meet for a coffee tomorrow. If they take an unusually long time to respond, you are likely to interpret that delay as indicative of their negative attitude – you have interpreted their turn timing (Bögels et al., 2020; Kendrick & Torreira, 2015). Speakers may also sometimes accidentally mistime their turns. In Section 2.4, I discuss how the concepts of pragmatic meaning and turn timing disfluency are potentially interwoven. Based on participants' ratings of their perceptions of the corpus data from Chapter 4, Chapter 5 then assesses whether the relationship between speech rate and turn timing is different depending on whether turns are perceived as 'on time', 'early' or 'late'.

Conversely, speech rate can affect our perception of other durational aspects of speech, but the hypothesis that it also affects our perception of turn timing does not appear to have been tested directly from a quantitative perspective. For this reason, Chapter 5 also presents a secondary analysis to test whether participants' perceptions of turn timeliness are affected by speech rate.

Chapter 5. Research questions

What is the relationship between speech rate, turn timing, and perceived turn 'timeliness'?

Is the relationship between speech rate and turn-transition timing in question-response sequences modulated by participant ratings of turn timeliness? In other words, is there a consistent relationship between speech rate and turn timing specifically when turns are perceived as 'on time'?

Do speech timing features (particularly speech rate) of the question (and/or responses) affect participants' ratings of response timing?

The study in Chapter 5 found that faster turns are with later turn timing specifically when the next turn sounds 'on time', contrary to the predictions by Wilson and Wilson (2005) and Garrod and Pickering (2015). Chapter 5 also found no evidence that speech rate affects listeners' perception of turn timing, but the duration of the final syllable before the

turn switch does appear to influence perceptions – when the final syllable is longer, turns are also perceived as being initiated later.

The rest of this thesis is organised as follows. Chapter 2 first reviews research on turn timing distributions, production models of turn timing, and turn timing perception and interpretation. Chapters 3, 4 and 5 then present the studies outlined in the sections above. Chapter 6 discusses the findings of the three studies, focusing on the implications for turn timing models and methodology.

Chapter 2. Literature Review

2.1 Introduction

The aim of this chapter is to situate the three following empirical chapters within the context of turn timing research. Section 2.2 first provides definitions of turn timing and related concepts and then reviews research on the empirical characteristics of turn timing distributions, particularly addressing the idea that turn timing is a ‘precise’ behaviour. Chapter 3 presents a systematic review that aims to contribute to this discussion on the scope of variation and precision in turn timing.

Section 2.3 outlines theoretical models of turn timing production that aim to account for the observed characteristics of turn timing, specifically focussing on *coupled oscillator models* and their predictions on the role of speech rate variation. It outlines how previous research has found inconsistent evidence, potentially due to a methodological divide between experimental and observational studies. The corpus study of speech rate and turn timing in Chapter 4 aims to bridge that divide.

Section 2.4 discusses how the concepts of perceived turn timeliness, turn timing fluency, and the pragmatic meaning of timing are potentially interwoven, and reviews the limited evidence on how turn timing is perceived. This section informs the methodology of the perception study in Chapter 5, which operationalised the timeliness of turns through participants’ ratings.

2.2 What is Turn Timing?

Conversational turn timing refers to the time it takes for interlocutors to respond to each other during conversations. It is usually measured as the time between the end of one speaker’s turn and the start of the next turn by a different speaker: the *turn-transition time* (TTT). To support the following discussion about turn timing theories and the methodologies used in this thesis, the next three sections provide more detailed working definitions of conversation, turn taking, and turn timing, and discuss the implications of adopting variations of those definitions. Sections 2.2.1 and 2.2.2 discuss why turn timing needs an explanation, and what characteristics of turn timing distributions have been observed.

2.2.1 Conversations, turn taking and turns

The phenomenon of interest in this thesis is *conversational* turn timing, as opposed to the timing of turns during other kinds of practices. Following T.P. Wilson et al. (1984) and Sacks et al. (1978) for now, I will define conversations as synchronous spoken exchanges between speakers during which the content, length, number, progression and crucially, timing, of each speaker's contribution is not fully determined by explicit rules or by one participant who exclusively manages these aspects. This definition excludes, for example, written communication, or exchanges of recorded voice messages. It also excludes any exchanges in which the progression and content of turns is completely determined by customary practice or one person, such as responsive readings in Anglican church services.

T.P. Wilson et al. (1984) and Sacks et al. (1978) also excluded interviews, on the grounds that interviews involve some kind of institutional expectation of how the exchange is organised. However, taking the view that authority and custom are always present to some extent (see Dingemanse & Floyd, 2014, but also T. P. Wilson et al., 1984 themselves), I use a more inclusive application of this definition. In the rest of this thesis, interviews, conversational games (e.g., Anderson et al., 1991), and conversations recorded in a lab, are all considered to fall under the umbrella of conversation, as long as they satisfy the criteria above. Experiments that aimed to study one of these aspects by constraining another therefore are not conversations, but evidence from such studies will still be discussed below.

Conversations are often conceived of as consisting of a pattern of turn taking, similarly to games like chess (Duncan, 1972; Sacks et al., 1978; Yngve, 1970; cf. Levinson & Torreira, 2015; T. P. Wilson et al., 1984). In the case of conversations, interlocutors take turns speaking.

One model that formalises this turn taking system was proposed by Sacks et al. (1978). They aimed to account for what they call 'grossly apparent facts' (Sacks et al., 1978, p14) about conversation, including their observation that turn taking appears organised to minimize gaps and overlaps. Their model of turn taking has been influential in the psycholinguistic literature on turn timing, so we will discuss its main components.

Sacks et al. (1978) conceive of conversational turn taking as an economic system, in which interlocutors are negotiating for a resource: the right to speak and be listened to at a particular moment, i.e., to have the current 'turn' or hold the 'conversational floor' (cf. Hayashi, 1991). Interlocutors use their shared understanding of the building blocks and rules of conversational turn taking to collaboratively allocate the conversational floor. The building blocks are turn constructional units (TCU), which are utterances that could function as a turn by themselves and can range in length. When a speaker reaches the end of the current TCU, and a speaker change could occur, given the form of the utterance (it is a transition relevant place; TRP), the following rules apply:

1. The current speaker selects the next speaker (for example by asking a direct question).
2. Another speaker selects themselves by initiating a turn.
3. The current speaker selects themselves by continuing to speak.

These rules are ordered, in the sense that rule 2 only applies if rule 1 was not employed during the turn, and rule 3 applies if rule 2 did not result in a speaker taking the floor. If, after the conversation reached a stage where rule 3 could apply, still no speaker took the floor, then the state of the conversation cycles back to rule 2.

Not all utterances are attempts to take the floor. Speakers can also utter backchannels during another speaker's turn: short utterances or non-linguistic cues initiated during another speaker's speech or inter-speaker pause, that do not take the floor but encourage the current speaker to continue (Yngve, 1970).

This approach by Sacks et al. (1978) is not the only possible approach to turn taking. They take what we might call a '*speaker-internal*' view of turn taking by presupposing a competent conversationalist's understanding of what features of a conversation can mean – that turn takes are distinguishable from backchannels, for example. A '*speaker-external*' approach might study conversation as the on-off pattern of speech activity, similarly to how we might study birdsong, without reference to symbolic meaning (e.g., Cangemi et al., 2023; Heldner & Edlund, 2010). For this purpose, the structure of conversations is captured adequately when all linguistic meaning is removed.

These views are not mutually exclusive (physically there is an on/off pattern of vocalisations within which we can distinguish functional units; T. P. Wilson et al., 1984),

but they affect operational definitions of concepts like ‘turn’. For example, an external view might not distinguish turns from backchannels that are initiated during a speaker pause, and treat all vocalisations as turn initiations, while an internal view would distinguish them. This distinction has consequences for how we study turn timing, as the next section discusses.

Further research has also introduced nuance to the conceptualisation of conversation as a turn-taking system, for example by treating non-linguistic gestures as potential TCUs, or as parallel to the turn taking system (for a review, see Dingemanse & Floyd, 2014). Additionally, alternative approaches to the structure of conversation have de-emphasised turn taking by not assuming that speakers are always oriented towards responding to each other (Campbell, 2007), or have adopted a non-discrete approach to analysing conversational structure by using techniques that detect patterns in the acoustic signal at the time-scale of conversations (Kello et al., 2017; Warner, 1979).

The next section discusses what turn timing is, considering the difference between speaker-internal and speaker-external approaches to turn taking.

2.2.2 Turn timing

Turn timing is the moment a speaker initiates their turn relative to events in the conversational context. It is usually measured as the time between the end of one turn and the start of the next turn by a different speaker. Throughout this thesis, I will refer to such measurements of turn timing as *turn-transition time* (TTT). Other terms to refer to this concept are, for example, *floor transfer offset* (De Ruiter et al., 2006) or *response time* (Templeton et al., 2022). Defined in this way, TTT can be positive and negative: a positive TTT means there is a silence or a gap between the turns, while a negative TTT means the turns overlap. From here onward, ‘early/late turn timing’ are used to refer to the general phenomenon, and ‘short/long positive/negative TTT’ when referring to measurements of turn timing¹.

How we identify the ends and starts of turns by which to measure TTT, has consequences for the interpretation of what turn timing reflects: an aspect of speech timing, or an aspect of conversational coordination. For example, if we take an external view and measure TTT

¹ Note the difference between the scales implied by these two concepts: turn timing goes from *early* to *late*, while TTT goes from *long negative* to *short negative* to *short positive* to *long positive*.

by identifying turn boundaries as the acoustic onsets and offsets of speech, TTT reflects the amount of silence or overlap between speech activity. This measure might be used to determine how well speakers avoid overlap or to measure a concept like conversational fluency (McCarthy, 2010). On the other hand, if we take an internal view, we might try to identify what is functionally the most likely turn end that a speaker was responding to. TTT then reflects a view on turn timing as a case of speech timing. The difference between these acoustic and functional operationalisations is illustrated in example 1): the blue square represents TTT measured functionally, and the yellow square represents TTT measured acoustically.

- 1) Q: Are you going out tonight? Did you see your parents yesterday by the way?
A: Yes, I'm going to see a movie.
[schematic example]

Quantitative corpus studies often use external, acoustic operationalisations of turn timing for the practical reason that they can be extracted automatically, but those interested in speech timing may be missing some detail that way, for example regarding sequences of conversation like in example 1. Moreover, if we distinguish turns from backchannels, or turn timing from speech initiation after a lapse in the conversation, not all speech initiation after speaker switches strictly speaking represents turn timing at all (i.e., not all speech initiation may be due to the same mechanism).

Corps et al. (2022) point out a similar problem with quantitative corpus studies that specifically aim to supplement experimental language processing studies on how conversationalists deal with the dual task demands of preparing their responses while also listening. These types of studies often measure TTT as speech activity offsets. Based on a corpus analysis, Corps et al. (2022) show that this approach disregards parallel talk, where speakers are not actually responding to their interlocutor's turn but continuing their own previous turn. Parallel talk may place different kinds of demands on a speaker's speech planning capacity, meaning that an acoustic operationalisation of turn timing without at least further functional analysis misrepresents processing time.

Even given these considerations, identifying the starts of turns is not necessarily straightforward: various approaches exist that differ in whether and what turn initial features, such as inbreaths, paralinguistic clicks, turn initial particles (such as "uhm"),

non-vocal gestures and gaze, or even the timing itself, are considered part of the turn (Tepperman, 2013; Kendrick & Torreira, 2015; Chaloff, 1987). An empirical example of why it matters how we identify the start of a turn is provided by Kendrick and Torreira (2015). They compared the timing of preferred and dispreferred responses – responses that are socially considered more favourable, such as accepting invitations, agreeing with statements or commitments to requests, are *preferred*, and those that are less favourable are *dispreferred* (Bilmes, 1988). They found that, contrary to common qualitative and experimental observations (Bögels et al., 2020; Couper-Kuhlen, 1993; Pomerantz, 1984), dispreferred responses were timed *earlier* when they measured the start of the second turn from turn-initial inbreaths and clicks. However, dispreferred responses were timed *later* if inbreaths and clicks were included in the gap, suggesting that this second operationalisation of turn timing better captured the intuition that dispreferred responses are timed later.

Relatedly, Knudsen et al. (2020) argue that turn initial particles serve to maintain appropriate turn timing while giving the responder more time to formulate a response. This argument entails that the choice of which point in time to measure turn timing from depends on the phenomenon of interest: if we are interested in speech initiation timing, the start of the particle is relevant, while if we are interested in response planning, the start of the base TCU (i.e., the TCU around which the turn is build; Kendrick & Torreira, 2015) is relevant.

The last sections argued that the way turn timing is operationalised has conceptual and empirical consequences. Rather than committing to one view throughout this thesis, the rest of this chapter and Chapter 3 will consider studies that used various operationalisations. Chapters 4 and 5 will mainly adopt operationalisations that represent turn timing as a form of speech timing, and Chapter 6 will return to this topic.

The next section discusses why turn timing needs an explanation, and reviews empirical evidence on the statistical characteristics of turn timing. It will show that the distribution of turn timing measurements tells us something about the phenomenon and the mechanism that produced it.

2.2.3 Why does turn timing need an explanation?

Turn timing is often characterised as *precise* (Holler et al., 2021; M. Wilson & Wilson, 2005), *smooth* (Buanzur et al., 2018; Corps et al., 2022; Riest et al., 2015), or *finely coordinated* (Girard-Groeber, 2015; Holler & Kendrick, 2015). These characterisations are not well-defined by themselves – they say nothing about the exact scope of variation, and can mean that turn timing is fast (Levinson & Torreira, 2015), but also that speakers time their turns very near to the end of the previous speaker’s turn (Heldner et al., 2011) – in other words, that they adhere closely to the no-gap-no-overlap principle.

Quantitatively, this idea is often supported with reference to a cross-linguistic study by Stivers et al. (2009). Stivers et al. found that across ten languages, the mean TTT in polar questions-response sequences (i.e., the timing of responses to questions that can be answered with ‘yes’ or ‘no’) was only 208ms. Moreover, the mean TTTs of polar question-response sequences in individual languages ranged from a mean TTT of 7ms for Japanese to a mean TTT of 469ms for Danish, all within 261ms from the cross-linguistic mean.

A common argument (e.g., Heldner & Edlund, 2010; Levinson & Torreira, 2015; M. Wilson & Wilson, 2005) why turn timing production needs a psycholinguistic explanation goes as follows: precise turn timing cannot simply reflect reaction time to the end of the previous turn, because most turns are initiated earlier than reaction time: minimal vocal reaction time alone, when speakers respond with a prepared syllable to a stimulus onset, is often estimated at around 200ms, on the basis of mean reaction times of 223ms and 213ms reported by Fry (1975). Responding to stimulus offsets tends to take longer than to stimulus onsets by at least 60 ms (Torreira & Bögels, 2022), and reaction time is longer in more complex contexts, such as for picture naming (Indefrey & Levelt, 2004). Consequently, if many turns are initiated earlier than e.g., 208ms after the end of the previous turn, they must involve something else than just reaction to the end of the turn.

The proportion of turns that are timed earlier than expected according to estimated verbal reaction times is illustrated in Figure 2.1. Figure 2.1 assumes that turn timing can be characterised as a normal distribution with a mean of 208ms and a SD of 513ms (the cross-linguistic estimates from Stivers et al., 2009; however, this assumption of normality will be discussed below). The blue dotted line indicates the minimal verbal

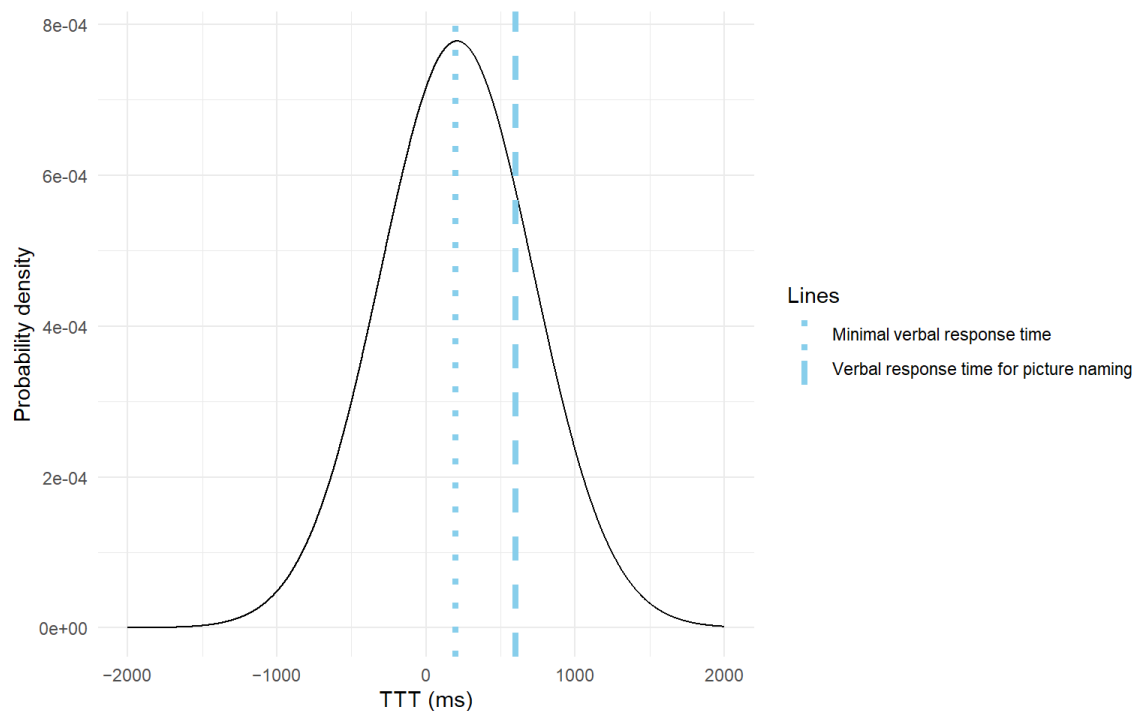


Figure 2.1 Probability density function of a normal distribution with mean = 200ms and SD = 400ms.

response time of 200ms estimated based on Fry's (1975) estimates, and the blue striped line indicates the average response time in a picture naming task of 600ms (Indefrey & Levelt, 2004).

As a consequence of the large proportion of turns that cannot be due to reaction time alone, speakers must already be performing some aspects of response preparation while listening to the previous turn. Levinson and Torreira (2015) argue that traditional models of speech comprehension and production cannot account for this necessity, because both comprehension and production require some of the same resources. Indeed, comprehension and production apparently compete for specific cognitive resources (Paucke et al., 2015) and neurological resources (Giglio et al., 2022; Lukic et al., 2021), and producing words while listening to other words reduces reaction time in experimental contexts (Ferreira & Pashler, 2002).

Besides pointing to small average TTTs, another argument in favour of the idea that turn timing is precise is framed around perception, not production (e.g., by Levinson & Torreira, 2015), since listeners are sensitive to, and interpret as meaningful, small variations in turn timing: listeners hear turn timing of 600ms or more as delayed (F. Roberts et al., 2011; F. Roberts & Francis, 2013), and detect TTTs of around 120ms as

gaps (Heldner, 2011). However, it is unclear whether that variation in production is under conscious control of the speaker and grammaticalized, or a ‘natural sign’ of the state of the speaker, i.e., it is not clear whether turn timing *production* or turn timing *perception*, or both, is the phenomenon we should describe as precise. Moreover, listeners can detect gaps in noise with much shorter durations, in the order of 4.2ms (Giannela Samelli & Schochat, 2008), suggesting that turn timing perception is potentially less precise than gap perception in general.

2.2.4 The scope of variation and precision in turn timing

The observations of precise turn timing production are often cited to motivate studies that aim to establish what determines turn timing, if not reaction to the end of the previous turn. However, the idea that turn timing is precise deserves some scrutiny for the following reasons.

First, Heldner (2011) argued that few turn transitions are perceived as ‘no-gap-no-overlaps’. They tested from what thresholds negative TTTs are perceived as overlaps, and positive TTTs are perceived as gaps, and found that overlaps shorter than ~120ms and gaps of less than 120ms are perceived as immediate turn transitions. Given this threshold, they also observed that no-gaps-no-overlaps are less frequent than gaps and overlaps larger than this threshold. In TTT distributions from three languages from Heldner and Edlund (2010) and (a gaussian model of) the data from the ten languages from Stivers et al. (2009), no-gaps-no-overlaps represent only 11.7% to 24.9% of cases. Heldner argued that this indicates the ‘precision’ of turn timing is frequently overstated.

Second, TTT distributions are often very wide: Heldner and Edlund (2010) report TTT distributions with ranges up to -3000ms to 5000ms, and Stivers et al. (2009) report that their ten samples have standard deviations between 371ms and 662ms. A ‘precisely timed’ behaviour should have a relatively narrow distribution, although of course there is no benchmark for what is ‘relatively narrow’. Along these lines, Heldner (2011) notes: “Clearly, this view [that conversation is well-organised to minimise gaps and overlaps] is only tenable if you ignore the fact that response times are distributed.” (Heldner, 2011, p512). Of course, this view does not imply that long overlaps and gaps are in fact favoured over no-gaps-no-overlaps, since the idea that conversation is fully precise or not precise at all are not the only two possible positions; it only implies that the exact extent to which conversation avoids gaps and overlaps successfully should be defined.

Third, there are other estimates of central tendency of turn timing distributions that are widely different from the mean 208ms TTT from Stivers et al. (2009) but not frequently cited to characterise turn timing. For example, Alpert et al. (1997) studied timing in interviews with neurotypical participants and participants with diagnoses of schizophrenia or depression. They reported a mean response time of 1500ms (presumably TTT) for the neurotypical group – more than three times the mean in the Danish polar question-response sample and 1292ms longer than the cross-linguistic mean observed by Stivers et al. (2009).

Possible explanations for the difference between these means could lie in the methodology of the studies or the characteristics of the data. On the one hand, the difference between the findings reported by Alpert et al. (1997) and Stivers et al. (2009) might be due to methodological differences such as in the manner of measuring turn timing. Unfortunately, Alpert et al. do not describe how they measured turn timing, so this possibility cannot be confirmed.

On the other hand, a second possible explanation is that the difference between Alpert et al. (1997) and Stivers et al. (2009) is related to another statistical characteristic of turn timing that has sometimes been observed, namely that turn timing distributions do not tend to be strictly normal distributions: Sørensen et al. (2021) note, based on a review of only three studies, that turn timing distributions tend to be right-skewed. As a measure of central tendency, the mean is more sensitive to outliers than the median and mode, and normal-like distributions with a right skew (i.e., a positive skew), have a higher mean than mode or median because of relatively infrequent but relatively large positive values. As a result, using the mean as a summary statistic to describe a skewed distribution is likely to lead to an overestimation of the most frequent observations in the distribution, and some researchers have suggested using medians or modes to describe turn timing distributions instead (Kendrick & Torreira, 2015; see also Heldner & Edlund, 2010, make this point for analyses that treat gap and overlap distributions separately, since those distributions are skewed and have moreover a fixed lower or upper boundary).

In the case of Alpert et al. (1997) and Stivers et al. (2009), the fact that Alpert et al. found a much larger mean than Stivers et al. may therefore be due to the presence of relatively many large TTT values in the former's data. On the one hand, we could regard these infrequent positive values as outliers, but on the other hand, we could consider them a

characteristic of the data that is captured meaningfully by the large mean. In either case, this large discrepancy should ultimately be accounted for by theories of language processing based on estimates of turn timing. More generally, if we consistently observe that turn timing distributions have a particular characteristic, such as skewness, that characteristic should be explained by a theory of the behaviour that produced that distribution.

Another, empirically motivated reason why the shape of TTT distributions matters comes from the previously mentioned study by Kendrick and Torreira (2015), who compared the timing of preferred and dispreferred responses. They quantitatively assessed whether dispreferred responses are timed later than preferred responses, as qualitative studies report (Couper-Kuhlen, 1993; Pomerantz, 1984). They observed that the most frequent dispreferred and preferred responses follow a TTT of approximately the same length (when TTT is measured from the end of the first turn to the start of the second turn, not considering inbreaths and clicks), but that that TTTs of dispreferred responses are distributed more widely and have a higher density after around 700ms. This pattern could result in a larger positive mean TTT for dispreferred responses, which would inaccurately confirm that dispreferred responses are timed later.

A theoretically motivated reason to care about the shape of turn timing distributions is that some theories of turn timing production explicitly predict that they should have a shape that reflects a cyclical underlying mechanism (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005). Sections 2.3.1 and 2.3.6 return to this prediction.

A fourth challenge to the idea that turn timing is precise is that it is widely observed that many factors affect turn timing: ranging from turn type or dialogue act (Knudsen et al., 2020; S. G. Roberts et al., 2015; Stivers et al., 2009), length (Heldner et al., 2011), the pragmatic meaning associated with turn timing (Couper-Kuhlen, 1993; Kendrick & Torreira, 2015; F. Roberts et al., 2011) and the presence of turn-taking cues (Beattie, 1978), to the speakers' psychiatric conditions (Howes & Lavelle, 2023) and background noise in the conversation's setting (Sørensen et al., 2021). More generally, Roberts et al. (2015) built a random forest model predicting turn timing from 32 variables, and concluded that no single type of variable was uniquely more predictive than the others. This view is in tension with, even if it is not contradictory to, the view that turn timing is precise. However, it is not clear for all those factors whether they have consistent effects,

which factors could explain the large differences in means reported, and consequently, to what *degree* turn timing is precise.

A final point of interest about variation and precision in turn timing is that turn timing is also regularly represented as *cross-linguistically* precise (Dingemanse & Floyd, 2014) based on the aforementioned study by Stivers et al. (2009), but that the degree and consistency of cross-linguistic variability is not known. Specifically, Stivers et al. (2009) observe, despite anthropological reports of differences in tolerance for silence in conversation across cultures, that five sequence-organisational variables they studied almost all affected TTT in the same way in all the ten languages. From this observation they conclude that turn timing is cross-linguistically organised according to the same principles.

However, on the one hand, Stivers et al. (2009) do observe numerical differences in the means, medians and modes of the languages, which the authors attribute to small cross-linguistic differences in listeners sensitivity to delay (see Section 2.4). Language is a significant grouping variable in their regression model, and using their reported means and standard deviations to calculate a one-way ANOVA from sufficient data also shows a significant effect of language, $F(9) = 20.04$, $p < .001$, although note that this test assumes that the data is normally distributed. A significant difference in TTTs has also been reported for American English and Finnish conversations, as well as for the behaviour of American English speakers and Finnish speakers in conversation with each other (Sneck, 1987).

On the other hand, the relative difference in TTT or in sensitivity to delay (see Section 2.4), or the magnitude of that difference between languages has to my knowledge not been replicated. In a study of 24 languages, Dingemanse and Liesenfeld (2022) report a cross-linguistic mean of 59ms. Although they do not state their operationalisation of turn timing or language specific measures of central tendency, in a figure of density plots, the languages do not appear to follow the same grouping as in Stivers et al. (2009). Similarly, Weilhammer and Rabold (2003) found significant differences in overlap lengths, but not pause lengths, between German, English and Japanese, but the weighted means of pause and overlap lengths of their language samples also do not follow the same grouping as in Stivers et al. (2009).

Besides this indication of cross-linguistic differences, it is also not clear if the range of turn timing variation in the data from Stivers et al. (2009) quantitatively represents the cross-linguistic variation in turn timing in general. Stivers et al. included data from specifically polar question-response sequences. They argue that turn timing in polar question-response sequences is representative of all turn timing based on a non-significant difference between turn timing in *all* question-response sequences and *all other* sequences in a corpus of Dutch telephone conversations. However, first, non-significant difference does not establish that two categories are functionally the same. Second, polar question-response sequences may contrast with open question-response sequences, rather than all question-response sequences with non-question-response sequences, since polar questions constrain the possible responses (Bögels et al., 2020; Knudsen et al., 2020). Third, turn switches in telephone conversations and polar question-response sequences in face-to-face multi-party conversations also share the feature that the next speaker often is selected by the current speaker. As Stivers et al. (2009) themselves show, when the next speaker is selected by speaker gaze, they initiate their turn earlier. Holler et al. (2021) also observe that responses in question-response sequences in multi-party conversations are timed earlier than other types of turn switches, apparently for this reason, and Beattie (1978) reports that responses to questions were later than responses to non-questions. Patterns of turn timing in polar-question response sequences may therefore not capture all aspects of variation².

In short, there appears to be scope for variation in turn timing, including cross-linguistic variation, but it is not clear to what extent that variation is representative and consistent across studies.

2.2.5 Summary and research questions

To summarise, the claim that turn timing is precise should be interpreted in the context of the discussion in which it is made (i.e., models of speech perception and production predict longer latencies), since empirical evidence can be interpreted to indicate the opposite: that turn timing is not precise but highly variable. Moreover, it is worth

² Additionally, as discussed in Section 2.2.2, not all speaker switches may be the result of the same mechanism, so the anthropological observations of differences in tolerance for silence can possibly not be tested quantitatively with reference to polar question-response sequences only.

investigating not just *if* turn timing is precise, but exactly what the scope of variation is, and what consistent patterns in turn timing distributions exist.

To contribute to this discussion, Chapter 3 presents an analysis of secondary data gathered as part of a systematic review, to address the following research questions:

1. Are assumptions about turn timing often cited in the literature supported by empirical evidence, and what is the existing scope of variation in turn timing?
 - 1.1. What estimates of turn timing are reported in studies of spontaneous conversation between adults?
 - 1.2. What operationalisations of turn timing are used and what are their differences?
 - 1.3. What factors are found to have a consistent effect on turn timing?
 - 1.4. To what extent is there variation between turn timing estimates that cannot be accounted for by factors considered within the studies?

2.3 Theories of Turn Timing Production

The previous section reviewed why turn timing is considered to be precise and in need of an explanation, although the exact degree of variability is to be determined. This section discusses psycholinguistic models that aim to account for how turn timing can be so precise, i.e., they aim to model the exact moment speakers initiate their turns.

Psycholinguistic models of turn timing can be divided into two types. These two types of models differ in how they propose that the mechanism that produces turn timing is structured. On the one hand, some researchers have theorised that turn timing is the result of entrainment of neural oscillators between speakers, and importantly, a perception-action link (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005). They entail that turn timing is produced by a cyclical mechanism. I will refer to these models as *coupled oscillator models*.

On the other hand, some models conceive of turn timing as the result of some combination of processing time (e.g., Levinson & Torreira, 2015), arising from prediction and/or reaction processes similar to Hick's law, or Donder's observations on the effect

of decision choice on decision time (Donders, 1969; Hick, 1952; cited in Levinson & Torreira, 2015). For convenience, I will refer to the latter as *prediction-reaction models*.

It is important to note, however, that the components of these two types of models are not all contradictory or mutually exclusive, as will be highlighted in the next sections as well. Additionally, researchers have also classified turn timing models along different dimensions. For example, the coupled oscillator models have been characterized, including by the researchers who first proposed them (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005), as prediction models, in opposition to reaction models (T. P. Wilson & Zimmerman, 1986), because they emphasise that speakers make predictions about their interlocutor's speech timing during conversations to identify upcoming turn ends. However, empirically, the key point that distinguishes these two approaches is that they make different predictions regarding the nature of the relationship between speech rate and turn timing: coupled oscillator models predict that turn timing is earlier after faster turns than after slower turns and vice-versa, while prediction-reaction models do not predict such a relationship.

Section 2.3.1 first outlines the coupled oscillator models and Section 2.3.2 then reviews indirect support for those models. To provide a contrast with the coupled oscillator models, Section 2.3.3 outlines prediction-reaction models of turn timing. Finally, Sections 2.3.5 and 2.3.6 focus on evidence for and against predictions regarding speech rate and the shape of turn timing distributions.

2.3.1 Coupled oscillator models

A model of the coupled oscillator type was proposed Wilson and Wilson (2005), who aimed to identify the neurological implementation of the behavioural rules for turn taking formulated by Sacks et al. (1978). Particularly, they were interested in explaining the cyclical application of the rules in cognitive-neurological terms. In doing so, they also aimed to account for some empirical observations, including the generalisation that 1) turns are timed smoothly, with minimal gaps and overlaps, 2) that simultaneous starts are rare (Jefferson, 1989), and 3) their own observations that TTTs appear to be in the range of the average length of a syllable and that there is a periodic component in turn timing distributions (T. P. Wilson & Zimmerman, 1986; this latter point is discussed in section 2.3.5).

Wilson and Wilson (2005) hypothesise that during conversation, endogenous neural oscillators in the auditory cortex of the listener entrain to the speech rate of the turn they are listening to, which in turn controls the timing of their own speech initiation through a perception-action link.

Endogenous neural oscillators are neurons or agglomerates of neurons, cyclically firing in phase with each other (Stiefel & Ermentrout, 2016). In this case, these agglomerates of neurons are located in the auditory cortex, so their activity should be detectable in electroencephalography (EEG) or magnetoencephalography (MEG) recordings, which only detect activity just under the scalp, as high power in the frequency at which the neural activity oscillates (Donoghue et al., 2022; Giraud & Poeppel, 2012). Wilson and Wilson (2005) point out that at the time of their writing, the activity of neural oscillators had been implicated in temporal perception and processing (e.g., Burle & Bonnet, 1999, and others cited in M. Wilson & Wilson, 2005), and they suggest that neural oscillators in the auditory cortex entrain to the rate of syllables. This unidirectional coupling of auditory oscillations to heard speech should be robust to some local irregularities in speech timing to track speech rate over the course of a turn.

To account for turn timing, Wilson and Wilson then propose that the entrained oscillator in the auditory cortex is itself coupled with oscillating activity in the motor cortex, forming a perception-action link. They suggest that the entrained activity in the motor cortex can be represented as a speech readiness function: the cyclical activity corresponds to increases and decreases in likelihood of initiating speech.

During conversations, because the speaker's speech entrains the auditory oscillator of the listener, interlocutors become linked in a system of entrained oscillators. Wilson and Wilson additionally hypothesise that the speaker's and listener's speech readiness function is anti-phase locked, meaning that the speaker's readiness to speak is maximal when the listener's is minimal, and vice versa.

Specifically, this last hypothesis represents the cyclical nature of the rules formulated by Sacks et al. (1978; see Section 2.2.1): rule 2 can only apply when the current listener's speech initiation readiness is at a maximum, and the previous speaker's readiness at a minimum, while rule 3 can only apply when the previous speaker's speech initiation readiness is at its maximum, and the listener's at a minimum. If no speech is initiated at

one of the states of the system, it cycles back to the other state. When none of the interlocutors in the conversation initiate speech, after some time the linked cycle breaks down because of the absence of entertaining speech. The resulting silence is a lapse.

A similar proposal was made by Couper-Kuhlen (1993; cf. Auer et al., 1999; Ogden & Hawkins, 2015), although not from a neurological or psycholinguistic perspective. Through qualitative analysis of metrical patterns around turn switches, Couper-Kuhlen argued that conversation is organised to maintain a continuous rhythmic structure across turn switches and that divergence from this structure has a communicative function. They also suggest that the cyclical application of the rules from Sacks et al. (1978) are applied in time with this rhythmic structure.

Importantly, Wilson and Wilson's model does not account for how the listener knows *that* the end of a turn is coming up (Wilson and Wilson suggest they use turn-ending cues to do that, as observed by e.g., Duncan, 1972; Duncan & Niederehe, 1974). It only accounts for how the listener estimates at what exact point in time a turn ending will occur, once they have identified that the turn end is coming up at a particular cycle of the system of entrained oscillators. The model also aims to account for the exact moment the listener can launch articulation of their own turn, once they have identified at which cycle of the system of entrained oscillators it is appropriate to start speaking.

Also seeking to identify an implementation of the rules of turn taking proposed by Sacks et al. (1978), Garrod and Pickering (2015) formulated a theory of speech comprehension and production during conversation that included a similar model of turn timing. They propose that interlocutors make predictions regarding the content and the timing of their interlocutor's turn using separate modules of their speech processing system, whose output the listener then integrates to determine the exact time the turn they are listening to is likely to end.

On the one hand, to predict the content of the current unfolding turn, listeners primarily use a prediction-by-simulation strategy. Based on knowledge and context, speakers simulate what their interlocutor is likely to say next. On the other hand, to predict the exact timing of that content, Garrod and Pickering (2015) hypothesise that listeners rely on entrained neural oscillations in their auditory cortex, which they integrate with their content prediction at a later stage.

Garrod and Pickering do not explicitly aim to explain the cyclical application of the rules formulated by Sacks et al. (1978) with their timing prediction module, only the generalisation that turns are often initiated after minimal gaps and overlaps. However, they do hypothesise, like Wilson and Wilson (2005), that entrained oscillators in the auditory cortex of the listener are coupled with oscillators in the motor cortex that control speech initiation, constituting a functional perception-action link like the one proposed by Wilson and Wilson (2005).

These two models differ somewhat in their emphasis, but when it comes to turn timing production, they both hinge on the role of a proposed periodic perception-action link. As a result, they both predict that speech rate is correlated to turn timing: TTT is expected to be generally shorter after a faster turn than after a slow turn.

The models also differ with regard to some more specific predictions made by the authors, some of which are also implicit in the other. First, Wilson and Wilson (2005) predict that the speech rate of speech across both inter-turn pauses and between-turn pauses should correlate, and that the speech rate of the second turn correlates with TTT.

Second, both models entail that the underlying mechanism that produces turn timing is cyclical. Wilson and Wilson (2005) predict that this structure should be detectable in observed TTT distributions. Because the frequency of their hypothesised periodic speech readiness function is around the time scale of a syllable and anti-phase locked, they furthermore explicitly predict that TTT should cluster at around half a syllable before and half a syllable after 0ms TTT. At the same time, TTTs of exactly 0ms are expected to be relatively rare, since the current speaker's readiness to initiate speech is at a maximum at the end of their turn, while the listener's is at a minimum. Garrod and Pickering (2015) do not explicitly predict such a clustering pattern or bimodality in TTT distributions, although the entrained perception-action link in their model also does require an underlying cyclical clustering of TTT of some kind, even if it is not expected to be visible in raw TTT distributions.

Third, Garrod and Pickering (2015) predict that the regularity of a turn should affect the timing of the subsequent turn, although they do not explicitly state how. Presumably, more irregular speech should be associated with less precise timing, and so should a less predictable final syllable duration. The former prediction is not strongly implied by Wilson

and Wilson (2005), who expect that the auditory oscillator is robust to irregularities. The latter would be an artifact of measuring TTT as the end of the first turn to the start of the second, since an unexpectedly long or short final syllable compared to the previous speech rate would occupy an unexpected amount of time during the cycle of the coupled oscillator system that spans the TTT.

Aside from similarities and differences in their behavioural hypotheses regarding turn timing, the models point to different anatomical structures as candidates for what substantiates their perception-action. Garrod and Pickering (2015) suggest a pathway through the midbrain (citing Giraud et al., 2007; Kotz & Schwartz, 2010). In a review of the role of the motor cortex in speech perception, Scott et al. (2009) proposes a similar role for a posterior pathway that projects from the auditory to the premotor cortex. Wilson and Wilson (2005), on the other hand, do not discuss the anatomy of the perception-action link.

Support for these proposals should come in the form of evidence of the perception-action link at a neurological level and converging evidence on the existence and function of entrained oscillators in the auditory cortex. The next section reviews such converging evidence from studies on neural entrainment to speech, as well as entrainment of speech rates between speakers, and convergence and covariance of turn timing between speakers.

2.3.2 Indirect support for coupled oscillator models

This section reviews empirical observations that provide indirect support for the entrainment models of turn timing by Wilson and Wilson (2005) and Garrod and Pickering (2015). Evidence for and against their predictions on speech rate and the shape of TTT distributions will be discussed in Sections 2.3.5 and 2.3.6.

First, entrainment of activity in the sensory areas of the cortex to rhythmic stimuli has been widely observed, including when listening to speech (Luo & Poeppel, 2007); neural activity, as measured by EEG, synchronises to semi-regularities in the amplitude-envelope of speech at multiple time scales (Bröhl & Kayser, 2021), including in frequencies from around 1Hz to 4Hz (customarily referred to as the delta band) and 4Hz to 8Hz (the theta band; Donoghue et al., 2022), which correspond approximately to the range of syllable rates (Pellegrino et al., 2011).

Some researchers have questioned whether the signal detected to be synchronous with regular stimuli is always direct evidence either of entrainment (L. Meyer et al., 2019), or of endogenous oscillatory neural activity (Haegens & Zion Golumbic, 2018); it could also be responsive activity that appears oscillatory because it responds to regular activity (e.g., steady-state event-related potentials; Luck, 2014; for an argument against, see Doelling et al., 2019). In favour of the idea that the relevant activity is evidence of entrained oscillators are observations that it takes some cycles to build up (Luo & Poeppel, 2007), and that the entrained signal is sustained for some cycles after the series of regular stimuli has ceased (Coffey et al., 2021; Van Bree et al., 2021; Xu & Ye, 2015).

This activity appears to play a role in comprehension, since, for example, comprehension is reduced when the degree of entrainment is disrupted by transcranial magnetic stimulation (Riecke et al., 2018). However, it is not clear exactly how the entrained neural activity is integrated in the speech processing pipeline – i.e., whether it reflects low-level acoustic encoding or higher-level perceptual processing (Kayser et al., 2015; L. Meyer et al., 2019).

On the one hand, entrainment is stronger when the speech is played forward than when it is played backward (Gross et al., 2013) and it is weaker when the auditory input is manipulated to reduce intelligibility, for example when syllable-rate fluctuations are removed (Doelling et al., 2014). The strength of entrainment is also modulated by whether a listener is attending to the speech or not, when they are exposed to a cocktail-party style stimulus (Golumbic et al., 2013; cf. Calderone et al., 2014; Obleser & Kayser, 2019). These observations suggest that the observed activity does not only reflect low-level acoustic processing.

On the other hand, listeners show a comparable degree of entrainment to speech in a language they understand than to speech in a language they do not understand (Song, 2018). Additionally, Kayser et al. (2015) inserted pauses in speech stimuli to make them irregular, and found that delta-band entrainment was reduced while theta-band entrainment was sustained, which they take to suggest that entrainment in the delta-band reflects a top-down process, while entrainment in the theta band reflects low-level acoustic encoding.

Nevertheless, entrained neural activity at the timescale of syllables has been implicated in speech processing as a framework for temporal prediction and ‘packaging’ of auditory information (Ding & Simon, 2014; Giraud & Poeppel, 2012; Luo & Poeppel, 2007; Teng et al., 2018), by determining the sampling rate of the acoustic input (Giraud & Poeppel, 2012), or by aligning phases of high sensitivity with salient moments in the acoustic input (Peelle & Davis, 2012). For example, Peelle and Davis (2012) suggest that the oscillatory activity reflects increases and decreases in perceptual sensitivity: at one point in the phase, the relevant part of the auditory cortex is more likely to pass on sensory information than at other times. This proposal is supported by evidence that the timing of stimuli relative to the oscillatory activity’s phase affects the likelihood of the stimulus being heard (e.g., Kubetschek & Kayser, 2021). Entrainment to the syllable rate of speech then has the function of aligning moments of high perceptual sensitivity to perceptually salient points in the syllable. As a result of this mechanism, by timing their turns in time with the previous turn’s rate, speakers facilitate speech perception for their interlocutors.

Wilson and Wilson (2005) point out that a consequence of their model is that speech rates across turn switches should correlate. Such entrainment of speech rate between speakers could similarly serve to facilitate perception. Evidence for that effect has indeed been observed: In experimental conditions, speakers’ speech rates are affected by a speech prime (Jungers & Hupp, 2009), and they adjust their speech rate to that of a confederate in scripted dialogues (Schultz et al., 2016) and conversations (Freud et al., 2018). In spontaneous conversation, interlocutors’ speech rates tend to be correlated (Xia, Levitan and Hirschberg, 2014) and their speech rates converge more closely after speaking for longer periods of time (Finlayson et al., 2012; Schultz et al., 2016), and are specifically correlated across turn switches (Finlayson et al., 2012). This convergence of speech rate serves, or at least relates to, a social function: people are more likely to cooperate after a conversation if their speech rate converged more closely (Manson et al., 2013).

In support of Garrod and Pickering’s (2015) expectation that the regularity of speech affects turn timing, speech also appears to become more regular around turn switches (Mooney & Sullivan, 2015; Ogden & Hawkins, 2015). In general, people behave more regularly when timing is at stake in a task (Lelonkiewicz & Gambi, 2020; Vesper et al., 2011).

Independently from speech rate, the timing of turns of interlocutors also converges across conversations (Edlund et al., 2009; Franich, 2023; Ten Bosch et al., 2005), tends to become shorter and more similar over the course of a conversation (Edlund et al., 2009; Shimojima et al., 2002; cf. Wehrle, 2021) and varies together during conversations (Edlund et al., 2009; Shimojima et al., 2002).

Apparently to the contrary, Pouw and Holler (2022) observe a *negative* lag-1 autocorrelation of turn timing during dyadic conversations: over the course of the conversation, TTT values are negatively correlated to the previous TTT value of a different speaker. Nevertheless, Pouw and Holler (2022) suggests that this observation reflects a compensatory effect, where speakers adjust their turn timing based on the behaviour of their interlocutor rather than on their own past behaviour. They interpret it as evidence for the general view, exemplified also by Wilson and Wilson's (2005) proposal, that timing during conversation is not the result of speech timing mechanisms that operate within individuals, but that it is co-produced by the interlocutors in such a way that they can be thought of as a single system (cf. Auer et al., 1999; Couper-Kuhlen, 1993). Along similar lines, hyperscanning studies have shown that interlocutors' brain activity correlates during verbal interactions (Ahn et al., 2018; Nam et al., 2020). Both these observations are compatible with, but not direct evidence for, the proposal that turn timing is due to a perception-action link that creates a system of coupled oscillators spanning across interlocutors, as suggested by Wilson and Wilson (2005), Garrod and Pickering (2015).

In general, these observations of correlation, convergence and synchrony can be interpreted to suggest that interlocutor's turn timing behaviour is linked at a mechanistic level. However, convergence of this type is not direct evidence for a perception-action link; Ten Bosch et al. (2005) suggest an alternative interpretation, namely that convergence of turn timing could be caused indirectly by the degree to which interlocutors are having a more or less engaging conversation, which may also apply to convergence of speech rate.

The presence of oscillatory activity that becomes entrained to the amplitude envelope of speech, or any effect that might have on auditory perceptual sampling, is also not sufficient to support Wilson and Wilson (2005) and Garrod and Pickering (2015). Without a perception-action link, entrained neural activity can still be implicated in turn timing,

for example if it reflects the mechanism by which listeners predict when a turn is likely to end, but then we would not expect any relationship between speech rate and TTT.

The next section on prediction-reaction models first briefly outlines an alternative view of the mechanism that produces turn timing that does not assume a perception-action link, and Sections 2.3.5 and 2.3.6 then review more direct evidence regarding speech rate and TTT distributions.

2.3.3 *Prediction-reaction models*

Unlike the coupled oscillator models by Wilson and Wilson (2005) and Garrod and Pickering (2015), some psycholinguistic theories of turn taking do not conceive of the mechanism that produces turn timing as due to a cyclical perception-action link. Instead, they predict observed TTTs are the result of a (linear) combination of processing time, potentially including comprehension and turn prediction processes, speech articulation preparation, inhibition of launching prepared speech motor plans, and reaction time to turn ending cues.

An example of a psycholinguistic theory of turn taking that entails a prediction-reaction model of turn timing was proposed by Levinson and Torreira (2015) and later developed by Barthel (2020). Levinson and Torreira outline a box-and-arrow model that aims to account for the fact that vocal reaction time in experimental contexts is longer than turn timing during conversation. In essence, their model is structured as follows:

Listeners monitor the incoming turn to recognise the dialogue act being produced. At this point, they already start planning their own turn, potentially down to the level of an articulation motor plan, but this prepared articulation plan is held in a buffer while the listener continues to monitor the incoming turn for syntactic completion cues, and if signalled, prosodic turn-ending cues. Once a turn-ending cue is recognised, articulation is released from the buffer. Barthel (2020) proposed several refinements to this model, including the suggestion that the listener actually releases their prepared articulation plan from the buffer once they reach a threshold of certainty that a transition relevant place is reached, which is influenced by their ongoing projections and monitoring of the pragmatic, syntactic and phonological progression of the turn. Note that these proposals differ from one in which the initiation of articulation is launched in anticipation of a predicted target.

A prominent discussion among researchers investigating the time-course of turn preparation and production concerns the timing of response planning – the ‘early- versus late planning’ debate (for a review, see Bögels & Levinson, 2017). Specifically, the question is whether speakers prepare their turn as early as possible or as late as possible. The former means that speakers must use potentially costly general attentional resources to both comprehend their interlocutor’s turn and prepare their own at the same time and use the same resources to do speech comprehension and production (Giglio et al., 2022; Lukic et al., 2021; Paucke et al., 2015), but have their response ready early enough to account for observed early turn timing. Levinson and Torreira (2015) fall on the ‘early planning’ side of this debate and propose that listeners are engaged in dual tasking: their cognitive and neurological resources switch rapidly between comprehension and production processes. A ‘late planning’ position is often attributed to Heldner and Edlund (2010), because they argue that many turns are in principle initiated late enough to be due to reaction time.

Support for prediction-reaction models comes from, among other sources, studies that demonstrate that an increase in processing load delays turn timing. For example, syntactically more complex turns are timed later in spontaneous conversations (S. G. Roberts et al., 2015). Presenting a distracting task during turn taking also increases turn timing (Boiteau et al., 2014). Furthermore, speakers respond earlier when they could start planning their utterance earlier, because the relevant linguistic information needed to start planning the response appeared earlier in the question, under experimental conditions (Magyari et al., 2017) as well as during interviews (Bögels et al., 2020). Under experimental conditions, participants also respond earlier when a lexical or pitch turn-ending cue is present than when it is not present (Barthel et al., 2017) and they respond earlier when the turn ending cue appears further away from the end of the turn (Torreira & Bögels, 2022). However, these effects could also be observed if a perception-action link governed turn timing. Indeed, Garrod and Pickering (2015) explicitly predict such effects of predictability.

The key difference between coupled oscillator models and prediction-reaction models is thus that prediction-reaction models do not predict a direct relationship between speech rate and subsequent turn timing. However, Roberts, Torreira and Levinson (2015) suggest that that faster turns are more difficult to process, which causes the next speaker to

initiate their turn later. On the basis of a descriptive observation from that same study by Roberts et al. (2015) that faster turns are responded to *later*, not earlier, in their data (see Section 2.3.6), Levinson and Torreira (2015) also speculate that listeners have more time to prepare a response when listening to slow than to fast turns, which allows them to initiate their responding turn earlier.

2.3.4 Summary of predictions

To summarise, the coupled oscillator models make the following predictions:

1. The speech rate of a turn is correlated with the timing of the subsequent turn: faster turns are associated with shorter TTT and vice-versa. I will refer to this hypothesis as the perception-action entrainment hypothesis.
2. TTT distributions are not normal distributions but have a periodic component.

The next few sections review evidence regarding these predictions. Section 2.3.5 first discusses evidence on the presence or absence of periodic structures in TTT. Section 2.3.6 discusses empirical reports on the relationship between turn timing and global and local speech rate – the prediction that Chapters 4 and 5 focus on.

2.3.5 The structure of TTT distributions

The coupled oscillator models propose that the mechanism that produces turn timing is cyclical. Based on this feature of their model, Wilson and Wilson (2005) predict that TTT distributions should show evidence of a periodic structure. More specifically, they also predict that TTT distributions should cluster at around half a syllable's length before and after 0ms TTT, while TTTs of exactly 0ms are rare.

Beňuš (2009) tested this prediction using a dataset of TTTs automatically extracted from a corpus of twelve spontaneous, task-oriented, dyadic conversations. They measured turn timing in two ways: as speech activity offsets, and as the latency of high energy peaks associated with pitch accents across turn switches. They observed, first, that histograms of the distributions did not have two peaks on either side of 0ms. However, the location of the highest density of their distribution of speech activity offsets was around 100ms, consistent with the prediction that TTT distributions should peak at around half a syllable's length.

To the contrary, Franich (2023) found that the distributions of positive TTTs of ten speakers in a yes-no guessing game each tended to be bimodal: Six of the ten had a larger and smaller peak, although both on the positive side of 0ms, while the other four had a heavy right shoulder (i.e., showed a second area of high density next to the peak, that did not include a peak itself).

However, although Wilson and Wilson's (2005) prediction seems clear and easy to test using visualisations, the degree to which a distribution is visualised with multiple peaks is influenced by how it is plotted (e.g., number of bins or smoothing function used), and Franich (2023) furthermore had few datapoints; only 370 observations across the ten speakers. Additionally, variation in the pre-turn switch speech rates could obscure the expected periodicity in observations of raw TTTs.

In response to this last consideration, Beňuš (2009) also presents a histogram of TTT normalised by the average duration of the syllables in the preceding turns, which they predicted should show clusters around -.5, .5, 1.5, etc. However, while his distribution peaks at the expected .5, there are no further clusters.

Another complicating factor about the prediction that TTTs should cluster before and after 0ms TTT, is that the multiple clusters might obscure each other. Turns initiated at earlier or later cycles of the coupled oscillatory system might be expected to be less frequent than those at a cycle just after the previous turn end to such a degree that they are not visible.

To account for this possibility, Wilson and Zimmerman (1986) tested for periodicity using autoregressive integrated moving average (ARIMA) modelling of normalised (by subtracting each 10ms bin from the previous) and smoothed TTT distributions from seven conversations. They found significant periodic components with phases between 80 and 180ms.

Similarly, O'Dell et al. (2012) also attempted to detect periodicity by fitting four mixture models of increasing complex combinations of normal distributions to a TTT distribution from only a single conversation, and found no evidence that the models with multiple components were a better fit than a model without.

In short, Wilson and Wilson's (2005) predicted clustering has not been widely observed, and Wilson and Zimmerman's (1986) evidence for periodicity has not been replicated by O'Dell et al. (2012).

2.3.6 *Speech rate and turn timing*

This section reviews evidence for and against the perception-action entrainment hypothesis – the prediction by the coupled oscillator models that faster speech rate is associated with shorter TTT. Both observational evidence from spontaneous conversational data and experimental evidence is considered. In the following discussion, note that the direction of the predicted relationship between speech rate and turn timing depends on the way speech rate is expressed: if speech rate is measured as the number of linguistic units per time unit (e.g., syllable per second), the coupled oscillator models predict a *negative* relationship. However, if speech rate is expressed as time per unit (e.g., average duration of syllables in milliseconds), the entrainment models predict a positive relationship.

Firstly, Beňuš (2009) used data from spontaneous conversations to test several of the predictions outlined by Wilson and Wilson (2005), including the prediction that the speech rate of a turn should correlate with the subsequent TTT (see also Section 2.3.5). Using a corpus of twelve task-oriented dialogues in American English, they automatically extracted measurements of speech rates (from lexically defined syllables) over inter-pausal units, and TTT, operationalised as the automatically extracted time between the onsets and offsets of subsequent inter-pausal units by different speakers. Each turn switch was coded for the type of turn switching behaviour by two annotators, with categories denoting whether they were smooth switches (i.e., not pause-interruptions or backchannels), 'butting-in', backchannels, interruptions, non-interruptive overlaps, interruptions during a pause, or continuations after a backchannel.

Through a series of univariate correlation analyses, Beňuš (2009) reports finding no significant correlation between speech rate and TTT correlate in the data overall, but a significant correlation in one of the twelve conversations, and in the specific cases of pause-interruptions. When using pitch-accent rate as a measure of speech rate, there was a significant overall correlation between pitch accent rate and the latency of the first pitch accent by the next speaker, again with the strongest correlation in pause interruptions. They interpret these observations as weak evidence in favour of Wilson and

Wilson's (2005) prediction, and speculates that timing is particularly important to manage pause interruptions, leading to a stronger correlation in that subset of the data.

Also using spontaneous conversational data in American English, this time from telephone dialogues from the Switchboard corpus (Godfrey et al., 1992), Roberts et al. (2015) observed an opposite relationship between speech rate and turn timing. They measured TTT as automatically extracted offsets between stretches of speech activity from different speakers, and 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. They then used a random forest analysis (Breiman, 2001) to rank predictors of TTT in their corpus and found that speech rate was the fourth most important predictor, after speech act, the duration of the first turn, and the duration of the second turn. However, a random forest algorithm does not explicitly model the direction and shape of the relationship between variables. It is an ensemble machine learning algorithm that involves sub-setting the data and predictors and training many decision trees. Random forest models can thus implicitly account for non-linear relationships and interactions. To explicate the relationship between speech rate and turn timing, Roberts et al. (2015) report (descriptively) that, contrary to the perception-action entrainment hypothesis, the fastest quartile of turns was followed by a TTT that was 100ms longer than the slowest quartile of turns, if they restricted the turn sequence context by excluded backchannels and short turns. Based on this observation, Levinson and Torreira (2015) suggest that speech rate indirectly affects turn timing by affecting the difficulty involved in processing the turn, as discussed in Section 2.3.3.

On the other hand, Fujiwara et al. (2007) observed, in a corpus of dyadic role-playing games in Japanese, a high correlation between a dialogue act's speech rate (morae/utterance length) and its TTT from the previous dialogue act, at $r=-1.00$, in a subset of dialogue acts that they term 'responses' in their paradigm. However, their TTT measurements captured not only turn timing, and may also represent within-speaker pauses.

On a smaller scale, two other observational studies found no evidence in either direction. First, O'Dell et al. (2012) hypothesised that scatterplots of turn timing against speech rate should show a fan-like structure, where each line represents one cycle of the entrained system. However, in their data from one conversation, they found no such structure.

Second, Kendrick and Torreira (2015) compared the timing of 195 preferred and dispreferred responses, and found that the TTT distributions of the two types overlapped, even though they expected dispreferred responses to be timed later. They suspected that that overlap in TTT distribution might be an artifact of speech rate differences between the two types but found no significant effect of speech rate in mixed linear regression models predicting TTT from speech rate. However, their data included many cases with potentially pragmatically marked turn timing, which may have obscured any neutral relationship between speech rate and TTT (see Section 2.4).

Stronger support for the entrainment models came from experimental studies. Notably, Corps et al. (2020) tested Garrod and Pickering's (2015) prediction that speech rate affects turn timing in a question-response paradigm. Their participants listened to prerecorded polar questions and were asked to respond 'yes' or 'no' as quickly as possible. The rate of the final word (which was monosyllabic), and the rate rest of the question, were independently manipulated by either compressing it by a factor of .5 or presenting it at the natural rate, creating 2 by 2 design with fast/slow context speech rate and fast/slow local speech rate conditions. With regard to the rate of the question, Corps et al. (2020) found that participants responded earlier to speeded than to non-speeded questions, in accordance with the perception-action entrainment hypothesis.

Interestingly, Corps et al. (2020) measured response time in two ways: as the time from the end of the final word of the question to the start of the response (corresponding to the usual operationalisation of TTT), and as the time from the start of the final word of the question to the start of the response. They argue that measuring response time from the offset of the final word confounds response time with the amount of time the participant had to prepare their response: during longer final words, they had more time to prepare, so non-speeded final words should be associated with faster response time by virtue of that extra preparation time. This consideration is especially important when observing the effect of the speeded vs non-speeded question context.

However, if we assume that participants are aiming to start speaking at the end of the final word (because they aim to minimize gaps and overlaps), then measuring response time from the start of the final word likewise confounds it with the duration of the final word, because speeded final words would be associated with faster response times because one component of the response time, the final word, is shorter. If so, this

operationalisation may not accurately reflect observe the effect of the final word rate. The contrast between these two measurements highlights that the usual operationalisation of turn timing as the time between the end of one turn and the start of the next, rests on the assumption that this span accurately reflects the speaker's speech timing. Chapter 6 returns to this issue.

Using their second, favoured operationalisation, Corps et al. (2020) found that a speeded final word was associated with faster response times, which they interpret to indicate that participants adjusted their speech rate representation after a single syllable. On the contrary, using the first, more conventional measurement, they found that a speeded final word was associated with a slower response time. A possible interpretation of this second effect is that final lengthening is a turn-ending cue (Duncan, 1972), and a more salient or earlier turn-ending cue causes speakers to time their turns earlier (Barthel, 2020; Torreira & Bögels, 2022). Another is that when the final syllable is shortened, response times may be slower because the coupled oscillator system has not adjusted to it (i.e., an effect opposite to local phase resetting). However, Corps et al. (2020) point out that in both those cases, an interaction between context speech rate and local speech rate would be expected, but they found no significant interaction.

In short, Corps et al.'s (2020) main effect of the speeded versus non-speeded conditions support the prediction that faster speech rates are associated with faster response times, contrary to the observation by Roberts et al. (2015), but as predicted by Garrod and Pickering's (2015) and Wilson and Wilson (2005). However, given the ambivalent results by primarily Beňuš (2009) and Roberts et al. (2015), it is not clear if they use that entrainment-type responding style in spontaneous conversation as well.

These conflicting results between the experiment and observational studies may thus relate to methodological differences. First, the demand characteristics of a question-response paradigm are likely to be quite different from the demands of spontaneous conversation and may therefor trigger a different production mechanism. Along these lines, Meyer et al. (2018) propose comparing the estimates of response times from experimental studies to TTT estimates from spontaneous conversations to determine whether the experiment successfully isolated an aspect of conversational behaviour. They suggest that the more similar the experimental estimate of response time to those from conversations, the better the paradigm captured turn timing behaviour. Meyer et al.

based their own comparisons on the estimates reported in the previously mentioned study of polar question-response sequences by Stivers et al. (2009). Corps et al. (2020) point to this argument and note that the TTTs they observe are longer than the expected 200ms (basing their expectation on Stivers et al., 2009), with means of 582ms and 452ms and modes of 713ms and 245ms, which they take as an indication that their experiment may not have evoked fully ecological turn taking behaviour. Furthermore, they report instructing their participants to respond as fast as possible. Conversely, experimental paradigms eliminate some of the information present in spontaneous conversations, such as turn ending cues or social- or common ground information. Both these aspects of experimental design may evoke behaviour that is either different in type or in degree from conversational behaviour.

Another difference between the experimental and observational studies may relate to the dialogue acts they investigated: the experiment by Corps et al. (2020) used a polar question-response paradigm, while the observational studies included a range of dialogue acts, and do not report using multivariate statistics to understand possible interactions between the dialogue act and the relationship between speech rate and turn timing. However, dialogue acts may differ in the degree to which speakers are oriented to responding (Corps et al., 2022). Specifically, questions typically select the next speaker, meaning that a source of potential complexity (the process of choosing whether to select yourself as the next speaker) is simplified. As spontaneous questions arguably have this characteristic in common with experimental trials, it is possible that turn timing in spontaneous question-response sequences may pattern with experimentally observed behaviour. The observation by Fujiwara et al. (2007) regarding the relationship between speech rate of the second turn and turn timing suggest a similar dialogue-act dependent effect. Such considerations lead some researchers of spontaneous conversation to focus on question-response sequences to observe speech timing effects (e.g., De Vos et al., 2015), and Meyer et al. (2018) suggest that experimental paradigms and spontaneous data are best compared when using the same type of dialogue act sequences

For these reasons, the corpus study presented in Chapter 4 tested the perception-action entrainment hypothesis specifically in spontaneous question-response sequences, the dialogue act types used in experimental paradigms, and uses multivariate statistical techniques to test if any observed relationship between speech rate and turn timing is

modulated by the type of question and response. It also uses hand-aligned data rather than automatically extracted measurements, aiming to represent speech timing as accurately as possible (see Section 2.2.2).

More recent studies align to some extent with this interpretation that experiments and spontaneous conversations may engage distinct behavioural mechanisms. First, in a study on various aspects of the speech of simultaneous interpreters during counselling sessions, Vranjes and Oben (2022) measured the timing of interpreters' speech relative to the end of the turn they were interpreting in 96 question-interpretation-response sequences. They observed that speech rate was a significant negative predictor in a mixed linear model predicting the timing of the interpreter's speech, in support of the perception-action entrainment hypothesis. Second, Franich (2023) measured the timing of responses in a dyadic yes-no-guessing game. They observe that questions with faster speech rates were associated with faster response timing in a mixed multiple regression analysis of 370 questions. They also found that longer final words in the questions were associated with faster response timing, similarly to Corps et al.'s (2020) findings when they used the end of the last word to measure response timing.

Both these studies align with Corps et al.'s (2020) experimental findings in favour of the perception-action entrainment hypothesis. However, while interpreting and guessing games are naturalistic interactions, it is unclear to what extent they both access the same behaviour as spontaneous conversation. The timing of interpreters' speech may involve different behaviour than spontaneous turn timing, just as experiments do, since the content of their turn is restricted. Similarly, a yes-no guessing game restricts the content of turns.

Along similar lines, Torreira and Bögels (2022) asked Dutch-speaking participants to say *ja* ('yes') as fast as possible in response to stimuli that consisted of a string of repetitions of the syllable /ma/. They found that participants responded earlier after a fast string (200ms per syllable) than after a slow string (300ms per syllable) but only when no prosodic manipulations similar to turn-ending cues were present. When the final syllable was lengthened 1.5 compared to the other syllables, or a phrase-final pitch contour was inserted, faster strings were associated with slower response times, possibly because the turn-ending cue appeared closer to the end of the stimulus in the fast case, leaving the participant less time to respond. Torreira and Bögels (2022) conclude that while

perception-action entrainment may constrain turn timing behaviour, this effect might be overshadowed by other factors, such as reaction to turn ending cues. They suggest that this might be the case in their experiment because they instructed the participants to respond as fast as possible. Observing that their response times had means and medians that are similar to estimates from spontaneous conversation (between 208ms and 284ms; Torreira & Bögels, 2022, p. 7), using a similar argument to Meyer et al. (2018), they suggest that speakers often have the same objective in conversation.

To summarise, corpus studies support a range of conclusions regarding the perception-action entrainment hypothesis, while evidence from experiments and more constrained interactional contexts more clearly supports the hypothesis.

2.3.7 Summary and research questions

Empirically, what distinguishes coupled oscillator models of turn timing from other models is the prediction that faster turns should be associated with shorter TTTs. Studies have come to opposite conclusions: Based on data from spontaneous conversations, Beňuš (2009) reported a negative relationship in only some subsets of the data, while Roberts et al. (2015) reported a positive relationship. Conversely, the experimental study by Corps et al. (2020) was in line with the perception-action entrainment hypothesis. These conflicting results may relate to methodological differences.

To investigate this apparent contradiction, Chapter 4 presents a study of spontaneous question-response sequences, addressing the following research questions:

2. Is speech timing (particularly speech rate) predictive of turn-transition time, and if so, what is the nature of their relationship, and is it modulated by other factors?

Hypothesis-testing

- 2.1 Is there an inverse relationship between the Question's Speech Rate and TTT?
- 2.2 Is there an inverse relationship between Final Lengthening (Final Rhyme Duration) and TTT?

Exploratory questions:

- 2.3 Is the relationship between the Question's Speech Rate and TTT best described as linear or as quadratic?
- 2.4 Are any observed effects of speech timing factors modulated by dialogue factors already shown to influence TTT (Question Length, Question Type, Response Length and Response Type)?
- 2.5 Is there an inverse relationship between the speech rate of the response and TTT?
- 2.6 Do we observe any reliable differences, in the above factors, between Dutch and English?
- 2.7 What is the shape of the TTT distributions in this data?

2.4 Turn 'Timeliness' and Turn Timing Perception

So far, Section 2.2 has discussed that turn timing is often considered 'precise', but that there appears to be considerable variability as well, of which the scope is worth defining. Section 2.3 described psycholinguistic models that aim to explain how turn timing can be precise to the extent that it is, and reviewed evidence for and against those models based on observations of the relationship between speech rate and turn timing.

However, as mentioned in Section 2.2.4, turn timing is also affected by many factors. Some of that variation in turn timing is salient to listeners and interlocutors as pragmatically meaningful delay (F. Roberts et al., 2011) or disfluency (Van Os et al., 2020). Models that aim to account for turn-timing precision therefore assume that there is an expected turn timing behaviour against which meaningful delays or disfluencies are interpreted. For example, coupled oscillator models can be understood to that the predicted relationship between speech rate and turn timing is different when turns are timed 'on time' than when they are late or early³. To test this prediction, we need a way to

³ Along similar lines, Levinson and Torreira present their model as accounting specifically for typical cases of turn timing: "The model therefore captures the most typical turn transition values observed in conversational corpora." (Levinson & Torreira, 2015, p. 14). To test it, it would therefore be useful to distinguish the cases it accounts for from those it does not account for independently from the predictions of the model.

distinguish timely turns from non-timely turns independently from the model's predictions (cf. Beňuš, 2009).

One way to do so is through perception studies, by operationalising listeners' perception of the timing of turns. However, turn timing perception has not been as widely studied as turn timing production, especially not from a quantitative perspective. Additionally, listeners' perceptions and interpretations of turn timing are potentially non-linearly related to each other, and also to other aspects of the linguistic context as well.

The next three sections review studies that have observed relationships between turn timing variation and listeners' interpretations of 'fluency', 'naturalness', 'smoothness' or 'timeliness', and how these interpretations may be related to each other and other aspects of turn switches. These studies have usually taken one of three possible perspectives. Some asked how turn-timing variation relates to listeners' interpretations of constructs such as the fluency, naturalness or smoothness. Other asked the related question of how specific response latencies are associated to pragmatic interpretations of delay in specific instances. From a third perspective, a very limited number of studies have asked how listeners *form perceptions of* turn timing, based on which they then form interpretations of its meaning⁴. The next few sections discuss each of these approaches in turn.

2.4.1 Turn-timing fluency and naturalness

Firstly, some studies have aimed to determine how constructs such as fluency, naturalness, smoothness or timeliness relate to TTT. Their observations cannot straightforwardly be applied to new spontaneous data, however. Depending on their approach, they can introduce several problems, as this section discusses.

One line of research aimed to determine what TTTs can be considered 'smooth' transitions by identifying a threshold from which turn timing becomes audible as an overlap or gap. For example, Walker and Trimboli (1982) asked participants to listen to 5 minutes of a continuous conversation and indicate at which turn transitions they heard a

⁴ Here, 'turn timing interpretation' refers to conclusions that interlocutors or listeners draw from turn timing variation. 'Turn timing perception' refers to how interlocutors or listeners come to those inferences from the perceptual input they receive. This contrast is not meant to imply that turn timing perception is either a lower-level or a higher-level perceptual phenomenon.

pause. They recommended that positive TTTs of up to 200ms should be considered smooth transitions.

Similarly, Heldner (2011) asked participants to indicate when they heard gaps and overlaps, and determined that overlaps shorter than 120ms TTT and gaps of shorter than 120ms TTT were perceived as immediate turn transitions. While informative about pause and overlap perception, the aim of these studies appears to have been to determine which turn transitions conform perceptually to the no-gap-no-overlap principle. However, assuming that ‘smooth’ transitions, operationalised as just outlined with a perceptibility threshold rather than a contextual acceptability threshold, are also ‘optimally’ timed, risks turning the no-gap-no-overlap principle from a descriptive generalisation into a prescriptive yardstick. Along these lines, Heldner (2011) also notes that the duration of ‘unmarked’ TTT tends to be longer than his observed 120ms gap detection threshold.

Another study has examined at what temporal threshold turns stop sounding ‘on time’ and start sounding delayed, rather than when their TTT is audible as a gap: Stivers et al. (2009) asked an informed annotator whether responses sounded ‘on time’ to them, in polar question-response sequences with gaps (positive TTTs), for each of ten languages. They concluded that the threshold for judging turns as delayed varies slightly between languages – they observed thresholds between 203ms TTT in Danish and 36ms TTT in Japanese, and the three languages with the longest mean TTT (Danish, ǂǂkhoe Hai||om and Lao) significantly differed in their thresholds. This approach implies that there is something like a competence-performance distinction (Chomsky, 1965) in turn timing. On this view, speakers aim for a target behaviour, while single listeners can refer to their knowledge of the language system to reliably judge the speaker's success at hitting that target.

Other studies have collected participant ratings of manipulated read human speech (Van Os et al., 2020) or synthesised speech (Itoh et al., 2009) in controlled experimental designs. They concluded that turn timing affects listeners’ interpretation of the fluency or naturalness of turn switches, but timing and fluency or naturalness do not stand in a straightforward linear relationship.

Specifically, Van Os et al. (2020) asked L1 listeners of Dutch to listen to question-response sequences spoken by L1 Dutch speakers, in which the TTT was manipulated to range between -600ms and 900ms with steps of 300ms, and the speed of the response was manipulated to be slowed by a factor 1.20 or speeded by a factor .83 to create a slow and a fast condition. The participants rated the fluency of the responses. Van Os et al. observed that the relationship between fluency ratings and turn timing was not linear, since long negative and long positive TTTs were rated as less fluent than TTTs at 0ms (however, the authors model this relationship as linear in their inferential procedure). This relationship between timing and fluency suggests that the scale from non-fluent to fluent may not necessarily be congruent with a possible scale of timeliness from early to late.

This relationship additionally interacted with the speech rate of the response. Faster responses were considered more fluent when they were timed later, while slower responses were considered more fluent when they were timed earlier. In a second experiment with the same paradigm, Dutch L1 listeners rated stimuli in Turkish-accented Dutch. In this experiment, the interaction between speech rate and TTT was additive – listeners rated late responses as more disfluent if they were spoken at a slow speech rate than if they were spoken at a fast rate.

Based on these observations, Van Os et al. speculate that their participants interpreted the fluency of turn timing relative to the speech rate of the response, although their results strictly speaking do not demonstrate this effect directly. Rather, as the authors also suggest, listeners seem to form their judgement of fluency in part with reference to how eager or how reluctant the combination of the two features made the speaker sound, where fast, early responses sound too eager, and slow, late responses sounded too reluctant. In other words, they suggest that the fluency of responses is a construct that appears to be composed of other social judgements that are interpreted from combinations of features of speech.

This observed variability of turn-timing expectations is not surprising, as turn timing also conveys meaning (Roberts & Francis, 2013) and evokes different expectations about the content of the upcoming turn in different contexts (Bögels et al., 2020). Additionally, a range of features may contribute not only systematically to actual turn-timing behaviour, but also to listeners' perceptions of fluency or timeliness. Such features could include listeners' idiosyncrasies, as well as linguistic aspects of the actual turn switch.

In line with that expectation, Itoh et al. (2009) found that ‘naturalness’ ratings of dialogues between humans were highest when the TTT was not manipulated: dialogues with TTTs manipulated to be different fixed or proportional durations were rated as less natural than the original TTTs. On the contrary, when the same dialogues were presented in synthesised speech, a fixed TTT at 500 ms was rated as most natural. Itoh et al. conclude that ‘a natural response time exists’ (Itoh et al., 2009, p. 1835), but it depends on a range of factors. They suggest that those factors might include the acoustic properties of the context that differed between the original speech and the synthesised speech, as well as properties of the discourse.

To summarise, interpretations of timing can potentially not directly be mapped onto perceptions of the presence of a pause. Furthermore, interpretations of turn timing fluency or naturalness are affected by other features of the context than just turn timing. Consequently, it is not clear how experimentally observed threshold of timeliness or smoothness can be applied to observational studies to distinguish between well-timed and badly timed turns without biasing the results of further analyses.

2.4.2 Interpretations of delay

As mentioned, turn timing does not only relate to impressions of fluency or smoothness, but is also related to more specific pragmatic meaning, as evidenced by a large body of qualitative work in the framework of Conversation Analysis (e.g., Couper-Kuhlen, 1993). From a quantitative perspective, a similar line of research to the previously mentioned studies on fluency or smoothness aimed to understand how responses with TTTs of particular lengths lead listeners to assign different pragmatic interpretations to those responses.

First, Bögels et al. (2020) asked participants to listen to question-response sequences while recording EEG. They found that preferred responses after a longer gap were associated with an earlier negative ERP compared to dispreferred responses, which the authors interpret to indicate that participants expected the preferred response to be timed earlier.

Using an explicit, behavioural paradigm to access listeners’ interpretations, Roberts & Francis (2013) asked participants to rate manipulated, acted dialogues on a scale denoting the responding speaker’s perceived ‘willingness’, ‘enthusiasm’ or ‘agreement’.

They found that listeners' ratings of responses as conveying agreement, enthusiasm or willingness drop after TTTs of around 600ms or more compared to shorter TTTs: responses after longer positive TTTs are interpreted as conveying less willingness, enthusiasm or agreement. As they presented their participants with stimuli in which only the TTT was manipulated, in a crossed experimental design, it appears that the duration of TTTs alone is enough to shift participants' interpretation, and no other cues are needed.

Another study by some of the same authors observed that listener's ratings drop as positive TTT durations increase (F. Roberts et al., 2011), both when they rate the speaker's willingness to comply with a request or when they rate the speaker's agreement with an assessment, but to different degrees in these different types of dialogue act sequences. Additionally, this effect happened in Italian, Japanese and American English alike, but to different degrees in each language.

These experimental findings suggest that timing expectations depend on context, and that the way listeners come to pragmatic interpretations of turn timing may be subject to a range of effects, including cross-linguistic differences and dialogue act. Additionally, the speech context may affect listeners' *perception* of the duration of delay itself, in addition to their interpretation, as the next section reviews.

2.4.3 Perception of turn timing and speech-rate normalization

Aside from asking how concepts like fluency, timeliness and pragmatic interpretations are related, we can ask how listeners perceive turn timing as early, timely or late in the first place. At face value, we may expect that listeners simply perceive the absolute duration of TTT and compare it to some expectation, that may depend on context, to come to an interpretation. However, there are also reasons to expect that turn timing perception is not absolute in that sense, as this section reviews. Specifically, the perception of other durational contrasts in speech has been widely shown to be subject to speech-rate normalisation.

Speech-rate normalisation is the effect that listeners' perception of a variety of durational feature of speech is depends on the rate of the speech in which the relevant feature is embedded: in faster speech-rate contexts, the threshold for perceiving a particular durational feature as long shifts to earlier. For example, the voice-onset time threshold for categorizing a plosive as voiced (Diehl et al., 1980), or the threshold for categorizing a

phoneme as long in a length contrast (Bosker & Reinisch, 2017), is earlier in fast speech-rate contexts. Even the perceived presence or absence of short function words, such as ‘or’, is affected by speech-rate normalization (Dilley & Pitt, 2010). Entrained neural activity at the timescale of syllables (Section 2.3.2) has also been hypothesised to be responsible for such speech-rate normalisation effects, by Wilson and Wilson (2005) and Garrod and Pickering (2015), and more recently Bosker and Ghitza (2018).

Speech-rate normalization effects hold under experimental conditions that mimic monologue, but also under the following conditions that are similar to the conversational context in which turn timeliness is perceived. First, speech-rate normalization experiments often compare speech rate conditions with artificially compressed or expanded speech, resulting in extreme speech rate values relative to the range of speech rates in natural speech. However, Nagao & De Jong (2007) demonstrated that these effects are also found when speech-rate conditions are constructed from unmanipulated recordings of strings of meaningless syllables spoken to a metronome, representing more naturalistic speech rate ranges.

Second, speech-rate normalization does not just affect categorical distinctions: continuous durational cues to prosodic boundaries are also interpreted with reference to the preceding speech rate (White et al., 2022).

Third, speech-rate normalization effects also appear to transfer across speakers. In an experimental study where the perception target was spoken by a different speaker than the rate prime carrier utterance, listeners still seemed to normalize the target for speech rate (Newman & Sawusch, 2009). Similarly, a listener’s own previous speech rate can have a rate-normalisation effect on target contrasts in experimental stimuli (Bosker, 2017b).

To summarize, speech-rate normalization effects appear to hold in contexts similar to turn timing: turn timing also represents a continuous contrast in meaning (cf. White et al., 2022), it is also found in the relation between speech by two different speakers (cf. Bosker, 2017b; Newman & Sawusch, 2009), and it is perceived in relation to spontaneously uttered speech-rate variation (cf. Nagao & De Jong, 2007). These findings suggest a hypothesis that turn timing may also be perceived relative to the previous turn’s speech rate.

Along this line, Couper-Kuhlen (1993) suggested that interactionally meaningful aspects of turn timing are constructed by the interlocutors in relation to the rhythm of the conversation as a whole. They suggest that this rhythm is carried by metrically prominent segments (stressed syllables that function as ‘beats’) and presents qualitative analyses to support this idea. They also report asking ten Conversation Analysts to identify pauses that they would mark on a transcript as potentially relevant and found that those that broke the metrical structure or encompassed a whole ‘beat’ were most often noticed.

Kendrick and Torreira (2015) consider this idea from a quantitative perspective. In the previously mentioned study comparing the timing of preferred and dispreferred responses, they expected to find that dispreferred responses are timed later than preferred responses (see Sections 2.2.2 and 2.3.6), but found that it was dispreferred *formats* that were delayed compared to preferred formats. Informed by Couper-Kuhlen’s (1993) framework, they speculate that the expected difference in timing between preferred and dispreferred turns may not lie in absolute timing, but in perceived timing, which may be affected by the rate of the previous speech. To test whether this unexpected observation was therefore an artifact of differences in speech rate between the preferred and dispreferred responses, they performed mixed linear regression analyses predicting turn timing from speech rate in 195 turn switches and found no significant effect. However, this analysis does not directly test whether *perceived* turn timing is affected by the preceding speech rate, since it does not account for differences in perceived timing.

Similarly, Roberts & Francis (2013) consider the possibility that the turn timing is subject to speech rate normalisation as an explanation for their finding that Italian, Japanese and American English speakers judge delays to different degrees at particular durations of TTT. They reject it as untestable on the grounds that speech rate in these languages is presumably intrinsically a different concept, because they are so-called syllable-timed, mora-timed and stress-timed languages respectively (cf. Couper-Kuhlen, 1993). However, the concept of a rhythm typology that describes an intrinsic difference between languages in what segment is the basic unit in its rhythmic structure is not currently widely supported, as empirical evidence better supports the idea that those variations in speech timing between languages arise from other differences in their phonology, such

as the complexity of consonant clusters (Dauer, 1983). Additionally, the neural tracking of speech rate that is hypothesised to be responsible for rate normalisation effects (Bosker & Ghitza, 2018) also happens over a range of frequency bands corresponding to the rate of syllabic or stress units, but independently from the language listened to (Song & Iverson, 2016).

A rate-normalization effect on turn-timing perception would also be expected by the coupled oscillator models of turn-timing behaviour for the following independent reason. The theories propose that turn timing *production* is done in relation to forgoing speech rate. If listeners can identify ‘fluent’ or ‘natural’ timing production, then they would be expected to base their interpretations of turn timing on the TTT relative to the foregoing speech rate. A speech-rate normalisation effect on turn timing perception is therefore implied by coupled oscillator models, although not vice-versa.

In short, there is little quantitative evidence for an effect of speech-rate normalisation on turn-timing perceptions, even though speech-rate normalisation affects the perception of many similar aspects of speech, and would be predicted to affect turn timing by the coupled oscillator models of turn timing (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005).

2.4.4 Summary and research questions

To summarise, not all turns are timed equally fluently (Van Os et al., 2020), naturally (Itoh et al., 2009), or smoothly (theoretically, at least; Heldner, 2011). Additionally, variation in TTT can convey pragmatic meaning (F. Roberts et al., 2011). The study presented in Chapter 5 aimed to operationalise the perceived timeliness of turns through listener ratings, to test whether the relationship between speech rate and turn timing predicted by the coupled oscillator models is modulated by salient variations in timing. This study opted to explicitly ask participants to judge timeliness of turns in turn switches from spontaneous conversations, building on the study by Stivers et al. (2009) with a large participant group, to enrich the corpus data with a perceptual variable. The main research question of Chapter 5 is:

3. What is the relationship between speech rate, turn timing, and perceived turn ‘timeliness’?

3.1 Specifically, is the relationship between speech rate and turn-transition timing in question-response sequences modulated by participant ratings of turn timeliness? In other words, is there a consistent relationship between speech rate and turn timing when turns are perceived as ‘on time’?

As this section discussed, the hypothesis that speech rate normalisation affects turn timing perception does not appear to have been tested. For this reason, Chapter 5 also presents a secondary analysis of the corpus data, targeting that question:

3.2 Do speech timing features (particularly speech rate) of the question (and/or responses) affect participants’ ratings of response timing?

2.5 Summary and Outlook on Empirical Chapters

This chapter has reviewed conversational turn timing research from several perspectives. First, the smoothness, precision and speed of turn timing are often emphasised in the literature, and many studies point to estimates of around 200ms to make this point. However, it is worth defining the degree of variation in turn timing systematically. To contribute to that endeavour, Chapter 3 presents a systematic review of turn timing samples.

Second, coupled oscillator models that aim to account for ‘smooth’ turn timing predict that faster turns are associated with shorter turn-transition times. Previous experiments and observational studies have found contrary results. Chapter 4 presents a corpus analysis of speech rate and turn timing in the dialogue acts often used in experimental paradigms – question-response sequences – but from spontaneous conversation, to investigate whether they pattern with the question-response sequences from experiments, or the general (weak) pattern in spontaneous conversation. Chapter 4 also investigates whether any relationship between speech rate and turn timing is modulated by other dialogue- or speech timing variables.

Third, the study in Chapter 5 operationalised the timeliness of turns to test whether perceived timeliness modulates the relationship between speech rate and turn timing.

This chapter also discussed implications of choosing particular operationalisations of turn timing. The systematic review in Chapter 3 considers whether the operationalisation used by included studies is responsible for systematic differences in turn timing

estimates. Chapter 4 and 5 use (mainly) an operationalisation that aims to access turn timing as a form of speech timing. Based on results from the rating study in Chapter 5, Chapter 6 returns to the topic of turn timing operationalisations.

Chapter 3. Turn Timing in Conversation: A Systematic Review

3.1 Introduction

Conversational turn timing is widely considered a ‘precise’ behaviour, in the sense that it is fast and apparently cross-linguistically so. However, estimates of central tendency vary between samples to a large degree, and multiple linguistic and non-linguistic factors are known to affect turn timing. There currently appear to be few systematic attempts to collect a comprehensive overview of the factors that have been found to affect it and to assess how robust those effects are, or to describe the range of variation of turn timing distributions reported in the literature. To fill this gap, we performed a systematic review of turn timing in spontaneous conversation between adult humans. This chapter presents an analysis of turn timing distributions and study-level variation in turn timing estimates.

The remaining subsections of Section 3.1 first briefly recap the discussion in Section 2.2 on turn timing distributions to outline the motivation of this study (Section 3.1.1), provide an overview of studies with similar aims (Section 3.1.2), and introduce the method of systematic reviewing (Section 3.1.3). Section 3.1.4 states the research questions and discusses specific methodological choices.

3.1.1 Motivation

The idea that turn timing is a precise phenomenon originates with a study by Sacks et al. (1978), who proposed that speech during conversation is organised to minimise gaps and overlaps. This idea is often supported with reference to a cross-linguistic study by Stivers et al. (2009), who reported a mean turn-transition time (TTT) across ten languages of only 208ms, with the mean TTTs of individual languages ranging from 7ms TTT for Japanese to 469ms TTT for Danish – at most 261ms from the cross-linguistic mean. This and similar observations led researchers to develop theories of language processing during turn taking that account for such fast timing (Garrod & Pickering, 2015; Levinson & Torreira, 2015; M. Wilson & Wilson, 2005) often arguing that the behaviour of turn timing needs to be understood as more than just reaction time (but for arguments disputing this view, see Heldner, 2011; Heldner & Edlund, 2010).

As section 2.2.3 discussed, however, the claim that turn timing is ‘precise’ is not well-defined by itself and can be challenged by a number of observations, such as that turn timing distributions tend to have a wide range (Heldner & Edlund, 2010; Stivers et al.,

2009), that many factors influence turn timing (S. G. Roberts et al., 2015), and that very different measures of central tendency are also reported (e.g., a mean of 1500ms in Alpert et al., 1997). These discrepancies, along with any consistent characteristic of turn timing distributions, should be accounted for by models of turn timing production.

One way to describe the scope of precision and variation in turn timing is to form specific hypotheses and to test them experimentally on new data using quantitative or qualitative methods. Another, more exploratory, approach is to systematically attempt to find patterns in turn timing estimates from existing samples first. This study took the second approach.

3.1.2 Previous research

Some previous research has similarly attempted to address these questions by synthesising secondary data. They have either focused on turn timing in a specific population or collated turn timing distributions in literature reviews. Following the former strategy, Nguyen et al. (2021) conducted a systematic review of turn timing specifically by infants in infant-adult interactions and found inconclusive evidence for any modulating factor other than the adult's turn timing. In the context of pupil-teacher interactions, Tobin (1987) reviewed studies that addressed the phenomenon of 'wait time', defined as the pauses between utterances in the interaction, and concluded that longer wait time was associated with better performance.

Following the strategy to synthesise data from spontaneous conversations between adults in general, Heldner and Edlund (2010) include an (apparently not systematically assembled) overview of some turn timing estimates from the literature in their discussion of what proportion of turn timing was late enough that they could in theory be due to speakers reacting to the end of the previous turn rather than to prediction. They report means between -460ms and 664ms, and medians, which are reported less frequently in their sample but appear more stable, between 264ms and 451ms. Similarly, Levinson and Torreira (2015) cite averages between -78ms and 480ms from ten samples reported in four studies (they do not specify what measure of central tendency they include).

Notwithstanding these efforts, there does not currently appear to be a systematic review of turn timing in spontaneous conversations between adult speakers. The study presented in this chapter was designed to contribute to the discussion by systematically

identifying the range of reported estimates of turn timing in adult populations, and to attempt to find consistent influences on turn timing, through the method of systematic reviewing.

3.1.3 *What is a systematic review?*

Systematic reviewing is a method that is used to answer empirical research questions by creating a comprehensive or well-delimited synthesis of existing research outputs. It is often used to collect previously reported results as secondary data in a further analysis. The design of systematic reviews aims to make their search for secondary data of this kind a reproducible endeavour. To do so, a protocol is written in advance and ideally preregistered. The protocol identifies the search strategy, including the databases accessed, which exact search string is used to search them, and on what date the search is conducted. It also identifies the criteria for including and excluding search results in the review, how many reviewers are involved, and how the identified data will be synthesised. Standard guidelines exist for reporting systematic reviews, such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Page et al., 2021) checklists. These methods and guidelines were also followed in the study in this chapter, which was preregistered (see Section 3.2) and is reported here following the PRISMA guidelines.

3.1.4 *Research questions and methodology*

To summarise, this study aimed to describe the variation between studies in the distribution of turn timing in adults and attempt to find patterns that might explain it. Its research questions were formulated as follows:

1. Are assumptions about turn timing often cited in the literature supported by empirical evidence, and what is the existing scope of variation in turn timing?
 - 1.1 What estimates of turn timing are reported in studies of spontaneous conversation between adults?
 - 1.2 What operationalisations of turn timing are used and what are their differences?
 - 1.3 What factors are found to have a consistent effect on turn timing?

1.4 To what extent is there variation between turn timing estimates that cannot be accounted for by factors considered within the studies?

This chapter aimed to address these questions primarily through an exploratory, qualitative synthesis of quantitative results of the systematic review. This synthesis addresses variation in turn timing estimates and distributions at the level of studies, not within studies. The search protocol of the systematic review was designed to also identify studies that report effects of particular variables only, but not overall turn timing estimates or distributions. These studies are not included in the synthesis in this chapter. They were excluded following the procedure in Section 3.2.5, in order to focus the synthesis on turn timing estimates and distributions, between-study variation in those distributions, and potential methodological effects on that variation.

The present study focuses on adult speech, as opposed to language development, complementing the review of infant turn timing by Nguyen et al. (2021). The phenomena of interest were limited to turn timing in spontaneous conversation, so turn timing in conversation-like experimental tasks is not considered. The reason for excluding experimental data, even when the experiment was designed to evoke a specific aspect of conversational behaviour, is that the experimental conditions are likely to introduce task-related demands that are different from the kind of demands placed on a speaker during spontaneous conversation. We exclude experimental tasks to avoid confounding our observations between two potentially different behaviours. The guiding principle is that each of the aspects of conversation described in Section 2.2.1 (content, length, progression, number and timing) should not be determined by the experimental paradigm to study timing. For example, studies were excluded for constraining the exact content of permitted turns to ‘yes/no’ in order to isolate effects on turn timing (e.g., Corps et al., 2020; A. S. Meyer et al., 2018), while studies that used conversation-eliciting games like the map task (e.g., Anderson et al., 1991) but did not completely limit the content of turns were included.

A few terminological points should be made first. Different researchers use different terminology to refer to the phenomenon of turn timing, and to operationalisations of turn timing. In this chapter, as in the rest of this thesis, other researchers’ terminology is translated to ‘turn timing’, whenever they refer to the behaviour, and ‘turn-transition time’ (TTT) whenever they refer to specific measurements.

Additionally, the word ‘record’ is used in this chapter, but not elsewhere in this thesis, to refer to a publication as it is indexed in a database. The publication itself is referred to as a ‘report’. From here onwards, the word ‘study’ is used to refer to a research project presented in a report that aimed to address a research question. A report can include multiple studies. The word ‘corpus’ is used to refer to datasets collected for either that specific study or for an unrelated report or public use. A corpus can be used in multiple studies, and one study can use multiple corpora. As Section 3.2.6 discusses in more detail, the basic unit of observation in this chapter is a sample of turn timing observations taken from a corpus and reported in a study.

In the remainder of this chapter, Section 3.2 describes the procedure used to identify the selection of relevant reports, Section 3.3 presents and discusses the results, and Section 3.4 draws conclusions in relation to the research questions and discusses limitations of this review.

3.2 Methods

This section describes the search strategy and selection procedure used in this review. They are reported following guidelines proposed by the PRISMA framework (Page et al., 2021). The PRISMA and PRISMA-S (which focuses on the search strategy) checklists are included in Appendix A1. A PRISMA diagram summarises the exclusion steps in Section 3.2.3.

The methods applied in this chapter, including the inclusion and exclusion criteria, search strategy, screening, and data extraction procedure, were preregistered on the Open Science Framework website (Foster & Deardorff, 2017) at <https://doi.org/10.17605/OSF.IO/KVUCZ> on the 31st of March 2023. The protocol was written using ProSysRef, a proposed systematic review registration form that is not discipline-specific (Van Den Akker et al., 2020), as a reference point.

The following sections describe the search strategy (Section 3.2.1), the inclusion and exclusion criteria (Section 3.2.2), the screening process (Section 3.2.3), the study quality assessment (Section 3.2.4), additional selection criteria used to identify the samples included in this analysis (Section 3.2.5), the data extraction process (Section 3.2.6) and the methods used to analyse the results (Section 3.2.7).

3.2.1 Information sources and search strategy

Four databases were searched. First, the Linguistics and Language Behavioural Abstracts (Cambridge Scientific Abstracts, Inc, 2023) database was used as a domain-specific database relevant to the topic of this review. It was searched through the ProQuest interface. Second, PsycINFO (American Psychological Association, 2023) was used to include sources related to behaviour but not specifically language science. Third, MEDLINE (Ovid Technologies, Ltd, 2023) was used because it indexes medically-oriented publications. Both PsycINFO and MEDLINE were searched through the OVID interface. Fourth, Scopus (Elsevier, 2023) was used because it is a discipline-independent database, and it was searched through the Scopus interface. No further search strategies, such as citation chaining search (i.e., screening the records cited by, or citing, included reports; Boland et al., 2017), were performed.

The topic of this review, turn timing, is addressed by researchers working in different disciplines, such as psychiatry (e.g., Alpert et al., 1997) or telecommunications (e.g., Kitawaki & Itoh, 1991), in addition to psycholinguistics (e.g., Levinson & Torreira, 2015). As is often the case with phenomena investigated in multidisciplinary contexts, there is no single agreed-upon terminology used to describe turn timing. To account for this inconsistency as well as possible, the search strategy was developed as follows. Its structure was as depicted in (1):

(1) Conversation AND Timing AND NOT Child, Animal, Robot

The search strategy included keywords denoting ‘conversation’ or synonyms, keywords denoting ‘timing’ and synonyms, and keywords excluding concepts that this review does not consider (see Section 3.2.2 below), such as ‘child’ or ‘experiment’. The exact search queries were developed as follows: A set of nine key records that covered a range of approaches to turn timing was identified in an informal scoping review (Table 3.1). The search queries were refined by iteratively testing whether they found those records, until all records that were indexed in each database were found by the search queries. Table 3.1 lists which records were found by the final search query or not indexed by a particular database. The search query was applied to the abstracts, title and keywords of records, but not the full text. The search was limited to reports in English; no other limits were set.

Table 3.2 (next page) summarises the search strategy. It lists, for each database, the interface used to search it, the search term used to search it, and the number of records identified when the search was performed on the 31st of March 2023.

Article	Scopus	LLBA	PsycINFO	MEDLINE
Stivers et al. (2009)	Yes	Not in database	Yes	Yes
Heldner & Edlund (2010)	Yes	Yes	Yes	Not in database
Ten Bosch et al. (2005)	Yes	Yes	Yes	Not in database
Roberts et al. (2015)	Yes	Not in database	Yes	Yes
Beňuš (2009)	Yes	Not in database	Not in database	Not in database
Yoshida et al. (2008)	Yes	Not in database	Not in database	Not in database
Pouw & Holler (2022)	Yes	Not in database	Yes	Yes
Templeton et al. (2022)	Yes	Not in database	Yes	Yes
Campbell (2008)	Yes	Not in database	Not in database	Not in database

Table 3.1 Records used to develop the search strategy.

3.2.2 Inclusion and exclusion criteria

Reports identified through the search strategy described above were included or excluded based on the following criteria. First, as mentioned, only reports written in English were included. No geographical or time limit was applied. Journal articles, books, theses and dissertations and technical reports indexed by the databases were considered for inclusion. Other sources of grey literature, i.e., sources of literature published outside the traditional academic outlets (Boland et al., 2017) were not searched.

Reports were included if they reported duration measurements of turn timing from spontaneous conversations between adults. The way duration measurements should be reported was not specified in the protocol, and we chose to consider any descriptive statistics (including correlation coefficients and quantitative characterisations of distributions), inferential statistics and visualisations of measurements as reports of measurements. However, for the purpose of the analysis presented in this chapter, only

the selection of reports that provided measures of central tendency and visualisation of distributions was included.

Database	Interface	Search Query	Nr. records
Scopus	Scopus	(TITLE-ABS-KEY (conversation* OR turn-taking OR dialogue OR "verbal interaction" OR "spoken interaction" OR "dyadic interaction" OR "spontaneous speech" OR "speech activity") AND TITLE-ABS-KEY ("turn tim*" OR "floor transfer offset" OR (respon* W/1 (latenc* OR delay OR tim* OR offset)) OR (gap W/1 length) OR (gap W/1 duration) OR (gap W/1 turn) OR (overlap W/1 length) OR (overlap W/1 duration) OR (overlap W/1 turn) OR (pause W/1 length) OR (pause W/1 duration) OR (pause W/1 turn) OR (timing W/2 turn-taking) OR (temporal W/2 turn-taking) OR (rhythm* W/2 turn-taking) OR (timing W/2 discourse) OR (temporal W/2 discourse) OR (rhythm* W/2 discourse) OR "conversational timing") AND NOT TITLE-ABS-KEY (infant OR child OR animal OR species OR robot* OR artificial OR agent OR chatbot OR "voice assistant")) AND (LIMIT-TO (LANGUAGE , "English"))	708
Language and Behavioural Abstracts	ProQuest	(Subject,summmary,title("Conversation" OR "Turn Taking" OR dialogue OR "Dyadic Interaction" OR "verbal interaction" OR "spoken interaction" OR "Spontaneous Speech" OR "speech activity") AND subject,summmary,title("turn time" OR "floor transfer offset" OR (respon* NEAR/2 (delay OR latenc* OR tim* OR offset)) OR ((gap OR pause OR overlap) NEAR/2 (duration OR length OR turn)) OR ((timing OR temporal OR rhythm*) NEAR/3 (turn-taking OR discourse)) OR "conversational timing") NOT subject,summmary,title(infant OR child OR animal OR species OR robot* OR artificial OR agent OR chatbot OR "voice assistant")) AND la.exact("ENG")	217

Table 3.2 (Part 1 of 2) Search strategy. For each database, this table lists the interface used to access it, the search queries used to search it and number of results for each search on 31-03-2023.

psycINFO	Ovid	1. Conversation/ or "Interaction (Social)"/ or Social Interaction/ or "Interaction (Interpersonal)"/ or (conversation* or turn-taking or dialogue* or spoken interaction* or dyadic interaction* or verbal interaction* or spontaneous speech or speech activity).mp. 2. (turn tim* or "floor transfer offset*" or (respon* adj2 (tim* or latenc* or delay* or offset*)) or ((gap* or pause* or overlap*) adj2 (duration* or length* or turn*)) or ((timing or temporal or rhythm*) adj3 (turn taking or discourse)) or "conversational timing").mp 3. (infant* or child* or animal* or species or robot* or artificial or agent* OR chatbot* OR voice assistant*).mp. 4. (1 and 2) not 3 5. limit 4 to english language	351
MEDLINE	Ovid	1. (conversation* or turn-taking or dialogue* or spoken interaction* or dyadic interaction* or verbal interaction* or spontaneous speech or speech activity).mp. 2. (turn tim* or "floor transfer offset*" or (respon* adj2 (tim* or latenc* or delay* or offset*)) or ((gap* or pause* or overlap*) adj2 (duration* or length* or turn*)) or ((timing or temporal or rhythm*) adj3 (turn taking or discourse)) or "conversational timing").mp 3. (infant* or child* or animal* or species or robot* or artificial or agent* OR chatbot* OR voice assistant*).mp. 4. (1 and 2) not 3 5. limit 4 to english language	192

Table 3.2 (Part 2 of 2) Search strategy. For each database, this table lists the interface used to access it, the search queries used to search it and number of results for each search on 31-03-2023.

Spontaneous conversations were described in the protocol as unscripted, live, linguistic (oral or signed but not written) interactions, including semi-constrained conversational tasks such as a map task (e.g., Anderson et al., 1991), in which the participants choose when and how to initiate turns. A conversation in which participants are given a set of discussion points or questions but are free to choose when and/or if to raise them, is considered spontaneous. On the other hand, when the content of turns is not chosen by the speakers or scripted in advance, the conversation is not considered spontaneous. Examples of such cases are turns by a simultaneous interpreter during an otherwise

spontaneous conversation (e.g., Vranjes & Oben, 2022), or conversations by speakers who were given the opportunity to practice the content of the whole conversation together before the recording (e.g., Makarova & Zhou, 2006). We included reports that present data from conversations between a study participant and a confederate who was instructed what questions to ask, but consider only the participants' turn timing (Bögels, 2020; Brosy et al., 2016).

Reports were excluded when they concerned only experiments that elicited conversation-like responses, if the participant was or could reasonably have been aware they are not engaged in a spontaneous conversation with a human, or when the conversations involved one or more non-human conversation participants (e.g., animals or robots). Additionally, because this systematic review is concerned with adult behaviour, not language development, studies that involve participants under 16 years old were excluded. This age limit was chosen because this review was conducted at a UK institution, where the lowest legal age for study participants to provide informed consent to participate is 16 (in Scotland; Lepola et al., 2016). For the present purpose, speakers who are 16 years old or older were assumed to be at an adult stage of their language development.

Reports were not excluded for involving participants with language-developmental disorders or psychiatric conditions, even though some studies show differences in turn timing behaviour between patient groups and control groups (Alpert et al., 1997; Howes & Lavelle, 2023). Similarly, we did not exclude reports that presented data from sign languages, even though there are possible differences between processing of sign language and vocal language, including the time scale of stages of processing (Caldwell, 2022), and potentially in the timing organisation of sign languages and vocal languages (Grosjean, 1979). The reason for not excluding such reports is that the present study aims to be agnostic regarding the mechanism that is responsible for turn timing production and does not assume that peculiarities of sign language or of non-neurotypical language use interact with that mechanism differently. Turn timing from these populations is informative about the scope of variation in turn timing behaviour in general, and as the planned synthesis of the data is qualitative, the extent of any differences between these populations will be visible to the reader.

3.2.3 Screening process

The search strategy described in Section 3.2.1 was performed on the 31st of March 2023, when it identified a total of 1468 records across the four databases. The number of records identified in each database is listed in Table 3.2 above. This section describes the selection procedure that was then performed to select eligible reports. It consisted of three steps: records were screened first by title, then by abstract, and then by their full text. All records were reviewed by the first reviewer (the author of this thesis) and a random sample of records was additionally reviewed at step 2 by a second reviewer (Evelyn Fischer; see below). Each step, along with the number of reports included and reasons for exclusion, is also illustrated in the PRISMA diagram (Page et al., 2021) in Figure 3.1 (next page). Rayyan (Ouzzani et al., 2016), a web application designed to support the reviewing process, was used to manage these steps. Of the several AI tools featured by Rayyan, only the duplicate detection tool was used.

First, duplicate records were identified using Rayyan's duplicate detection tool. 478 records were flagged as duplicates, one of which was automatically removed by the tool and verified by the first reviewer, and the remaining 477 were manually removed by the first reviewer, leaving 990 unique records.

Selection step 1 Next, the first reviewer screened the unique records by title. The author(s) and journal name were hidden from the reviewer at this step. 255 reports were excluded based on their title, for being: 1) not on conversation, 2) not on humans, 3) not on timing, or 4) qualitative. Examples of reports that were excluded based on their title are Xue et al.'s (2021) 'Design and synthesis of 2D rGO/NiO heterostructure composites for high-performance electrochromic energy storage', a study on material science, and Franco-Gaviria et al.'s (2022) 'Quantifying resilience of socio-ecological systems through dynamic Bayesian networks', a study on the management of natural resources. These two records, along with the other 253 records excluded at this point, were considered unlikely to report conversational turn timing measurements.

Selection step 2 Next, the first reviewer screened the remaining 735 reports based on their abstracts. To assess the risk of bias in the selection process, a second reviewer also reviewed a randomly selected 20% sample of the 735 reports in step 2, amounting to 147 reports. Of those 147 reports, 20 (13.6%) received different decisions from the two reviewers: eleven were excluded by the first reviewer but included by the second

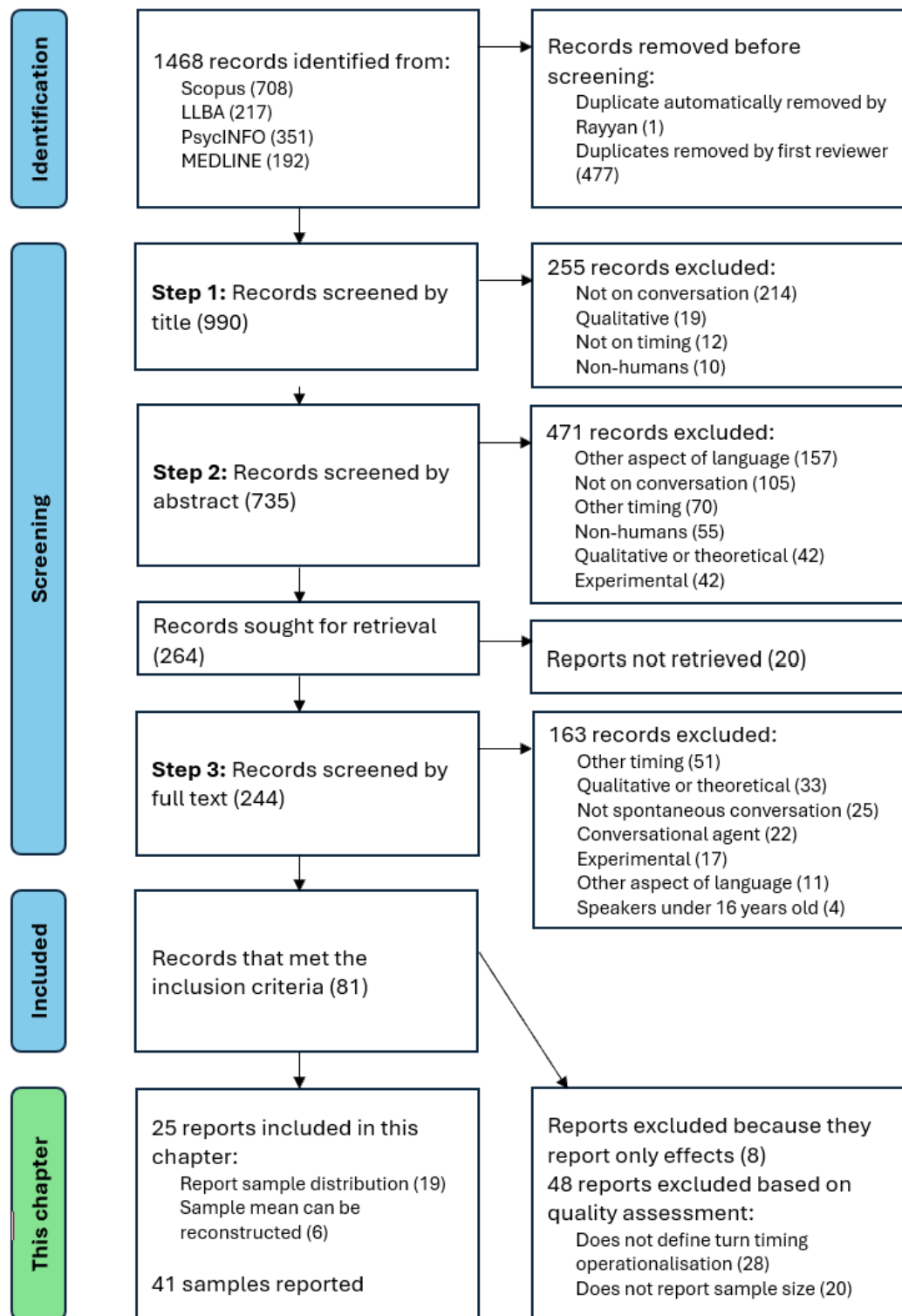


Figure 3.1 Adjusted PRISMA flow diagram (Page et al., 2021) indicating the number of studies included and excluded at each stage.

reviewer, while nine were excluded by the second reviewer but included by the first reviewer. Disagreement between the two reviewers was resolved through discussion. The two reviewers reached a consensus on all cases of disagreement. The consensus decision for three reports was to exclude them, and for 17 reports to include them. Only

one of these was included in the final selection. The disagreement resolution process was not taken into consideration by the first reviewer while reviewing the remaining 80% of records at this step. These were reviewed before the two reviewers compared their decisions – the second reviewer’s role was to provide a reliability check for the review process.

In total, 471 reports were excluded after all disagreements were resolved, leaving 264. At this step, reports were excluded because their abstracts indicated that they were 1) not on conversation, 2) not on humans, 3) theoretical work only, 4) qualitative, 5) experimental, 6) on another aspect of language, or 7) on the timing of another kind of behaviour. Examples of reports that were excluded based on abstracts are a journal article by Basset et al. (2001), which states in its abstract that it is concerned with movement of the velum during spontaneous speech, and a journal article by Nevat et al. (2014), which states in its abstract that it is concerned with written words. The abstracts of both examples are cited and discussed in Appendix A2.

Selection step 3 The full text of the resulting 264 reports was sought for retrieval. We attempted to access their full text through Newcastle University’s institutional subscriptions and collections and online pre-print and PDF repositories such as arXiv (Ginsparg, 2023). 20 reports could not be found, leaving 244.

The first reviewer screened the remaining 244 reports based on the full text, and excluded a further 163, leaving 81 included reports. A report was excluded at this step for the following reasons.

- 1) It reports measurements of a different kind of timing behaviour, including other types of conversational timing (e.g., on the duration of pauses after a turn-initial filler word; Beňuš, 2009b).
- 2) It reports on another aspect of speech (e.g., on polyrhythmicity; Andrews, 2021).
- 3) It uses only qualitative methods, including Conversation Analysis. Conversation-Analytic studies were excluded, even though they often present transcripts that include TTT measurements, because the fragments of their total dataset are usually specifically chosen to exemplify a pattern, rather than to represent that pattern quantitatively (e.g., Nevile, 2007).

- 4) It reports a study in which speakers under 16 years old are included in the conversations (e.g., Peltonen, 2022).
- 5) It reports the behaviour or construction of artificial conversational agents. This category included studies that build an artificial conversational agent and report its turn-timing behaviour. Others used human turn timing data to train machine learning models, but did not provide any TTT values, did not describe the direct relationship of that turn timing to any other variable, and did not use interpretability methods to describe how the turn-timing feature affected their model (e.g., Sakuma et al., 2023) – consequently they included no measurements of human turn timing.
- 6) Its data were not from spontaneous conversations. This category includes studies on written communication and rehearsed dialogues (e.g., Makarova & Zhou, 2006).
- 7) Its task setting included experimental constraints, for example by prescribing the exact content of turns (e.g., A. S. Meyer et al., 2018).
- 8) It presents exclusively theoretical work (e.g., Sandamirskaya & Schöner, 2008).
- 9) It reports turn-timing measurements in such a way that they cannot be distinguished from other phenomena. For example, in some studies, turn timing was measured as the time between a question and the relevant answer, even if that time included several turn switches with clarification requests (e.g., Bóna & Bakti, 2020; Timbrook et al., 2022). An exhaustive list of reasons to categorise reports in this group as excluded for reason (9) is given in Table 3.3 (next page), along with numbers of reports in each category.

3.2.4 Quality assessment

The quality of the reports identified through the selection procedure, as it relates to the purpose of this study, was then assessed using the three criteria stated in the preregistered protocol. General concerns about the consistency of each report were also highlighted and will be discussed in Section 3.2.6 on the data extraction process.

First, we assessed whether the report included a description of their operationalisation of turn timing. This criterion was included because without this information, it is not always possible to establish whether the report fell in exclusion category 9) listed in the

Issue	Number of reports excluded for this reason
Turn timing measurements can include multiple turns.	3
Turn timing is reported as the percentage of time in the conversation that is spend in turn-transition time, without including the number of measurements.	3
Turn timing is measured as annotators' impressionistic ratings of delay without giving duration measurements.	2
Within- and between speaker pauses are not distinguished.	3

Table 3.3 Reasons for excluding reports based on their operationalisation of turn timing, along with counts.

previous section. Of the selection of 81 reports, 28 did not indicate how turn timing was operationalised.

Second, we assessed whether the report included a description of the type of conversation or conversational task. This criterion was included to identify any studies that refer to an experimental or non-linguistic task as ‘conversational’. All 81 reports met this criterion, although some specified more information than others. For example, not all reports clearly indicated whether the conversations were dyadic or included multiple speakers. However, they all provided enough information to establish that they were involved in spontaneous conversations as defined in Section 3.2.2.

Third, we assessed whether the report indicated the sample size of their analysis. This criterion was included so the effect of sample size on the distributions could be assessed. Of the selection of 81 reports, 20 reports did not indicate any sample size.

3.2.5 Further exclusion steps for the purpose of this chapter

The analysis in this chapter is concerned with the scope of variation of turn timing distributions and what characteristics of the studies might explain this variation. For this reason, the selection of 81 reports that met the systematic review’s inclusion criteria was further narrowed down using the following additional criteria.

First, to ensure the comparability of the data, we took the most conservative approach to the three main quality criteria discussed in Section 3.2.4: for the purpose of this chapter,

reports that do not meet the quality criteria (missing operationalisation of turn timing, missing description of the conversation, and missing sample size) were excluded from the analysis. 50 reports were excluded for this reason.

Second, of the 33 reports that met the inclusion criteria and passed the quality assessment, some include data that are suitable for the first aim of the systematic review, because they report turn timing distributions, and some for the second aim, because they report effects. Only reports that presented the overall turn timing distribution of a study were included. Reports that only present effects on turn timing without summary statistics, or insufficient information to reconstruct overall summary statistics (see Section 3.2.6), were excluded.

3.2.6 Data extraction process

In total, the analysis in this chapter includes 25 reports, which describe the distribution of 41 turn timing samples. Data were extracted manually from each report by the first reviewer. The authors of reports were not contacted to ask for raw data or for summary statistics that were not included in the report. No automated tools for data extraction were used. When the report included supplementary materials with relevant summary statistics, these were also extracted, but when raw or pre-processed data were available, these were not analysed.

Data were extracted from each reported sample of turn timing observations. Whenever a report or study included datasets with conversational data from different languages and by different speakers, these are referred to as separate corpora and treated as separate samples in this chapter. For example, Heldner and Edlund (2010) used three publicly available corpora to answer one research question, and Stivers et al. (2009) used ten datasets in ten different languages by different speakers. These are treated respectively as three and ten different samples used in two studies. However, Sørensen et al. (2021) used a dataset of conversations by first-language speakers of Danish in Danish and English. This is treated as one sample because they involved the same set of speakers. The following data were extracted for each sample:

First, the following summary statistics of the sample distribution were extracted when they were available: the mean, median, mode, and measures of spread, skewness (the degree to which the distribution is asymmetrical toward the positive or negative side) and

kurtosis (the degree to which the distribution has a sharper or flatter peak, and sometimes thicker or narrower tails). Where needed, the measurements were converted to milliseconds⁵. The method of measuring the mode was also recorded when it was reported, but only two reports included this information.

Visualisations, including histograms and density plots, were also extracted. In seven reports, a histogram or density plot was included in the report, but the mode was not. In those cases, the mode was estimated from the plot using WebPlotDigitizer, a webtool designed to reconstruct data from plots (Rohatgi, 2017). Skew and kurtosis were not calculated if they were not available, because those measures require the raw data, but a heuristic was used to assess skew and visualisations of the distributions were assessed for kurtosis qualitatively (see Section 3.2.7).

Whenever a report included the mean and sample size of all subgroups in a sample, but no overall mean, the overall mean was calculated as the weighted average of the group means. However, when a summary statistic did not represent an aspect of individual TTT measurements, for example because it was the unweighted means of two means, or the percentage of the conversation spent in TTT, it was not included. When a report included a sample size for only some of its samples, only the sample of which the sample size was available was included here. Proportional data, such as the percentage of gaps and overlaps in a distribution, were also not extracted. When data were missing, no interpolation or other attempts to reconstruct missing data were made, besides those already described. Secondary data cited in the reports were also not included.

Some limits to the comparability of the extracted numerical data should be noted. First, in total, the method of calculating the mode was only available for two samples. The mode of a continuous distribution is not a determined value (i.e., the most frequent value), but depends on how it is calculated: for example, through finding the point of highest density in a probability density curve calculated in a particular way, or through binning data in bins of a particular size. Particularly the latter method is sensitive to the bin size. As a result, the modes are comparable to a limited and unknown degree.

⁵ One report (Mertens & de Ruiter, 2021) only included the unit in which turn timing was reported for some summary statistics, but not all. In that case, we cross-referenced figures reported in different sections of the report and its supplementary materials and assumed they reported seconds when no unit was specified.

Additionally, one study (Stivers et al., 2009) reports medians that are all multiples of 100ms TTT. These are likely values derived from binned data rather than raw medians. They also report modes that appear to be calculated with different methods: the modes in table S3 of their supplementary materials are multiples of 100ms TTT (0, 100 or 200), while the peaks in their histograms appear to lie in the bin from -250ms to +250ms TTT, and in the heading of the figure that presents those histograms, they state: “All distributions are unimodal with the highest number of transitions occurring between 0 and 200 ms.” (Stivers et al., 2009, p. 10589). We used the values reported in table S3 in the supplementary materials.

In addition to summary statistics and visualisations, the narrative descriptions of the conversation characteristics, the types of turn switches included in the sample, and the operationalisation of turn timing were also extracted. Section 3.2.7 describes how they were coded into categorical variables. The corpus and language of each sample were also included as variables.

Finally, as part of the systematic review, the description of participant characteristics was also extracted, as well as the total number of speakers, the exact number of speakers per conversation, the number of conversations, any statistical techniques used, and the factors reported to affect turn timing. These data were not used in this chapter.

The categories of extracted data, along with descriptions of their type, are listed in Table 3.4 (next page; note that the names of some data types, and their groupings, were changed in table 3.4 compared to the protocol for easier presentation). Some data were transformed or coded (see Section 3.2.7 and 3.3). Data were processed using Excel (*Microsoft Excel*, 2018).

3.2.7 Synthesis methods

The data were synthesised using the following approaches. First, to assess the variation in turn timing distributions across studies (research question 1.1), the variations in the means, medians, modes, standard deviations and sample sizes of the 41 samples were visualised and summarised by calculating the median, range and interquartile range of those values across the samples. For example, we identified the median value, the range and the interquartile range of the means of all samples. The visualisations extracted from

Variable	Type	Description
Conversation type	Conversation	The task conditions of the conversation (levels: free, interview, task-oriented).
Eye-contact	Conversation	Whether the speakers could make eye-contact during the conversation (levels: yes, no, both, unknown).
Dyadic	Conversation	Whether the conversations were dyadic or involved two or more speakers (levels: yes, no, both, unknown).
Naturalistic	Conversation	Whether the conversation was elicited for the purpose of the study, such as free lab-based conversations, or whether it had a purpose that was part of the participant's life and was recorded in the environment that suited that purpose, such as police interviews or BBC interviews (levels: yes, no, unknown).
Familiarity	Conversation [excluded]	Whether the speakers knew each other before the time of recording or not (levels: yes, no, unknown).
Dialogue Acts	Methodological	The timing of which dialogue acts was included (levels: all, no backchannels, only question-response sequences, only repair sequences, only echoic responses).
TTT Values	Methodological	Whether all TTT values were measured, or if a cut-off point was used (levels: all, gaps, overlaps, ≥ 750 , > 120 , > 200). Note that if outliers were removed, this information is not captured by this variable.
Turn Boundaries	Methodological [excluded]	What types of signals were considered part of turns, and therefore not of the TTT measurement. (Levels: not filled pauses, all vocal or manual gestures, only speech components, not laughter, unknown.)
Language	Language	What language the participants were speaking.

Table 3.4 Categories of extracted data.

the reports were assessed qualitatively for skewness and kurtosis. As an additional heuristic of skew, the medians and modes of each distribution were compared to the mean, were distributions with a higher mean than median or mode were considered right-skewed, and those with a lower mean as left-skewed.

Second, we conducted a narrative synthesis to identify the similarities and differences between the turn timing operationalisations used in the 25 reports (research question 1.2).

Data type	Explanation
report id	APA style reference to report.
sample id	Reference to report followed by a keyword that identifies the sample, if the report included more than one sample.
corpus	If publicly available, the same conversations may be analysed by different authors, so corpus name is recorded.
language	Language(s) used in the conversations.
participant characteristics	Narrative description from the report of characteristics of the speakers e.g., their age or any neurological conditions.
conversation characteristics	Narrative description from the report of characteristics of the conversation, such as their type (e.g., task-oriented naturalistic dialogue, map task, free conversation with prompts), situation (e.g., recorded in an indoor public space, lab, etc.), and modality (e.g., whether the speakers could have eye-contact, spoke over the telephone or face-to-face, etc), and whether they were dyadic.
operationalisation of turn timing	Narrative description of the operationalisation of turn timing.
dialogue acts included	Description of what types of turn switches were included, e.g., question-answer sequences, all proper turn switches, or all speech-activity switches.
unit of measurement	Milliseconds, seconds, etc.
number of conversations	Numeric value.
number of participants	Numeric value.
number of participants per conversation	Numeric values.
mean	Numeric value.
mode	Numeric value.
mode (method)	Description of how mode was computed.
SD	Numeric value.
median	Numeric value.
kurtosis	Numeric value.
skewness	Numeric value.
sample size	The number of turn timing measurements included in the sample.
data available	Whether and where raw data is available
graph available	Whether authors have included visualizations of turn timing data.
effects on turn timing	Description of any variable that affected turn timing.
statistical analysis	Statistics used in the report.
author's views	Theoretical perspective the authors ascribe to, if any.

Table 3.5 Variables coded from the narrative description of the conversations and methods in each report.

Third, the extracted narrative descriptions of the conversation characteristics and the operationalisation of turn timing for each sample were coded into categorical variables. These resulting variables are listed in Table 3.5 (previous page) and discussed in Sections 3.3.4 and 3.3.5. Not all reports provided information regarding each variable, and variables that had ‘unknown’ as the majority class were excluded.

Forest plots were then used to visualise potential effects of the variables on measures of central tendency of the samples (research questions 1.3 and 1.4). Only the forest plots that support the discussion in this chapter are presented in the main text, while the forest plots with all other possible groupings are included in Appendix A4 for completeness.

Although the 41 selected samples were not homogeneous in their operationalisation of turn timing or other aspects of their design, they all met the selection criteria of this review. Additionally, several samples came from the same corpus, so an analysis of variation across the entire selection risks bias introduced by repeated observations of the same data. For these reasons, we assessed forest plots and summary statistics for two subsets of the samples as well as the entire selection of samples. They are listed in Table 3.6 and were defined by the following criteria:

Subset	No. reports	No. samples	Synthesis methods
Subset 1	8	9	• Discussion of potential effects of variables through forest plots.
Subset 2	15	27	• Discussion of potential effects of variables through forest plots.
Entire selection	25	41	• Discussion of potential effects of variables through forest plots. • Discussion of distributions through visualisations. • Discussion of distribution skew using mean and median/mode comparison as a heuristic. • Discussion of the effect of sample size on measures of central tendency and spread through scatter plots.

Table 3.6 Type and number of reports and samples included in each synthesis subset, and the synthesis methods used for each subset.

Subset 1 comprises the most homogeneous subset of samples, namely only those that include measurements of turn timing for all dialogue act types in the conversation (e.g., not only question-response sequences), and all TTT values (or applied a symmetrical lower-and upper limit). When multiple studies used the same corpus to collect a sample that met these criteria, only the earliest published sample is included in subset 1. Only nine samples met these criteria. Subset 2 includes all samples from subset 1, as well as samples that represent a subset of dialogue acts in the conversations (e.g., only question-response sequences) and all TTT values. If multiple studies used the same corpus but made different decisions about which dialogue acts to include, both samples were included in subset 2. Of the 41 selected samples, 27 samples met the criteria to be included in subset 2.

Finally, the degree to which turn timing samples from the same language and from the same corpus group together is also evaluated using forest plots.

Data were transformed and plotted using the statistical program R (R Core Team, 2022) and the packages ggplot2 (Wickham, 2016) and tidyverse (Wickham et al., 2019).

3.3 Results and Discussion

The following sections present and discuss the resulting syntheses. First, Sections 3.3.1 to 3.3.5 describe characteristics of the selected samples, including the report type and publication year (3.3.1), the corpora and sources of data (3.3.2) and the languages represented in the samples (3.3.3). Section 3.3.4 discusses the differences and similarities between turn timing operationalisations, and Section 3.3.5 presents the distribution of conversation-related variables across the samples.

Next, Sections 3.3.6 to 3.3.12 discuss variation in the sample distributions, including characteristics of the sample distributions themselves (3.3.6 and 3.3.7) and potential effects of the variables on measures of central tendency of the samples in subset 1, subset 2 and the entire selection (3.3.8 to 3.3.12).

3.3.1 Selection results and report characteristics

25 reports were included in the final selection. 17 are journal articles, one is a PhD thesis, and seven are papers published in conference proceedings. They were published between 1978 and 2023. Figure 3.2 (next page) illustrates the distribution of their publication date over time. Reports on turn timing distributions in this selection became

more frequent in recent decades. The median publication year is 2011, so half of the reports were published in the final quarter of the period. For ease of presentation, the full list of reports is included in Appendix A3. Table A 3 to Table A 8 contain the aim of each report in general, and its results regarding turn timing specifically.

The 25 selected reports included measures of central tendency for 41 turn timing samples - 19 reports described the distribution of a total of 36 turn timing samples directly, and six more means could be reconstructed from reports that presented group means.

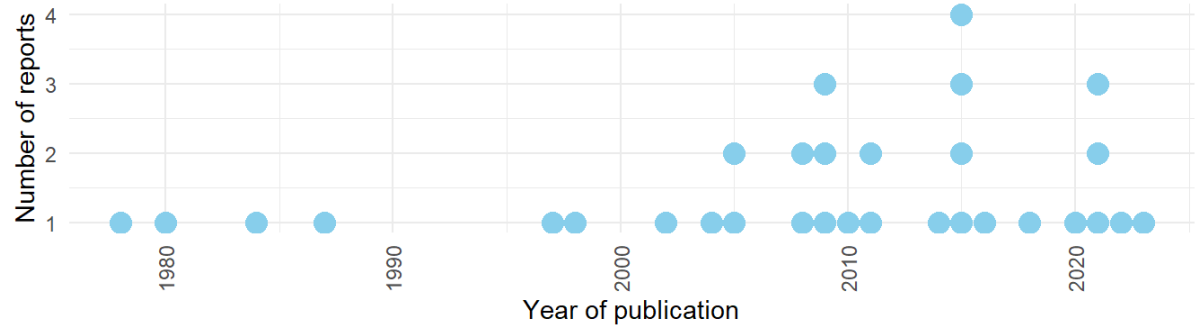
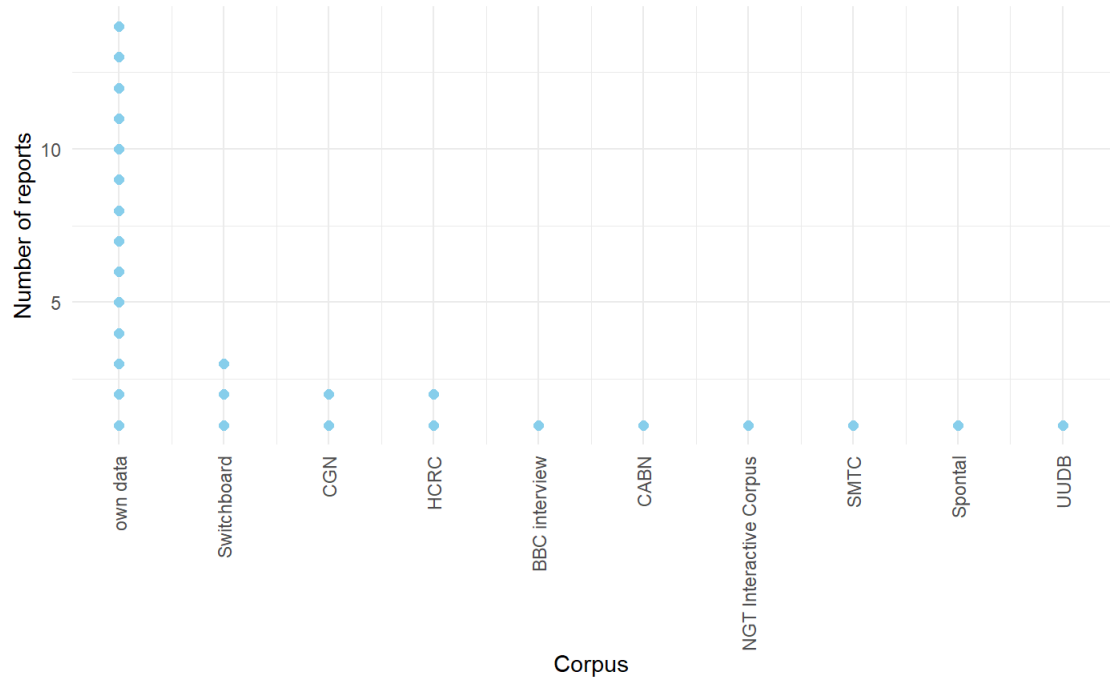


Figure 3.2 Frequency of reports published per year.



3.3.2 Data sources and corpora

14 reports indicate that they used either their own data or data from a specific other study by either the same or different authors, and eleven indicate that they used an existing public corpus or constructed a public corpus. Nine corpora were used in total. Figure 3.3 (previous page) illustrates the frequency of use of each corpus.

Three corpora were used more than once. These were the Switchboard corpus (Godfrey et al., 1992), the HCRC map task corpus (Anderson et al., 1991) and the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN; Oostdijk, 2000), and the NGT Interactive Corpus (*Nederlandse Gebarentaal*, Sign Language of the Netherlands, NGT; De Vos et al., 2015). Section 3.3.12 evaluates the consistency between repeated observations from the same corpus.

3.3.3 Summary of included languages

Overall, conversational data from 15 languages was reported. Figure 3.4 illustrates the number of reports that include data from each language⁶. In five reports, the language in which the conversations were held was not reported. Three of those described the participants as speakers of Japanese, so the conversations were assumed to be in Japanese (Mori, 2009; Nagaoka et al., 2005; Ozaki et al., 2008). One report identified the speakers as celebrities who grew up in the UK (Princess Diana and Martin Bashir; Kowal & O’Connell, 2022) and one presented a study that was conducted while the author was affiliated to an institution in a majority English-speaking country (Beattie, 1978), so the conversations were also assumed to be in English. Given these assumptions, English was the most frequently studied language, followed by Japanese and then Dutch⁷. Additionally, for the 13 reports that included English conversational data, the variety or accent of English was sometimes specified. Figure 3.5 illustrates the distribution of reports on English varieties.

⁶ Note that some reports include data from multiple languages, so the total number of data points in figure 3.4 (37) does not equal the number of reports (25). In five reports, the language in which the conversations were held was not reported. In those cases, we made assumptions about the language, as described in Section 3.3.3.

⁷ Similarly, when considering only the 32 reports in which the language was explicitly stated, English was most frequently studied (11), followed by Dutch (four).

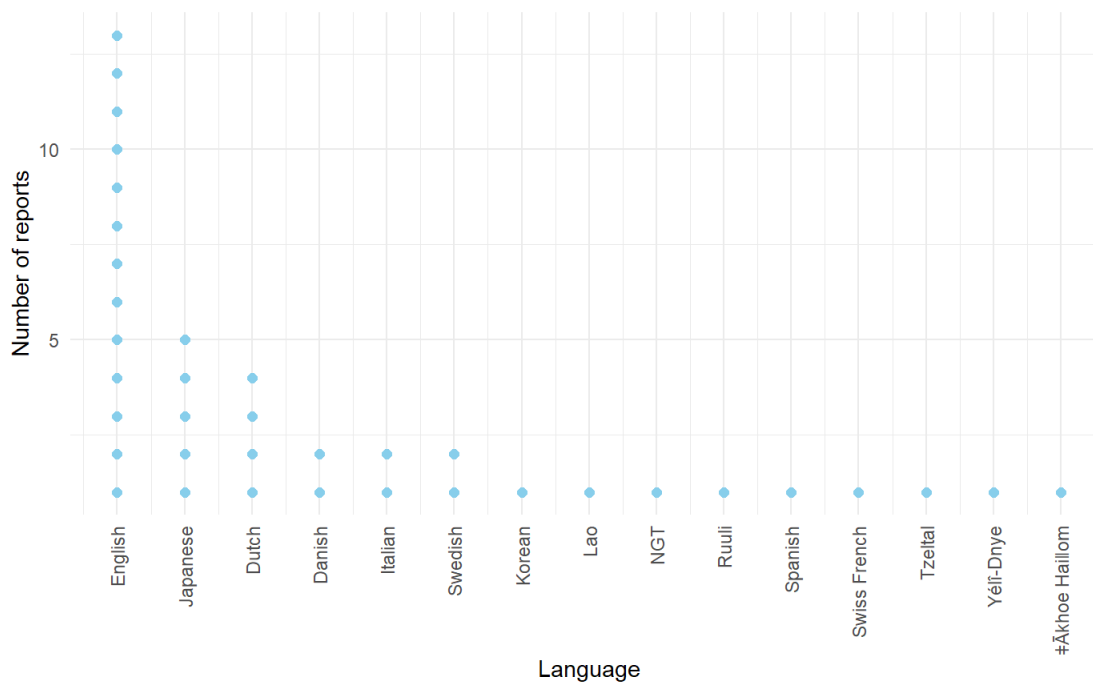


Figure 3.4 Frequency of reports that include conversational data from each language. NGT stands for ‘Nederlandse Gebarentaal’ (Sign Language of the Netherlands).

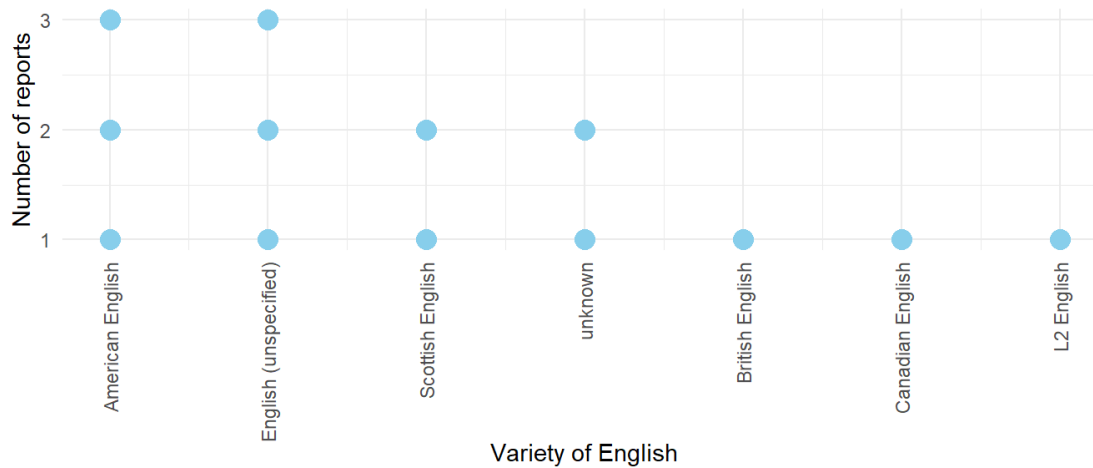


Figure 3.5 Frequency of reports that include conversational data from a particular dialect of English, of the 20 reports that include data in English.

3.3.4 Operationalisations of turn timing and methodological variables

This section discusses similarities and differences between the ways the included reports have operationalised turn timing. It also describes how that information was coded into categorical variables.

Fifteen reports operationalised turn timing as the time between the first and the second turn. Three reports operationalised turn timing as the pause between turns by different speakers, explicitly excluding overlaps. The remaining seven used a conversational state

operationalisation, similar or identical to the one proposed by Heldner and Edlund (2010), in which the conversational state is identified as the presence or absence of speech activity by each speaker, and turn timing is the time it takes to switch from a conversational state where only one speaker speaks to a conversational state in which only another speaker speaks. These definitions essentially amount to the same: they use the end of turn one and the start of turn two as the relevant points in time.

The ends and starts of the turns were identified through a range of methods, from annotator judgements to speech activity detection algorithms. Table 3.7 lists these methods and the number of reports that used them.

Method of identifying turn boundaries	Used by no. of reports
Completeness judgements by annotators	3
Auditory and/or visual inspection	7
dB threshold	1
Speech recognition algorithm	5
Existing boundaries in corpus	4
Not specified	5

Table 3.7 Methods by which the boundaries of turns were identified, with the number of reports that used each method.

The reports also appeared to differ in what component of behaviour or the speech signal they considered part of a turn, although 19 of the 25 did not explicitly specify their decision at all. Five of those used a speech or silence detection algorithm to identify turn boundaries, but the method by which the algorithm identified turn boundaries was not made explicit.

Of the reports that did explicitly describe how they identified turn boundaries, only three specified aspects of how they identified the start of the second turn but not the end of the first turn: Brosy et al. (2016) reported two operationalisations of turn timing, one that included all turn-initial filled pauses in the second turn, and one that excluded them. In the second, all turn-initial filled pauses were part of the TTT. Similarly, Fadden (2008) excluded turn-initial filled pauses from the second turn, so all turn-initial filled pauses

were part of the TTT. Finally, Stivers et al. (2009) considered the onset of a vocal or gestural response as the start of the second turn.

Of the three reports that did not limit their criteria to the start of the second turn, Buanzur et al. (2018) state that they included particles, inbreaths and clicks as part of turns. Nagaoka et al. (2005) used a 60dB threshold for detecting the presence of speech, and excluded laughter from the start of the second turn. More explicitly, De Vos et al. (2015) compared three operationalisations of turn timing in conversations held in Sign Language of the Netherlands, and identified the ends and starts of turns by the phonological status of gestures in the initial and final strokes.

Altogether, 22 reports do not explain how they identified the end of the first turn, although the end of a turn is not a self-evident point in time. It could refer to the end of a functionally defined turn-constructual unit, potentially followed by more speech material by the same speaker but judged by an annotator not to be part of the turn that the second speaker is responding to. An example of such a turn end is the start of the blue box in example 1 from Section 2.1, reproduced below. It could also refer to the offset of any stretch of speech activity. Even in the latter case, the exact point of speech offset cannot always be clearly identified, since the transition from speech signal to non-speech is often not a precisely identifiable moment, especially when two turns overlap.

- 1) Q: Are you going out tonight? Did you see your parents yesterday by the way?
A: Yes, I'm going to see a movie.
[schematic example]

In short, many reports did not provide precise phonetic operationalisations of turn timing. This is not necessarily a problem given the aim of the original study: for example, a study that aims to develop automatic processing tools to detect turn endings during human-robot interactions may want to tolerate imprecision so that their results are more likely to be robust to noise in actual applications (e.g., Chowdhury et al., 2014). However, it does mean that the effect of turn timing operationalisations on the sample estimates we collected can only be assessed to a limited extent – although we initially coded the operationalisations of turn timing into a categorical variable ('Turn Boundaries') to group samples by what features were considered part of the turn, we excluded this variable from further analyses because its majority class was 'unknown'.

For this reason, the decision to exclude reports that did not provide operational definitions of turn timing at all, but include those that provided a phonetically imprecise definition, results in a somewhat arbitrary selection, as many otherwise eligible reports were excluded.

However, not excluding reports that do not provide operational definitions of turn timing would introduce the following problem: as mentioned in section 3.2.3, some excluded reports concern turn timing, but mention including some within-speaker pauses in their measurements (e.g., Bóna & Bakti, 2020). Another used a definition that allows multiple turn switches before the relevant answer to be included in the TTT duration (Timbrook et al., 2022). These could only be identified because they did describe their operationalisation of turn timing. For this reason, we considered excluding reports that did not provide a definition the most principled way to exclude studies that refer to a different phenomenon.

Besides these differences in how turn timing was operationalised, the reports also differed on two other methodological points. The first methodological difference is what dialogue acts they included in their sample. Some measured TTT for all turn switches in their conversational data, while some measured TTT only for a particular type of dialogue act. Most samples only include question-response sequences (19), and nine include all dialogue act types. Other selection choices, namely excluding question-response sequences, excluding backchannels, including only echoic responses (responses that repeat part of the previous turn) and only repair sequences (turn sequences in which one speaker asks or invites the other to clarify what they said; Schegloff et al., 1977), are each represented in only one sample.

The second methodological difference relates to what values of TTT were included in the sample – some included all values, some included only positive values, with a gap between the first and second turn, and others used other cut-off points.

The choice of dialogue acts and TTT values included in the samples were also coded into categorical variables, which were then used to define subset 1 and subset 2 for subgroup analyses. Table 3.8 (next page) lists the frequency of each level.

Methodological variable	Level	Count
Dialogue Acts	All	18
	Question-response sequences	18
	No backchannels	2
	Echoic responses	1
	No question-response sequences	1
	Repair sequences	1
Turn Boundaries	Unknown	23
	All speech and gestures	11
	Speech components	3
	Not filled pauses	2
	All gestures	1
	Not laughter	1
TTT Values	All	30
	>0 or ≥ 0	6
	<0 or ≤ 0	2
	>-750	1
	>120	1
	>200	1

Table 3.8 Number of observations of each level for each methodological variable.

3.3.5 Conversation variables

The narrative descriptions of the conversations from each report were coded into five categorical variables, which are listed in Table 3.5 as conversation variables. Figure 3.7 illustrates the relative frequency of each level of each conversation variable. Note that the number of samples per level is not equal; there were 15 samples from dyadic conversations, for example, but only one sample from exclusively triadic conversations. Additionally, not all reports included the relevant information. As the majority class of Familiarity is ‘unknown’, it is excluded from further analyses.

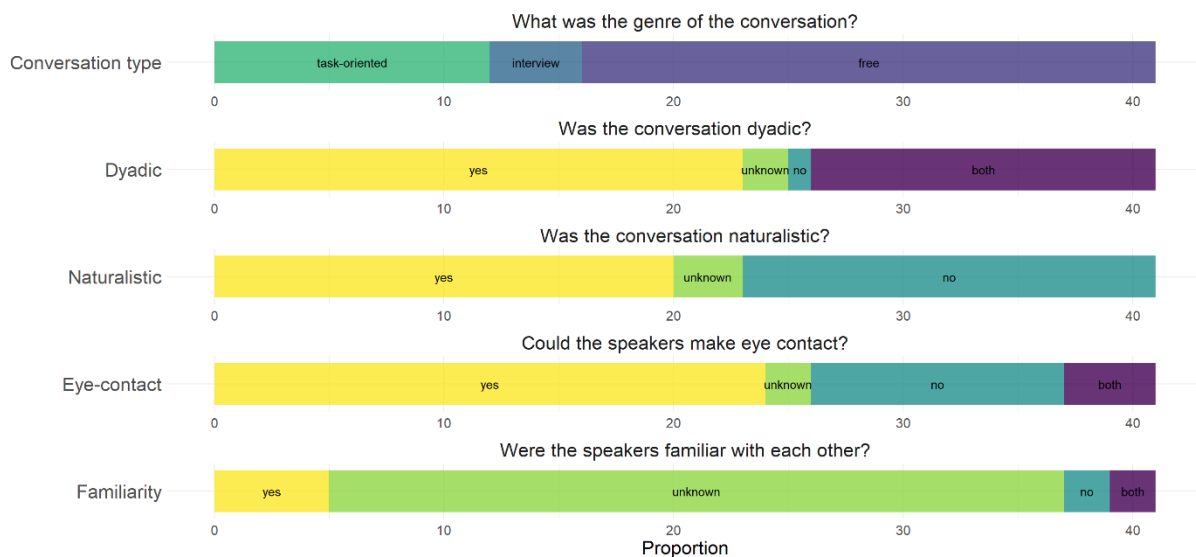


Figure 3.7 The proportion of the 41 samples in each level of each conversation variable.

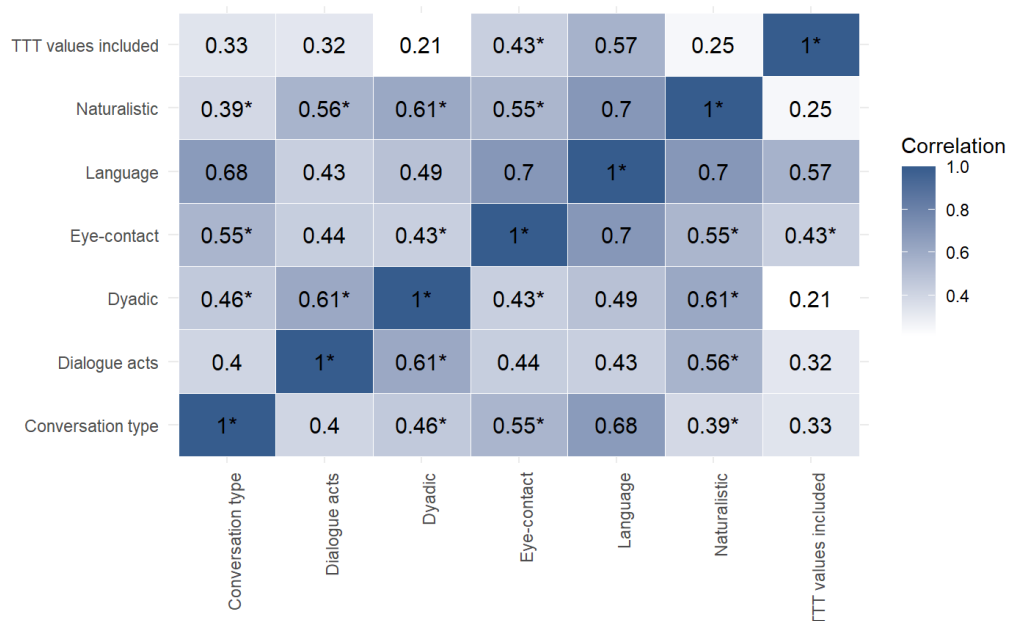


Figure 3.6 Heatmap depicting associations between the conversation- and methodological variables, measured by Cramer's V. Statistically significant correlations at $\alpha=0.05$ are marked with an asterisk (*) (The levels of the variable 'TTT values included' were grouped into the categories: 'gaps', 'overlaps' and 'all').

There were strong associations between the variables, possibly because of the small sample size and unequal spread of variable levels. Figure 3.6 illustrates the strength of associations between each conversation variable, as well as methodological variables (see Section 3.2.6) and Language. The strength of the associations is expressed in Figure 3.6 as Cramer's V, which measures the strength of associations between categorical variables and ranges from 0 (weak association) to 1 (strong association). Due to the strong associations, it is not straightforward to attribute differences between samples to the variables, as Sections 3.3.8 to 3.3.12 discuss in more detail.

3.3.6 Central tendency

Figure 3.8 shows the mean, median and mode for each sample, if it was reported or could be reconstructed from a plot or group means when it was not reported. The samples are ordered by sample size. The labels on the y-axis of Figure 3.8 are the sample IDs. Table 3.9 provides a summary of the means, medians, modes and the standard deviation per sample. It lists the median, range and interquartile range for each summary statistic. The values in Table 3.9 (and similar tables in later sections) are aggregated from the summary statistics of individual samples.

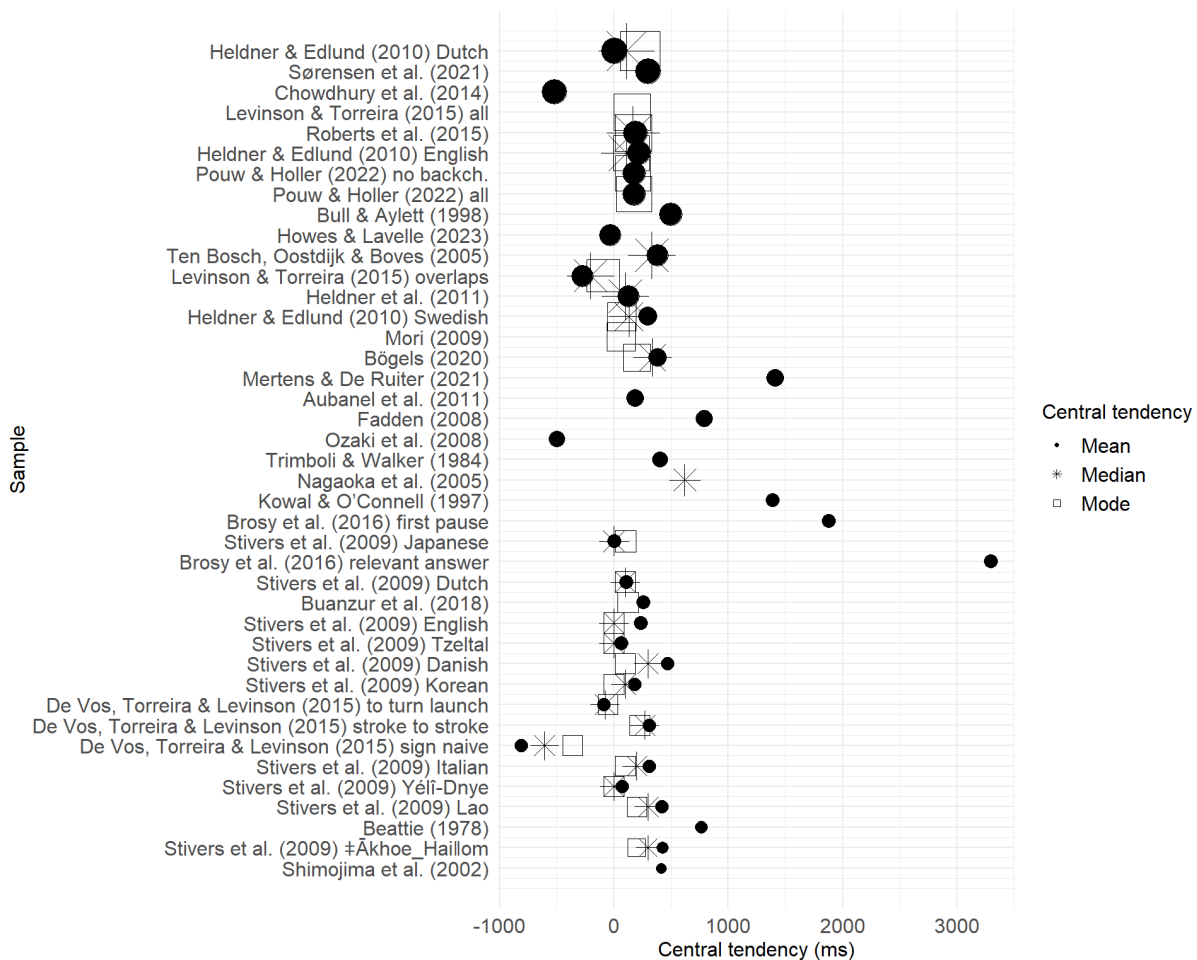


Figure 3.8 Means, medians and modes of all included samples, arranged by sample size. Point size reflects the natural logarithm of the sample size.

As expected, given that the mean is not necessarily considered a stable measure of central tendency for turn timing distributions (see Section 2.2), the means range much more widely (by 4112ms) compared to the medians (1227ms) and the modes (544ms). However, the mean is the most frequently reported measure of central tendency. Figure 3.8 also illustrates that most values lie within the 0 to 500ms range and the wide ranges

are due to a few extreme positive values. This same pattern is reflected in the relatively narrow interquartile ranges as shown in Table 3.9.

Metric	Range	Median	Interquartile range	No. reported
Mean (ms)	-812 to 3300	248	81 to 419	38
Median (ms)	-607 to 620	111	0 to 292	22
Mode (ms)	-316 to 228	100	0 to 171	24
Standard deviation (ms)	310 to 5340	531	458 to 724	22
Sample size	71 to 43374	532	196 to 6790	41

Table 3.9 Summary of the distribution of five summary statistics across samples in the entire selection of samples.

The included samples differ widely in size. The smallest sample has only 71 observations, while the largest has 43374 – more than 610 times as many. Figure 3.9 a-d (next page) show the sample size on the y-axes, against the mean, median, mode and standard deviation on the x-axes. The vertical line indicates the median value for each statistic. Samples with smaller sample sizes are more frequent, since observations cluster towards the origin of the y-axis in Figure 3.9a-d. Figure 3.9a-d also allow us to assess whether measures of central tendency converge with larger sample sizes, which would suggest that the variability of measures of central tendency may be partly due to sampling noise in small samples. Figure 3.9 supports this interpretation to some extent, since the measures of central tendency are more widely distributed towards the bottom of each plot.

A caveat to this interpretation, however, is that smaller samples also often represent a more specific category of turn switches. For example, the smallest sample with 71 observations only included TTT measurements for echoic responses (responses that repeat part of the previous turn, Shimojima et al., 2002). The larger variability in measures of central tendency and spread can therefore also be due to a specific characteristic of the included observations. This would imply that the increased variance in small samples is not just an artifact of the sample size. This possibility is also supported by the trend that samples included in subset 1 (i.e., those that included all values of TTT and all turn switch types; depicted in green in Figure 3.9) are more evenly spread around the vertical lines than the others.

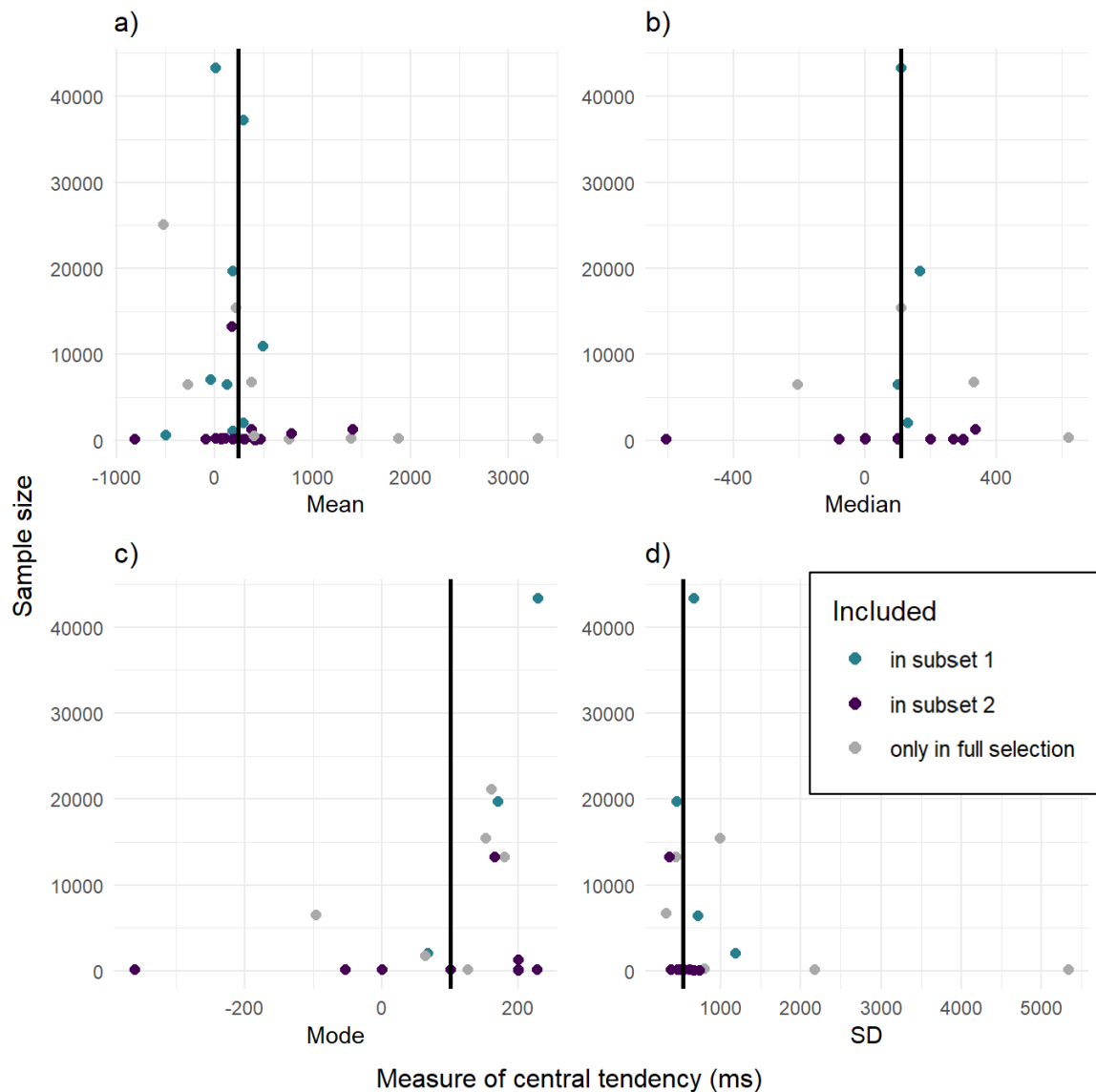


Figure 3.9 Measure of central tendency by sample size for the mean (a), median (b), mode (c) and SD (d) of each sample. The vertical line indicates the median of the means, medians, modes and sample sizes across samples.

Because the included samples are not heterogeneous in this sense, Sections 3.3.8 to 3.3.12 discuss the subsets of the included samples to attempt to find patterns that might be responsible for the diverging values.

3.3.7 Kurtosis and skewness

Only nine reports visualised the distribution of samples as histograms or density plots, which amounted to 22 of the 41 samples. These are reproduced with permission in Figure A 10 to Figure A 30 in Appendix A5. The plots in Figures A 10 to A 30 all depict distributions with a single peak. They also tend to have higher peaks and thicker tails than normal distributions, indicating that they are leptokurtic.

Sørensen et al. (2021) suggested that turn timing distributions are right skewed, based on a literature review of three studies. The plots in Figures A 10 to A 30 do not clearly support their generalisation, although the most frequent group does appear to be right skewed (nine were right skewed compared to six left-skewed and six approximately symmetrical distributions). Interestingly, their skewness in either direction does not necessarily seem to be the result of relatively many positive outliers, but of the relatively high density on either the positive or negative side of the point of highest density. While distributions are thus not straightforwardly bimodal with modes on both sides of 0ms TTT, as Wilson and Wilson (2005) predicted (see Section 2.3), they do show some sign of multiple clusters.

Another heuristic for assessing skewness, besides qualitative inspection of visualisations, compares the relative size of the mean to the median or to the mode of a distribution. Of the samples for which both the mean, and the mode or median, were available, Table 3.10 lists the frequency of right- and left-skewed samples. By this heuristic, most samples were right skewed, although many were close to symmetrical, as illustrated by Figure 3.10.

Metric	No. left skew	No. right skew	No. both reported
Median vs Mean	4	17	21
Mode vs Mean	6	16	22

Table 3.10 Skewness estimated by whether the median/mode is higher than the mean (left skew) or lower than the mean (right skew).

Of the four clearly left-skewed distributions in Figure 3.10, three may diverge from the others because of methodological differences. First, Levinson and Torreira (2015) included only overlaps (i.e., negative turn-transition times) in their sample. Second, De Vos et al. (2015) studied turn timing in Sign Language of the Netherlands (*Nederlandse Gebarentaal*; NGT), using three different operationalisations of turn timing. Of the three resulting samples, only one is right skewed. Incidentally, the authors conclude that their right-skewed sample was measured with an operationalisation that is most comparable to how turn timing is measured in vocal languages, because it did not include preparatory movements in the turns (and, according to authors, because its central tendency is closest to 200ms). In short, these three left-skewed samples could be considered

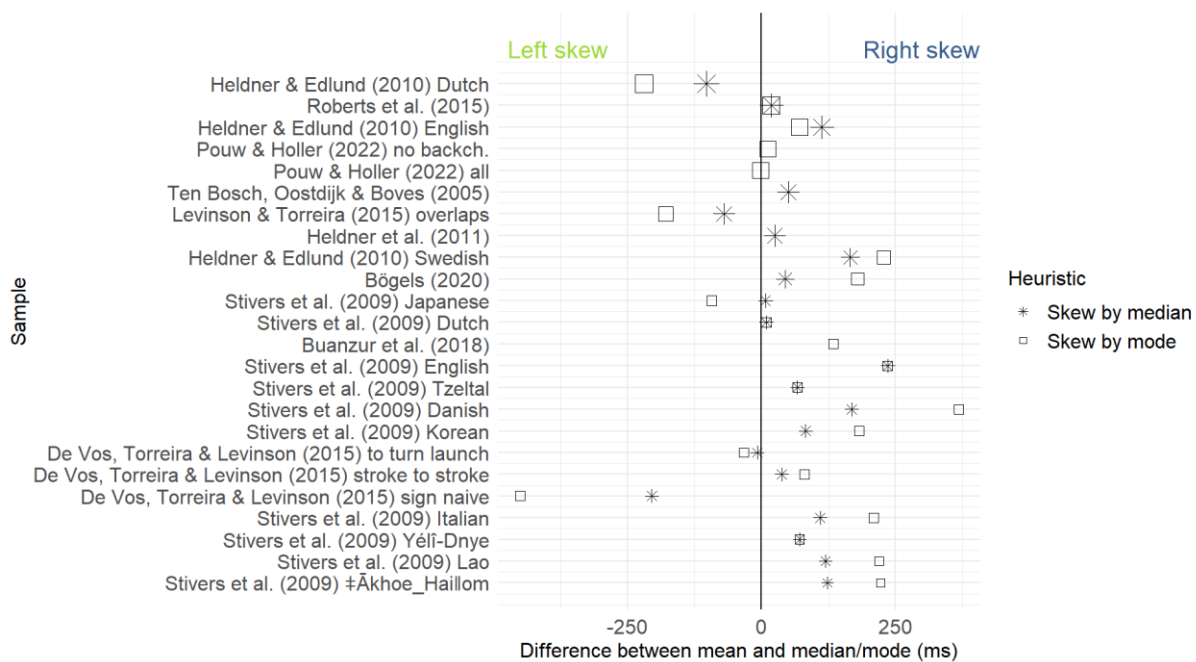


Figure 3.10 The size and relative difference between the mean and median/mode as heuristic for assessing skewness. Point size reflects the logarithm (base 10) of the sample size.

different enough from the others to support the idea that they do not break the generalisation that turn timing distributions tend to be mostly, but not overwhelmingly, right skewed.

This characteristic may be due in part to the fact that turn timing distributions have a soft lower-bound, since the second turn cannot start before the start of the first turn. A mixed-methods analysis that combines qualitative discussion of observations sampled from different regions in the distribution could help clarify this issue.

3.3.8 Subset 1

The following few sections discuss potential effects of the variables on measures of central tendency of the samples in subset 1, subset 2 and the entire selection of samples, starting with subset 1. Subset 1 consists of the samples that include all dialogue act types, all turn timing values of one or more speakers, and only the earliest published analysis of each corpus if two samples from that corpus were otherwise eligible for inclusion in this subset. Only nine distributions met these criteria. Table 3.11 summarises the means, medians, modes and standard deviations of each sample in this subset by the range, median value and interquartile range of those summary statistics. The measures of central tendency of the samples are distributed more narrowly than in the

entire selection, although the means still vary widely, from -495ms to 493ms TTT. The medians and modes are very close to the often-cited 200ms, but only three were reported.

Metric	Range	Median	Interquartile range	No. reported
Mean (ms)	-495 to 493	187	8 to 297	9
Median (ms)	100 to 168	127	108 to 140	4
Mode (ms)	68 to 228	169	118 to 198	3
Standard deviation (ms)	448 to 1182	685	608 to 827	4
Sample size	607 to 43374	7070	2060 to 19754	9

Table 3.11 Summary of the distribution of five summary statistics across samples in subset 1. The range, median, interquartile range and no. reported are listed for the sample means, medians, modes, standard deviations and sample sizes.

Figure 3.4 illustrates the distribution of the measures of central tendency in this sample, colour-coded by Conversation Type. Although this sample is very small and generalisations should be made with caution, the means of free conversations appear to be distributed more narrowly than the means of task-oriented conversations. Indeed, the means of free conversations are all within the 8-190ms range, while the means of task-oriented conversations all fall outside that range.

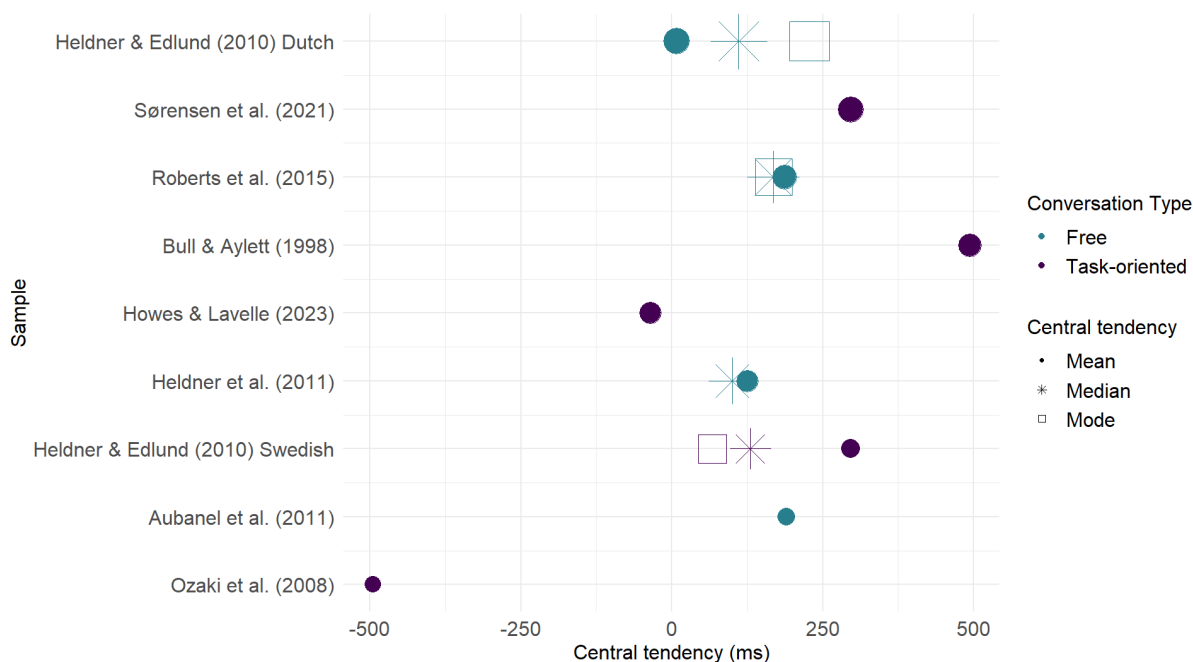


Figure 3.11 The mean, median and mode of samples in subset 1, arranged by sample size. The sample estimates are colour-coded by Conversation Type. Point size reflects the natural logarithm of the sample size.

The three task-oriented dialogues with means *higher* than this range concerned conversational tasks designed to collect spontaneous speech, namely map tasks (Bull & Aylett, 1998; Heldner & Edlund, 2010 Swedish) and the DiapixUK task (Sørensen et al., 2021). The two task-oriented dialogues with means *lower* than this include conversations between a navigator and a driver on the phone while on route (Ozaki et al., 2008), and a debate task, in which triads of participants had to discuss an imaginary scenario and decide which character they would throw out of a hot air balloon (Howes & Lavelle, 2023).

Subset 1 thus tentatively suggests a generalisation that the means of free conversations are more narrowly distributed than the means of task-oriented conversations, but since subset 1 includes only nine samples, this generalisation would need to be supported with more observations. However, if it holds, it could indicate that conversational tasks, compared to free conversations, allow more or larger divergences from the usually expected turn timing. Specifically, the means of the samples appear to range more widely, rather than their medians or modes, so this divergence from an expected turn timing might particularly occur in relatively rare, specific pragmatic contexts. In other words, specific conversational tasks may allow very long gaps or very long overlaps, which would pull the mean of the distribution higher or lower but have little effect on the modes or medians. Alternatively, the two task-oriented conversations with *lower* means both involve potentially stressful situations (debate and driving), which could affect turn timing in general, not just in specific pragmatic contexts. Along this line, language comprehension generally tends to be reduced during driving (Becic et al., 2010).

On the other hand, some other aspects of the specific included samples of task-oriented conversations may explain their wider range as well. Specifically, the conversations by Howes and Lavelle (2023) may have a large negative mean TTT because they are triadic rather than dyadic; Holler et al. (2021) argue that the presence of competition for the conversational floor reduces the observed TTTs because only the first response is measured when two speakers respond. Note also that some of their conversations include speakers with a diagnosis of schizophrenia, since the objective of their study was to compare turn taking behaviour between this group and a control group. However, conversations with only control speakers have a lower mean TTT than conversations that include a speaker with a diagnosis (-60ms vs. -7ms), so the diagnosis of some of the speakers does not seem to relate to the overall low mean.

With regard to the sample from Ozaki et al. (2008), the low mean TTT could also in part be attributable to the relative frequency of overlapping backchannels in Japanese (Cutrone, 2005; Morita, 2001), although it is not clear how the authors treated backchannels. However, their mean is still considerably more negative than other Japanese estimates (see section 3.3.11), making unlikely that the mean observed by Ozaki et al. (2008) is due to the language alone.

3.3.9 Subset 2

Subset 2 consists of all samples that included all values of TTT and any selection of dialogue acts. Repeated measurements from a single corpus were also included if they involve a different selection of dialogue acts. 27 samples met these criteria. Table 3.12 summarises their measures of central tendency and spread. Figure 3.12 (next page) shows the distribution of measures of central tendency in subset 2. Panel a) shows all samples from subset 2, and panel b) excluded four outliers: Ozaki et al. (2008), De Vos, Torreira and Levinson (2015) ‘sign naïve’, Fadden (2008) and Mertens and De Ruiter (2021). Figure 3.12 is colour-coded by Dialogue Acts to visualise the effect of including samples that only included some types of dialogue acts. Of these, only the sample with repair sequences is notably different from the others, with a larger positive mean.

Metric	Range	Median	Interquartile range	No. reported
Mean (ms)	-812 to 1410	190	69 to 398	27
Median (ms)	-607 to 336	106	0 to 251	18
Mode (ms)	-361 to 228	100	0 to 192	18
Standard deviation (ms)	356 to 1182	513	470 to 661	16
Sample size	71 to 43374	247	190 to 4283	27

Table 3.12 Summary of the distribution of five summary statistics across samples in subset 2. The range, median, interquartile range and no. reported are listed for the sample means, medians, modes, standard deviations and sample sizes.

The measures of central tendency of samples of question-response sequences do not appear to be distributed differently from the means of samples that include all turn switches. The four most extreme samples were removed from panel b of Figure 3.12, to more clearly illustrate the similarity between these two groups. The mean of means of



Figure 3.12 The mean, median and mode of distributions in subset 2, arranged by sample size. The sample estimates are colour-coded by Dialogue Act. Point size reflects the natural logarithm of the sample size.

samples with only question-response sequences is 192ms, and the mean of means of samples with all dialogue types is 120ms.

This observation is contrary to some previous studies that did find a difference between question-response sequences (e.g., Beattie, 1978, found that questions were responded to 1117ms slower than non-questions), but is in line with some claims that question-

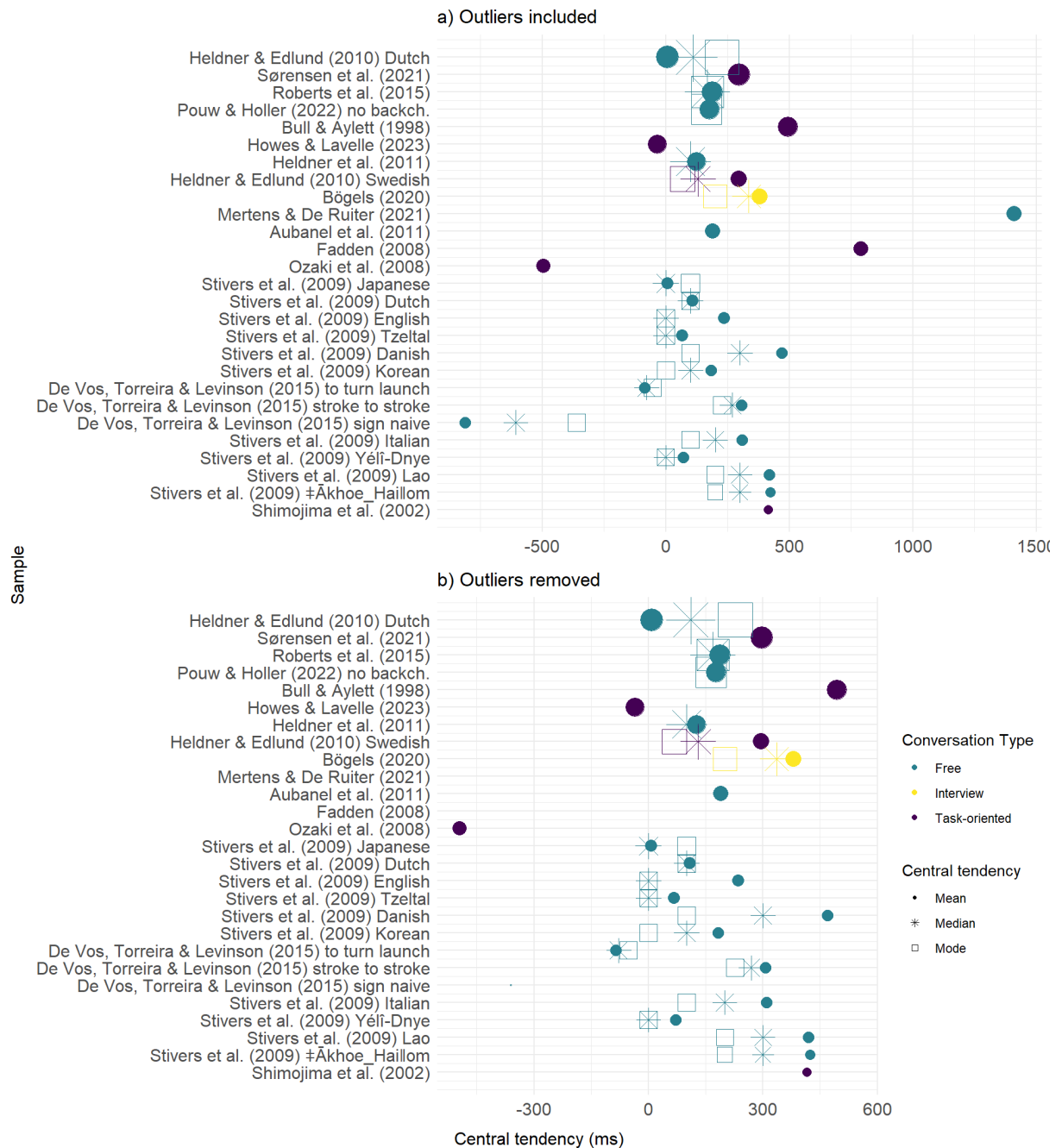


Figure 3.13 The mean, median and mode of distributions in subset 2, arranged by sample size. The sample estimates are colour-coded by Conversation Type. Point size reflects the natural logarithm of the sample size.

response sequences are at least quantitatively representative of all turn timing for some purposes (Stivers et al., 2009), because there is no evidence for a difference.

In subset 1, the means of task-oriented conversations are more widely distributed than the means of free conversations, as illustrated in Figure 3.13. Panel a) shows all samples from subset 2, and panel a) shows all samples from subset 2, and panel b) excluded three outliers: De Vos, Torreira and Levinson (2015) ‘sign naïve’, Fadden (2008) and Mertens

and De Ruiter (2021). Figure 3.13 again shows the means, medians and modes of subset 2, now colour-coded by Conversation Type. The same pattern does not appear in subset 2.

One possible reason why the same pattern may not be replicated in subset 2 is that many of its samples concerned only specific dialogue act types. Thus, if specific pragmatic contexts in task-oriented conversations causes their means to diverge more widely than in free conversations, then restricting the type of dialogue act of a sample may overrepresent or underrepresent unusual contexts. Another possibility is that the samples with a restricted set of dialogue acts had smaller sample sizes, which is associated with more widely varying measures of central tendency (see section 3.3.6). However, the generalisation that the means of task-oriented conversations are more widely distributed is based on only the nine observations in subset 1.

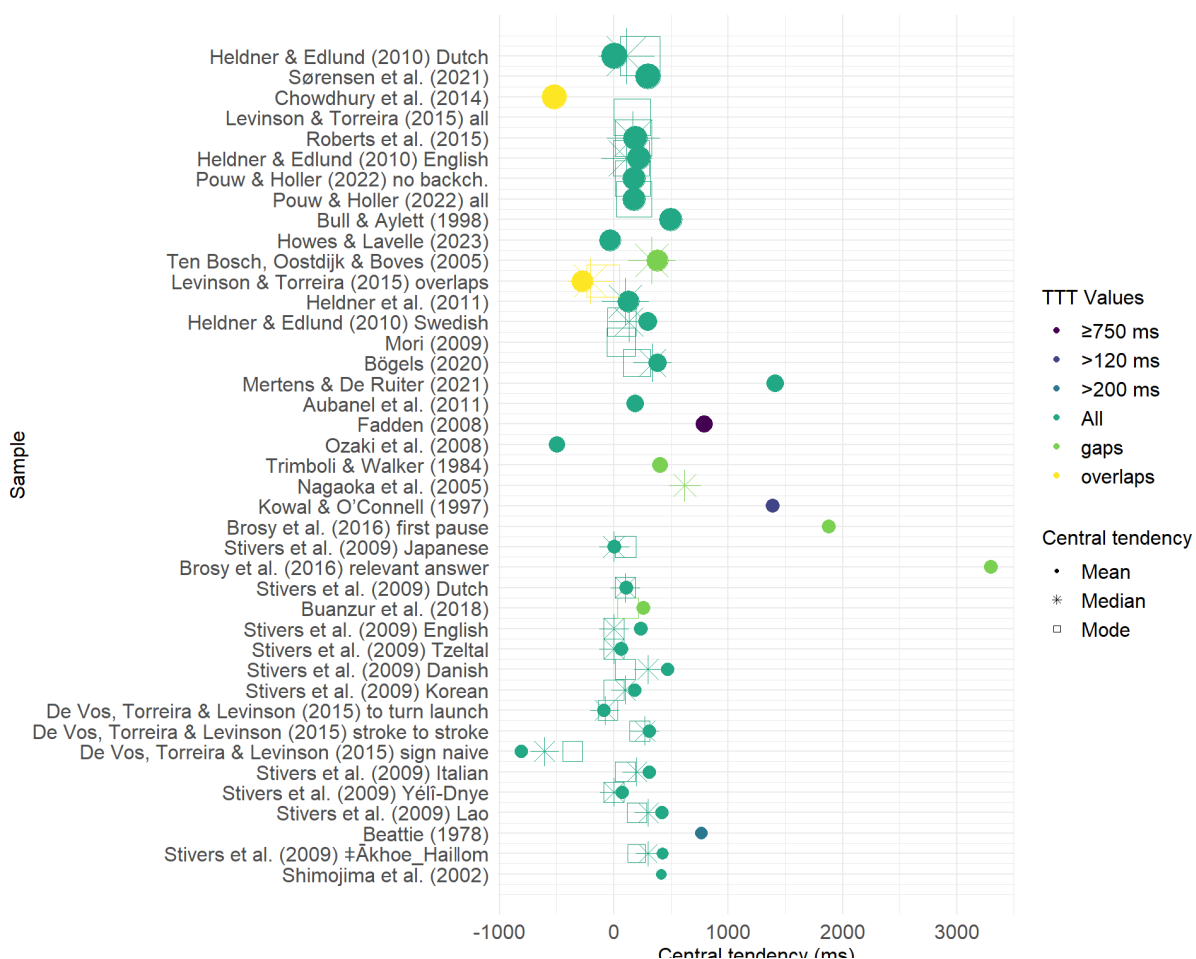


Figure 3.14 The mean, median and mode of distributions in the entire selection of samples arranged by sample size. The sample estimates are colour-coded by TTT Values. Point size reflects the natural logarithm of the sample size.

3.3.10 Entire selection

Subsets 1 and 2 consisted only of samples that included all values of TTT. To visualise the effect of including samples with only certain TTT values, the forest plot in Figure 3.14 shows, for the entire selection, which TTT values are included. The pattern in Figure 3.14 is mostly as expected: samples that report only gaps tend to have a higher mean, and samples that report only overlaps tend to have lower means, respectively, than samples that included all TTT values. Some of the exceptions concern samples that include all TTT values and nevertheless have very high or very low measures of central tendency. These are samples that have already been discussed above: an exception towards the high end is the sample by Mertens and De Ruiter (2021), who include all values of TTT but only repair sequences. Exceptions on the low end are De Vos et al.'s (2015) 'sign naïve' sample, who compared alternative operationalisations of turn timing, and Ozaki et al. (2008), who studied navigator-driver interactions.

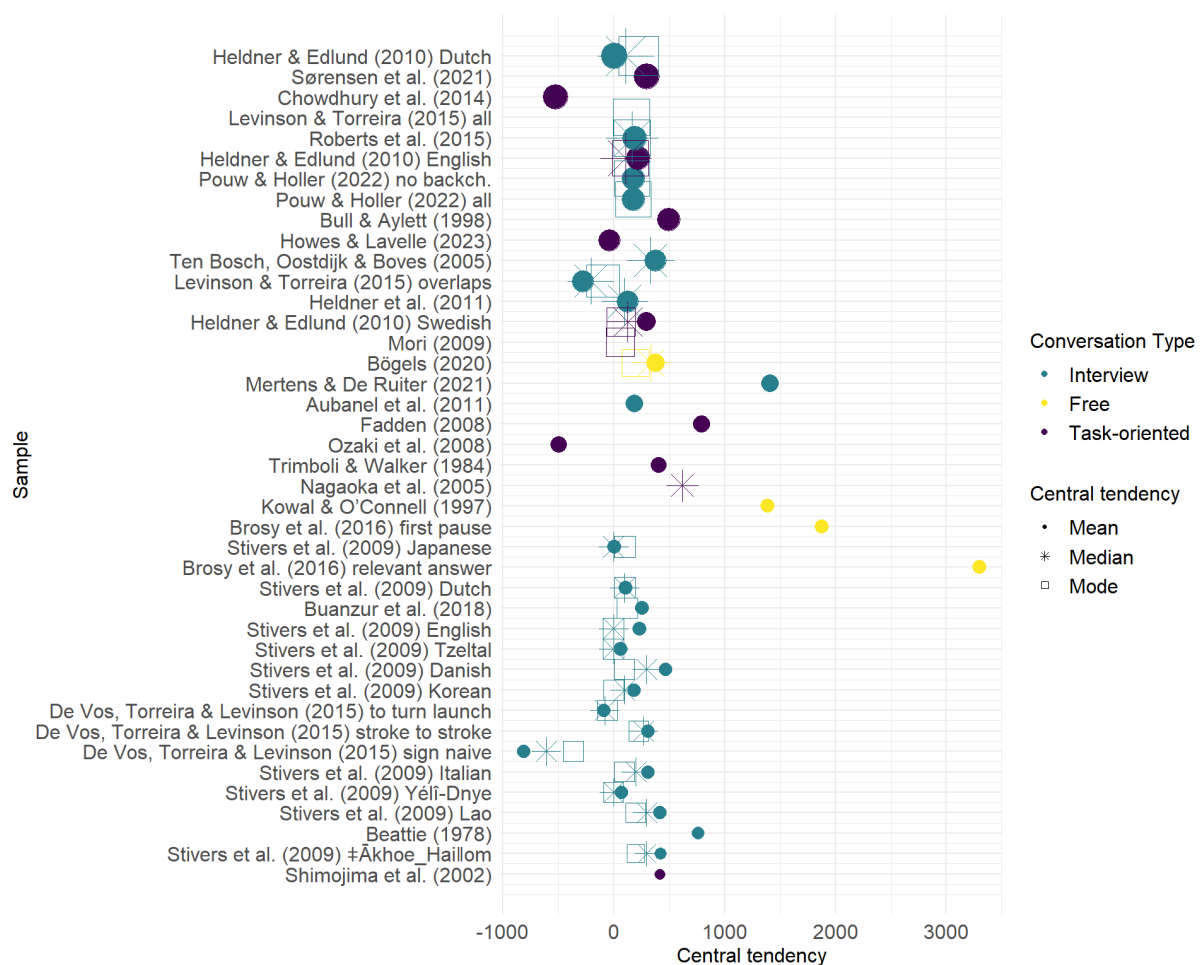


Figure 3.15 The mean, median and mode of distributions in entire selection, arranged by sample size. The sample estimates are colour-coded by Conversation Type. Point size reflects the natural logarithm of the sample size.

To assess the effect of Conversation Type on the measures of central tendency across the entire selection, Figure 3.15 again includes all 41 samples, now colour-coded by Conversation Type. As in subset 2, the pattern that the means of task-oriented conversations are spread more widely than the means of free conversations is not clearly supported by Figure 3.15. At face-value, the means of interviews appear relatively large. However, there are only four samples from interviews, and three of those include only gaps. The most extreme, Borsy et al.'s (2016) 'relevant answer', includes filled pauses in the TTT measurements.

3.3.11 Languages

As discussed in Sections 2.2.2 and 3.1.1, Stivers et al. (2009) found only small differences between turn-transition times across ten languages. Although their overall conclusion leans to the view that turn timing follows universal constraints, they do speculate that there might be language- or culture-specific calibration to listeners' sensitivity to delay based on their annotators' judgements. More concretely with regard to production, Sneek (1987) found that Finnish speakers have longer turn-transitions times than speakers of American English. This difference is 385ms, which is they report is statistically significant. Although not conclusive, these observations suggest there may be small but robust differences between languages or communities in how they time turns.

To assess whether language-related differences reliably exist across studies, we included all samples from subset 2 from a language that is represented at least twice in the entire selection of 41 samples, and plotted the means, medians and modes of the remaining samples, grouped by language, Figure 3.16. summarises the means and indicates the range of means per language. Panel a) shows means, medians and modes per sample arranged by sample size and colour-coded by Language. Point size reflects the natural logarithm of the sample size. b) shows the mean of means and range of means per language. Figure 3.16 does not suggest that language is a meaningful predictor of turn timing in this subset, since the samples do not group together by language. However, besides representing only four different languages, these samples differ in terms of the dialogue acts and conversation types included, as well as turn timing operationalisation. The lack of evidence of a grouping by language in Figure 3.16 therefore primarily indicates that a potential language-related effect, if it exists, is not robust to the variation imposed by other variables, including methodological variables.

3.3.12 Repeated observations from the same corpora

As mentioned, the entire selection of samples contained multiple samples from the same corpora. To avoid biasing our results with repeated observations of the same data, subset 1 only included the sample from each corpus that was published earliest, and subset 2 only included samples from the same corpus if a different decision was made by the authors about which dialogue acts to include. This section evaluates the extent to which samples from the same corpus group together. If they tend to group together, this

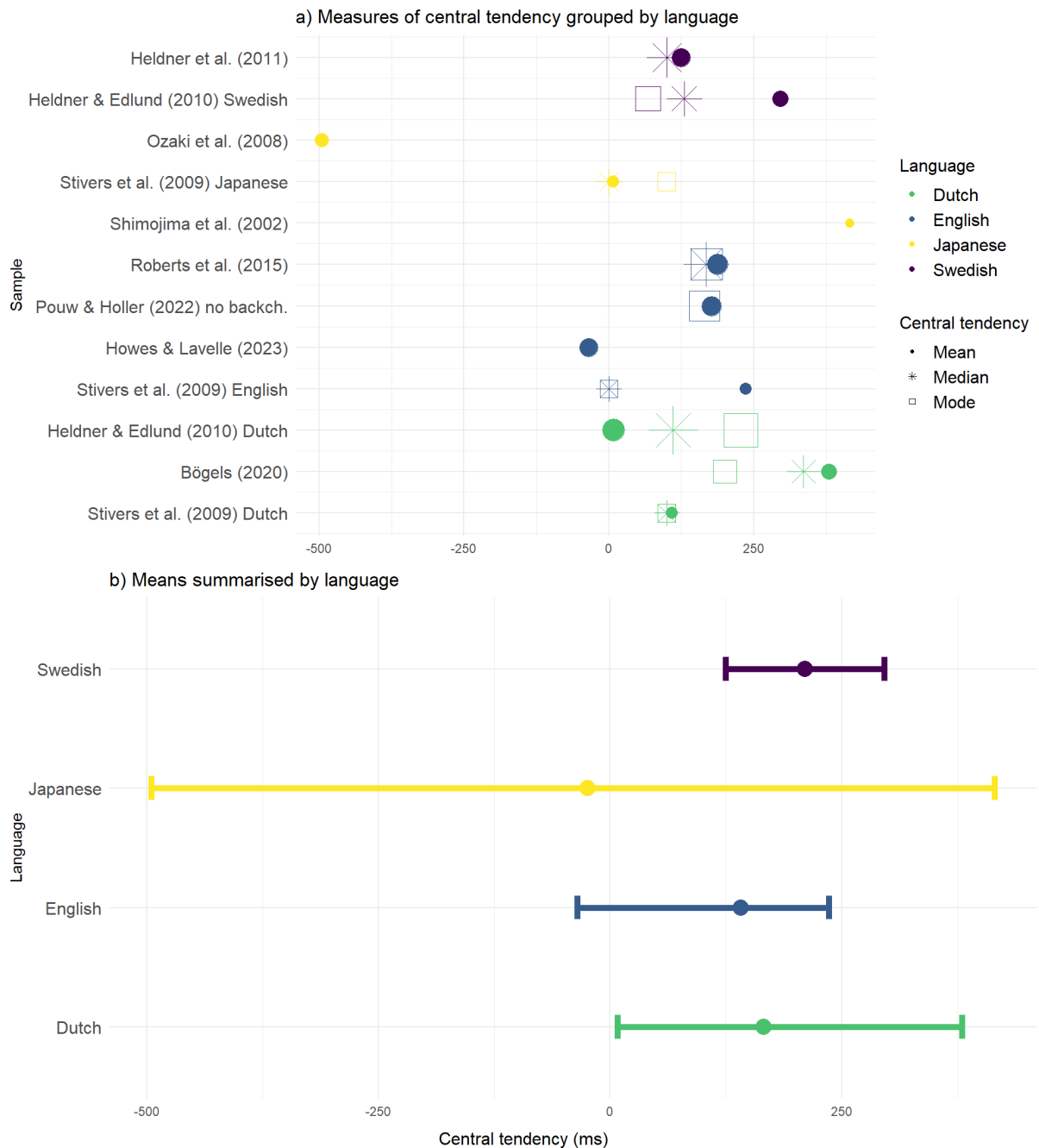


Figure 3.16 Measures of central tendency per language for all samples in subset 2 that were from a language that was included at least twice in subset 2.

may indicate that repeated observations do bias our results, but also that turn timing measurements are robust to variation across studies. Conversely, if they do not group together, this indicates that a large degree of variation of turn timing estimates across studies might be attributable to methodological differences.

Our data provide limited evidence on this issue. As illustrated in Figure 3.17, three corpora were used more than once. Additionally, the NGT Interactive corpus (of sign language) was also used in three samples. However, all three samples came from the report on Sign Language of the Netherlands that aimed to compare three alternative turn timing operationalisations. Similarly, Brosy et al. (2016) report two samples acquired by two turn timing operationalisations. As their comparison was already the objective of the original studies and has been discussed in sections 3.3.5 and 3.3.7, it is not discussed further here.

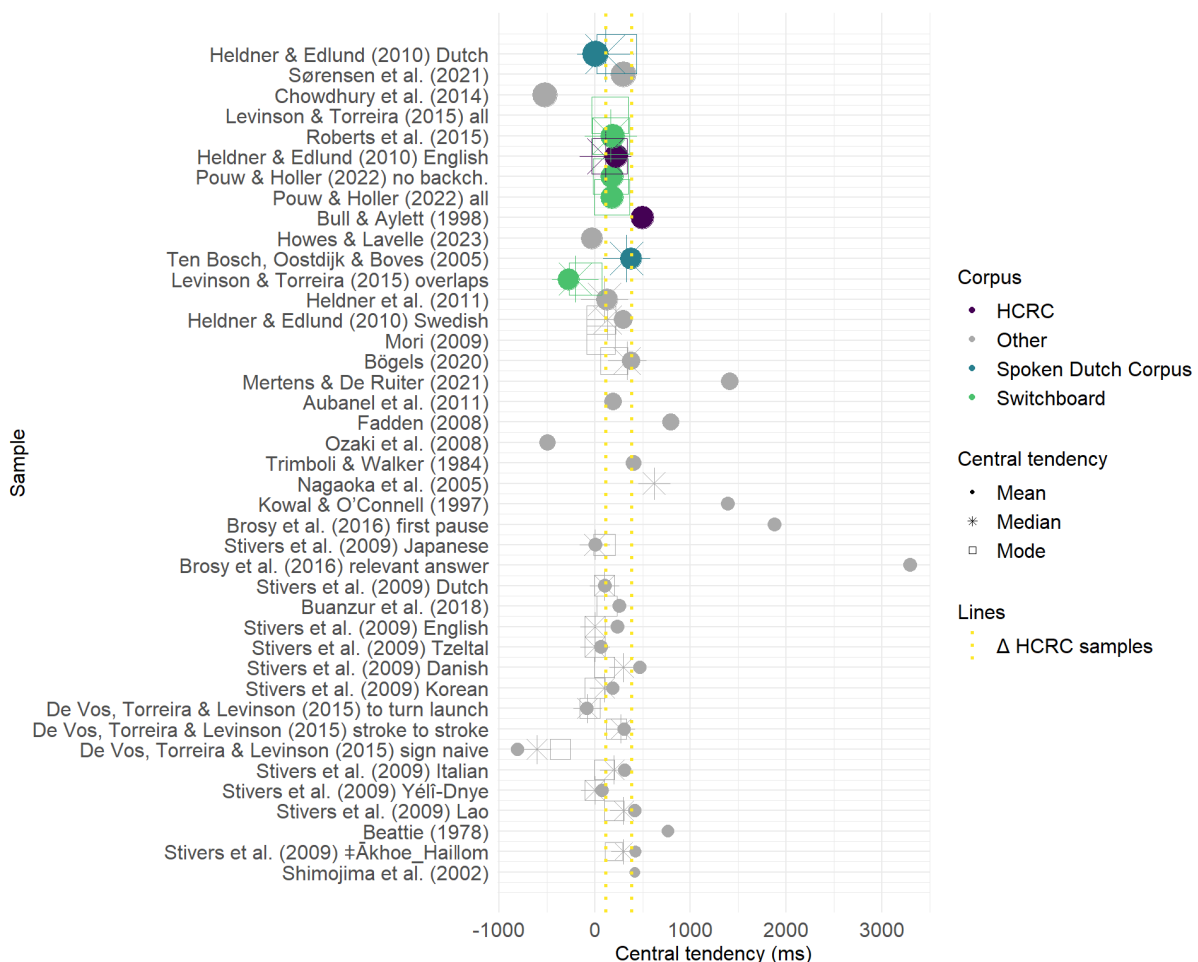


Figure 3.17 The mean, median and mode of distributions in the entire selection of samples, arranged by sample size. The sample estimates are colour-coded by Corpus. Point size reflects the natural logarithm of the sample size.

Of the remaining three corpora, the Switchboard corpus was used by Roberts et al. (2015), by Levinson and Torreira (2015) for two samples, and by Pouw and Holler (2022) for two samples. The HCRC corpus was used by Bull and Aylett (1998) and Heldner and Edlund (2010). The CGN was used by Ten Bosch et al. (2005) and Heldner and Edlund (2010). To compare samples from the same corpora, Figure 3.17 depicts all measures of central tendency, this time colour coded by whether they used one of these three corpora.

The samples from the CGN, Heldner and Edlund's (2010) Dutch sample and Ten Bosch et al.'s (2005) sample, do not group together particularly closely, with a difference in their means of 372ms and a difference in medians of 219. Neither do the samples from the HCRC, Heldner and Edlund's (2010) English sample and Bull and Aylett's (1998) sample, with a difference in mean of 284ms. For comparison, the dotted yellow lines in figure 3.18 indicate a range of 284ms around the median of the means, to illustrate that a large proportion of the other samples fall within this range: Of the 39 other samples, 15 have a mean that lies within this range (from 105 to 390 ms).

The means and medians of four of the five samples from Switchboard are more similar. This similarity is likely to stem from the fact that they took the same or very similar approaches to extracting the data: Roberts et al. (2015) and Levinson and Torreira (2015) both applied the operationalisation of turn timing used by Heldner and Edlund (2010) and identified turn boundaries using the existing segmentations in the corpus (note also that there is an overlap in authorship between Roberts et al., 2015, and Levinson and Torreira, 2015). Pouw and Holler (2022) use the already extracted data from Roberts et al. (2015). The 'overlaps' sample from Levinson and Torreira (2015) diverge from the others because it includes only overlaps, and the small differences between the others may be due to slightly different data filtering choices.

The larger difference between Heldner and Edlund's (2010) Dutch sample and Ten Bosch et al.'s (2005) sample might be attributable to the fact that Heldner and Edlund (2010) used a lower and upper duration limit for including TTT of -2000ms to 2000ms, while Ten Bosch et al. (2005) did not. They also used different sub-corpora of the CGN and are therefore not strictly speaking observations of the same data. Ten Bosch et al. (2005) used only the telephone conversations from the corpus, whereas Heldner and Edlund used both telephone conversations and face-to-face conversations. However, Bull and

Aylett (1998) reported that turn timing in conversations without eye-contact is significantly faster, by 157ms, than in face-to-face conversation, so it is unclear why turn timing in Ten Bosch et al.'s (2005) telephone-only selection is slower than in Heldner and Edlund's (2010) combined selection.

The difference between Bull and Aylett's (1998) sample and Heldner and Edlund's (2010) sample from the HCRC corpus may be due to methodological differences. They both operationalised TTT as the difference between the end of one turn and the start of the next. However, for their English sample, Heldner and Edlund (2010) defined turns as automatically detected stretches of speech activity, Bull and Aylett (1998) defined turns as made up of one or more 'game moves'. Their coding scheme defines game moves as "[...] an utterance, a partial utterance, or a group of utterances which convey the same, specific intent." (Isard et al., 1993). The coding scheme does not specify how move boundaries are identified. The difference between Heldner and Edlund (2010)'s English sample and Bull and Aylett's (1998) sample could therefore lie in the fact that, in the former, a speech activity switch could occur during a game move, which the latter would not have measured as a turn switch.

To summarise these observations on samples from the same corpora, this selection of samples provides limited information on the stability of turn timing estimates across different methodological choices, as only three corpora were used more than once, and the differences are attributable to selection choices. However, considering that the two samples from the HCRC corpus differ relatively widely compared to the variation between all samples in the selection, it seems that differences in sample estimates caused by slightly different selection procedures are potentially very large.

3.4 Conclusion

This chapter aimed to describe the range of variation of turn timing distribution across studies and to identify similarities and differences between those studies that might explain that variation. It did so through a qualitative analysis of secondary data from 41 turn timing samples identified through a systematic review. Because the 41 samples were not homogeneous, two subsets were also analysed: subset 1 consisted of the samples that included all TTT values and in which all dialogue acts were represented, and subset 2 consisted of samples that included all TTT values and any selection of dialogue

acts (e.g., only question-response sequences). In the following paragraphs, the findings are summarised with reference to the research questions stated in Section 3.1.4.

1.1 What estimates of turn timing are reported in studies of spontaneous conversation between adults?

The measures of central tendency of the samples ranged from -812 to 3300ms, although the majority clustered between 0 and 500ms. In all subsets, the means of the samples ranged more widely than the medians and the modes, consistent with previous observations that the mean is not a stable measure of central tendency for turn timing distributions. In subset 1, the medians and modes were particularly narrowly distributed between 68 to 228ms, somewhat close to the frequently cited average TTT of 200ms, but this variation was much wider in subset 2 and in the entire selection. Additionally, the turn timing distributions were generally, though not consistently, right skewed. This skewness was often due to a heavy shoulder, potentially indicating clustering. The distributions also appeared to be leptokurtic, having high peaks, thick tails, and narrow shoulders.

1.2 What operationalisations of turn timing are used and what are their differences?

The reports included in this review all operationalised turn timing as the time between the end of the first turn and the start of the second (resulting in a negative value in the case of overlaps), although they differed in how they identified turn boundaries and what they considered to be a turn boundary. Many reports did not define turn boundaries at all, particularly the end of the first turn. In other words, although all reports used comparable operationalisations of turn timing, many did not provide precise phonetic definitions of these operationalisation.

1.3 What factors are found to have a consistent effect on turn timing?

1.4 To what extent is there variation between turn timing estimates that cannot be accounted for by factors considered within the studies?

No clear predictors of the variation between samples emerged from this analysis. The type of conversation appeared to affect the means, but not the medians or modes, of samples in subset 1. Specifically, the means of task-oriented conversations were more widely distributed than the means of free conversations. This pattern suggests that task-

oriented conversations are more likely to involve “outliers”, i.e., occasional very long overlaps or silences, compared to free conversations. Qualitative analyses comparing extreme TTT values in task-oriented and free conversations could help clarify how task demands give rise to this pattern. However, the same pattern did not emerge in subset 2 or the entire selection of samples, potentially because those were less homogeneous. The remaining variation between samples was not explained by the language spoken in the conversations. Additionally, we found some evidence that repeated observations from the same corpus can vary widely.

This lack of clear patterns in the effects on turn timing distributions across studies could, in part, be due to substantial unidentified methodological differences between the studies. It could also be due to the size of our selection of studies – with only 41 samples, there were relatively few samples for each level of the variables of interest, and many of the variables were correlated. In this study, we opted to apply our quality criteria strictly and exclude reports that did not define turn timing or did not report sample sizes, to obtain a selection of samples that was as homogeneous as possible. As a result, many otherwise eligible samples were not included. A secondary analysis that uses a more permissive application of these criteria would involve a larger selection of samples, in which correlated variables may be less problematic. Additionally, the search strategy used in this study to identify eligible turn timing samples was designed with feasibility in mind but may have missed relevant reports due to limitations of the search query or because some reports were not indexed in the databases. An extension of this review, using citation chaining (screening the records cited by, or citing, included reports) as an additional search strategy, is likely to identify a larger selection of eligible samples. Finally, including other types of data, such as raw data obtained from authors, or proportional data, could support a more fine-grained analysis of differences between distributions.

In conclusion, while this analysis highlighted some tendencies, such as the clustering of measures of central tendency between 0 and 500ms, distributions tending towards a right skew and high kurtosis, and a potential effect of conversation type, it also suggests that methodological variation between studies may have a significant effect. Moreover, there seems to be little consistency, and especially specificity, in how turn timing is operationalised between studies. For this reason, the effect of methodological choices

on the outcome of the studies cannot be clearly identified. To address this problem, further research on turn timing could aim to provide definitions of turn timing and turn boundaries that are not just pragmatically and lexically, but also phonetically and acoustically specific. Chapter 6 returns to the issue of turn timing operationalisations.

Chapter 4. Speech Rate and Turn Timing in Spontaneous Question-Response Sequences

Parts of this chapter were previously published in:

Hoogland, D., White, L., & Knight, S. (2023). Speech Rate and Turn-Transition Pause Duration in Dutch and English Spontaneous Question-Answer Sequences. *Languages*, 8(2), 115.

4.1 Introduction

This chapter examines how speech rate relates to turn timing in spontaneous conversation. Its main aim is to test the perception-action entrainment hypothesis by coupled oscillator models of turn timing, which predicts that faster speech rates are associated with faster turn timing and vice-versa (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005). Previous studies that tested this hypothesis have come to contrary conclusions, possibly due to methodological difference between speech production experiments and observational studies of spontaneous conversational data, or because of interactions between the effects of speech rate and other predictive factors of turn timing. For this reason, this study investigates the perception-action entrainment hypothesis in spontaneous question-response sequences – the dialogue context that experimental paradigms have attempted to model – and asks whether there is an interaction between global and local speech rate and dialogue factors.

Section 4.1 first summarises the discussion in Section 2.3 on models of turn timing to motivate this study (Section 4.1.1), then states this chapter's research questions (Section 4.1.2) and discusses methodological choices made in its design (Sections 4.1.3 to 4.1.7).

4.1.1 Motivation

We can divide models of turn timing into two types: the *prediction-reaction* models, which propose that turn timing is the result of a combination of processing times (e.g., Levinson & Torreira, 2015), and the *coupled oscillator* models, which propose that turn timing is the result of neural entrainment between interlocutors and a perception-action link (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005). The main difference between

these two types of models is their predictions regarding the role of speech rate in turn timing. The coupled oscillator models' main hypothesis is that turns uttered with faster speech rates are related to shorter turn-transition times (TTT): the perception-action entrainment hypothesis. Coupled oscillator models also predict that turn timing distributions show periodic patterns or clusters in density above and below 0ms TTT.

In contrast, prediction-reaction models do not explicitly predict a particular turn timing distribution or relationship between speech rate and turn timing, although some of its proponents have speculated that faster turns are associated with *slower* subsequent turn timing because faster turns are more difficult to process (Levinson & Torreira, 2015; based on S. G. Roberts et al., 2015; see below). This effect may or may not be linear, since very slow turns are likely also harder to process (Bosker & Ghitza, 2018), which might result in a quadratic effect of speech rate on turn timing. On the other hand, slower turns also leave more time for the listener to process their content, and to identify turn ending cues (cf. Torreira & Bögels, 2022), which would promote a positive linear relationship.

Studies that aimed to test the perception-action entrainment hypothesis reported contradictory evidence. On the one hand, a corpus study by Beňuš (2009) found weak evidence in favour, while another by Roberts et al. (2015) concluded that although speech rate is the fourth most important predictor of turn timing out of 32 predictors, its relationship with turn timing appears to be *positive*, so contrary to the model's predictions. On the other hand, experimental studies, particularly one by Corps et al. (2020), found that participants responded faster to faster questions, in favour of the perception-action entrainment hypothesis.

As discussed in Section 2.3, this apparent contradiction between previous studies' findings might be due to methodological differences. These methodological differences include the choice of taking an experimental or observational approach, since spontaneous conversation and turn-taking tasks meant to mimic conversation may differ in their demand characteristics. Others differences are the choice to study a particular type of turn sequence, such as polar question-response sequences in an experimental paradigm (Corps et al., 2020) or all speech activity offsets (Beňuš, 2009; S. G. Roberts et al., 2015), or the choice of statistical analysis tools. Particularly, neither Beňuš (2009) nor Roberts et al. (2015) used multivariate statistical methods to investigate whether the

observed relationship is modulated by other factors, while more recent experimental results by Torreira and Bögels (2022) suggest it might be.

To examine this issue more closely, the corpus analysis presented here looks at a specific subcase of turn timing in spontaneous conversation, namely the timing of responses to questions – question-response sequences may constitute behaviour that is more like experimental tasks, since those tend to use a question-response paradigm. Within this constrained class of dialogue acts, this study tests whether faster speech rate is associated with shorter TTT, as predicted. As Corps et al. (2020) found that speeded of final word of questions were associated with faster timing independently from the speed of the question (using one of two operationalisation; see Section 2.3.6), this study also tests whether the local rate of the final syllable rhyme influenced turn timing, in addition to the global speech rate of the turn.

Aside from the primary aim of testing the perception-action entrainment hypothesis, this study also tests whether any association between speech rate and turn timing can best be modelled as linear or quadratic, whether that association is modulated by turn length and subcategories of dialogue acts, and whether the speech rate of the response is associated with its turn timing. It also explores whether the shape of TTT distributions show signs of clustering, as predicted by the Wilson and Wilson (2005).

The next section summarises this study's research questions and discusses its methodology.

4.1.2 Research questions and methodology

To summarise, this chapter primarily aims to test the prediction-action entrainment hypothesis, asking the following research questions:

2. Is speech timing (particularly speech rate) predictive of turn-transition time, and if so, what is the nature of their relationship, and is it modulated by other factors?

Hypothesis-testing

- 2.1. Is there an inverse relationship between the Question's Speech Rate and TTT?

2.2. Is there an inverse relationship between Final Lengthening (Final Rhyme Duration) and TTT?

This study also addresses the following exploratory questions:

Exploratory questions:

2.1 Is the relationship between the Question's Speech Rate and TTT best described as linear or as quadratic?

2.2 Are any observed effects of speech timing factors modulated by dialogue factors already shown to influence TTT (Question Length, Question Type, Response Length and Response Type)?

2.3 Is there an inverse relationship between the speech rate of the response and TTT?

2.4 Do we observe any reliable differences, in the above factors, between Dutch and English?

2.5 What is the shape of the TTT distributions in this data?

The following sections discuss the methodology of this study.

4.1.3 Question-response sequences

First, as mentioned, this study focusses on question-response sequences, in part because experimental studies that found evidence in favour of the prediction-entrainment theories used a question-response paradigm (Corps et al., 2020). In other words, if those experimental tasks usefully correspond to conversational behaviour, it seems most likely that they correspond to question-responding behaviour.

Other reasons for using question-response sequences are that turn timing in this context has been studied widely from a quantitative perspective, so other effects on timing are relatively well described (both experimentally, e.g., A. S. Meyer et al., 2018; and in spontaneous conversation, e.g., Buangur et al., 2018; Stivers et al., 2009; see also Chapter 3). Although the previous chapter did not find that question-response sequences have different measures of central tendency than other sequences, question-response sequences in dyadic conversations are also contexts in which speakers are likely oriented towards responding, as the next speaker is selected by virtue of the form. This

feature has led researchers to focus on question-response sequences to measure features of speech timing (De Vos et al., 2015), and might make timing mechanisms less sensitive to noise (cf. Couper-Kuhlen, 1993, p. 178, who expects turn switches to be more rhythmically constrained when their progression is determined).

Notwithstanding this focus on speech rate and turn timing in question-response sequences, we also analysed the automatically extracted TTT distributions of all speech activity switches in the corpus, in addition to the manually extracted TTT distributions of question-response sequences (see below).

4.1.4 Question type, response type, question length and response length as modulating factors

In addition to testing the effects of speech rate on turn timing directly, this study also investigates the role of four additional variables. These include two variables related to dialogue act, namely question type and response type, and two variables related to timing, namely question length and response length.

First, the experimental paradigm of Corps et al. (2020; see also Franich, 2023) included only polar questions – questions that can be answered with “yes”/ “no”. Polar questions may potentially differ in timing from other questions because they constrain the possible responses, which may alter the processing demands placed on the responder (Bögels et al., 2020; Knudsen et al., 2020), although we should note that the actual processing demands in spontaneous conversation are likely to be subject to many more influences. To allow a comparison to the experimental studies, but at the same time compare the effect of dialogue acts, we included all questions and used a variable ‘question type’ to distinguish between polar questions and others.

Second, we compared responses that answered the question to non-answer responses. Response type was included because it has consistently been shown to affect TTT in question-response sequences: specifically, non-answer responses are slower than responses judged to answer the question in a range of languages (Buanzur et al., 2018; Stivers et al., 2009).

This study also explores the influence of two additional timing-related variables: question length and response length. Question length was included because Roberts et al. (2015) found turn length to be a relatively important predictor of TTT: thus, longer turns (and very

short turns, which are excluded in this study; see below) have longer TTT relative to turns of intermediate length. Additionally, question length may interact with the relationship between speech rate and TTT, on the one hand because longer utterances allow listeners more time to entrain to the local speech rate. Response length was also included, for the similar reason that shorter turns have been observed to be timed faster and closer to the end of the previous turn in spontaneous conversation (Heldner et al., 2011), and were faster in experiments, possibly because of processing demands (A. S. Meyer et al., 2018).

By including question- and response type and length in this analysis, it is possible to compare the contribution of the primary variables of interest, speech rate and final rhyme duration, relative to widely established predictors of TTT. Of these four, Roberts et al. (2015) found, in their random forest analysis that ranked the importance of 32 variables in predicting turn timing, that response type, question length and response length are the three variables ranking higher than speech rate, which came fourth. By using multiple regression, rather than a series of bivariate correlation analysis as in the study by Beňuš (2009) or a random forest analysis as in the study by Roberts et al. (2015), we can also more directly assess interactions between the variables of interest and these additional variables.

4.1.5 *Linearity*

The coupled oscillator models 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. Therefore, using linear methods, as previous studies did, relies on the assumption that either 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 process is linear in line with prediction-reaction models. This study adopts this assumption.

4.1.6 *Manual extraction of data*

This study uses a relatively small, hand-aligned dataset, to complement previous corpus studies that used large, automatically extracted datasets (Beňuš, 2009; S. G. Roberts et al., 2015). It aimed to ensure that the speech timing variables (speech rate, final syllable rhyme duration, TTT) were measured as consistently and accurately as possible. Additionally, 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; see section 2.2.3). However, we compare the distributions of manually extracted TTTs of question-response sequences to distributions of automatically extracted TTT distributions of all speech activity switches in the corpus.

4.1.7 Languages and corpora

Finally, this study includes two corpora from different languages, Dutch and English, and considers potential differences between the two languages in an exploratory analysis, noting that Stivers et al. (2009) found a difference in the mean TTT of samples from these two languages of around 127 ms. The primary question regarding language is whether any speech rate effect on TTT is modulated by language. However, any observed presence or absence of differences between Dutch and English should not be over-interpreted, because 1) Dutch and English are typologically related and prosodically and syntactically similar and 2) those differences may relate to corpus-specific factors as much as to language per se. The previous chapter also discussed that the observed differences between these languages do not appear to be robust across studies. To distinguish the effect of language from the effect of other global characteristics of the corpora, a larger dataset, ideally with multiple multilingual corpora, is needed.

The following sections first describe the study methods (4.2), a preliminary analysis of the corpus data (4.3.1 to 4.3.6), and the results of the experimental and exploratory analyses (4.3.7 to 4.3.11), which are discussed in section 4.4.

4.2 Methods

4.2.1 Corpora

The data analysed in this study were taken from two conversational corpora: The *Instituut voor Fonetiek Amsterdam Dialogue Video Corpus* (Institute for Phonetics Amsterdam Dialogue Video Corpus, IFADV, Van Son et al., 2008) contains conversations between Dutch speakers, and the Leverhulme corpus (White et al., 2012) contains conversations between English speakers.

The IFADV and Leverhulme corpora were selected because they are comparable in many aspects; they both consist of free lab-based conversations between two speakers who are either friends or colleagues in university environments (as students or coworkers). However, any two corpora collected under different circumstances, by different research

teams, and for different purposes, are likely to be different to some extent, and some of those differences might affect speech timing or turn timing. This section describes key characteristics of the corpora and highlights potential differences.

The **Dutch** IFADV corpus contains 20 lab-based dyadic conversations, each of which is a 15-minute continuous sample from a longer conversation. The published corpus includes audio recordings in stereo WAV file format, video recordings of the participants, and Praat TextGrids with annotations for dialogue act, gaze direction, orthographic transcriptions, and automatic word-level alignments. This study includes annotation data that was gathered with reference to the WAV files and orthographic transcriptions, dialogue act and phoneme alignments. The video recordings and gaze annotations were not used.

Conversation id	Speaker age	Speaker gender	Relationship	Duration (min)
DVA3E	64	M	Friends, Colleagues	15
	29	F		
DVA4C	29	M	Colleagues	15
	62	F		
DVA6H	59	F	Colleagues	15
	65	M		
DVA7B	30	F	Colleges, Distant	15
	53	M		
DVA8K	20	F	Friends	15
	20	F		
DVA9M	21	F	Friends	15
	20	F		
DVA10O	18	F	Close, Informal, Friends	15
	19	F		
DVA11Q	28	F	Close, Informal, Friends	15
	28	F		
DVA12S	19	F	Friends	15
	20	F		

Table 4.1 Overview of interlocutor characteristics in the Dutch IFADV corpus

Of the 20 conversations of the IFADV, 9 were included in this study. The included conversation IDs are listed in Table 4.1 (column 1). Conversations DVA1A and DVA2C were not included because they were used in a pilot analysis, and conversation DVA5G

was published without an associated transcription file. Only 9 conversations were included to match the approximate total length of the Dutch sample to that of the English sample (see below).

Of the 18 Dutch speakers, 14 were female. They ranged in age from 18 to 65 years ($M = 34$ years) and spoke standard Netherlands Dutch with various regional accents, mainly accents associated with areas in the central or northern Netherlands. All pairs were described as friends or colleagues (Van Son et al., 2008). Table 4.1 describes speaker characteristics.

The **English** Leverhulme corpus consists of eight dyadic free lab-based conversations, as well as read speech and task-oriented dialogues. Only the free conversations are included in the analysis presented here. The conversations varied in length between 16 and 22 minutes, as described in Table 4.2. The corpus includes Praat TextGrids with, for each speaker, a tier with orthographic transcriptions, indicating instances of overlap, silences, and silent and filled pauses. It also includes annotator judgements on which speaker holds the conversational floor, and which speaker is currently attending to the other – i.e., pauses between turns and backchannel behaviour are marked differently from pauses between turns that occupy the conversational floor. Only the orthographic transcriptions were used in this study.

Conversation	Duration (min.: sec.)
fr-s01s02	21:39
fr-s03s04	21:14
fr-s05s06	21:09
fr-s07s08	21:03
fr-s09s10	19:23
fr-s11-s12	17:03
fr-s13s14	17:04
fr-s15s16	16:17

Table 4.2 Overview of conversation duration in the English Leverhulme corpus. Speaker demographic information is not available

The 16 English speakers all spoke Standard Southern British English (SSBE). The individual participants' ages, gender or relationships to their interlocutor were not described, but they are reported to be students at the University of Bristol with ages

between 18 and 30 years old, so the age range was smaller than in the Dutch corpus. They were at least acquainted with each other through participating in a series of speech- and conversational tasks as part of the corpus construction, but some appear to have known each other prior to the recording. For example, speaker s15 refers to a person not previously named in the conversation, while, in contrast, speakers s01 and s02 ask each other about their university degree or hobbies in a way that suggests they did not know each other well enough to know such personal information prior to the recording. In contrast, in the Dutch corpus, all dyads are reported to be acquainted with each other through contexts unrelated to the corpus collection.

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. The speaker pairs in both corpora were given some suggested topics for discussion and occasionally used them.

4.2.2 *TTT measurements in the full conversations*

All speech activity switches were extracted from the conversations automatically from the pre-existing utterance-level alignments of transcripts in the following way. For each utterance U1, it was first determined if its onset occurred during silence, or during the utterance of the other speaker as indicated by the transcript. If U1 started during silence, the offset of the nearest previous utterance U0 was identified. If U0 was by a different speaker, the time between U0's offset and U1's onset was recorded as a positive TTT. If U1 started during another utterance U0, the time between U1's onset and U0's offset was recorded as a negative TTT.

4.2.3 *Sampling of question-response sequences*

For each corpus, an exhaustive set of question-response sequences was also collected as follows. All questions were identified and delimited by their syntactic and/or prosodic structure, and by their apparent response-evoking purpose within the context of the dialogue. This includes utterances with the syntactic, lexical, or prosodic form of a question, but that pragmatically appeared to be a request for a statement of agreement, rather than asking for information. In Dutch, the speech act annotations and orthographic transcriptions were consulted to identify questions, but the annotator's (the author of this thesis) judgement was the criterion for inclusion. In the Dutch corpus, the boundaries were placed in accordance with the guideline in Section 4.2.4. In the English corpus, the

points at which questions end and response start, were matched to the boundaries already marked in the corpus transcription.

Next, questions with the following characteristics were excluded. First, those questions that did not conclude in a turn switch, but in a continuation by the same speaker were not included. In this study, only conversational activity that included vocalizations was considered a turn. In other words, instances where the second speaker (henceforth *responder*) responded to a question with a gesture only, followed again by speech from the person asking the question (henceforth *questioner*) were not included. Although gestures can function as turns by themselves within the structure of a conversation (cf. Dingemanse & Floyd, 2014; Stivers et al., 2009), in this study they were not considered. This decision was made because one of the aims of this study is to test the proposal by Wilson and Wilson (2005) that oscillations entrained to auditory input controls speech initiation, which they hypothesise to have the function of optimizing perceptual salience of speech. If this hypothesis is correct, then paralinguistic gestures may not be timed in the same way as (linguistic) speech events. This is an assumption, since gestures may well be used to maintain entrainment, if the visual and auditory perceptual systems are linked in the appropriate way. However, only the Dutch corpus contains information about the timing of gestures, so a pragmatic second reason for not including gestures was for the sake of comparability between the Dutch and the English corpuses.

Second, utterances with the form of a question that did not appear to be intended to prompt a directed response were excluded. These utterances included questioning backchannels (e.g., “oh really?”) that were followed by a continuation by the responder of a previous turn rather than a response. This category also includes questions that were reported speech, such as in example 1).

- 1) Speaker 1: And then she asked: ‘Did you see her again?’ To which I said
Speaker 2: mhm.
[schematic example]

The response to the question was identified as the start of the speech by the responder until the nearest point of syntactic, pragmatic and prosodic completion. Often, this was all the speech by the responder up until the next turn by the questioner. However, in the following cases the end of the turn was not considered the end of the response.

First, when the responder responds to a question, then pauses, and then changes the topic, the speech before the pause was considered the response.

Second, when the response was interrupted or overlapped by speech from the questioner, the moment the responder stopped speaking was considered the end of the response. Backchannels by the questioner were not considered interruptions of the next turn by the questioner.

Third, questions were often followed by a multi-utterance elaboration by the responder. In such cases, the first syntactically complete utterance followed by a pause was considered the response. In the case of responses to polar questions, when the responder responded with a yes or no, a short pause, and then another utterance that aligns with the question as in example 2), the second utterance was considered part of the response.

2) Q: Did you go?

R: Yes, I did go actually.

[schematic example]

There were two reasons for not including the entire stretch of speech activity by the responder until the questioner's next speech are as follows, which aligned with the assumptions of the double oscillator models and prediction-reaction models respectively. First, during a long-elaborated response, the responder may change speech rate, meaning that the average speech rate of the whole turn is less representative of the speech rate in the domain relevant to the turn switch than the speech rate of the first utterance alone. Second, a long turn that contains many pauses may not be planned by the speaker before the start of their turn. The whole length of that turn is therefore less likely to reflect a relevant aspect of speech planning than the length of the first utterance.

4.2.4 Speech- and language measurements

For each question-response sequence, the following variables were measured.

Turn-Transition Time Turn timing was operationalised as the duration of the interval between the end of the first turn (the question) and the start of the second turn (the response) measured in milliseconds (TTT). As is conventional, gaps and overlaps are measured on a single scale, i.e., TTT is positive when there is a gap between the question

and the response, and negative when there is overlap, although in the analysis presented here, they are considered both together and separately.

The start of the answer was identified as the initiation of the first speech sound of the response. We excluded paralinguistic clicks or inbreaths, but included filled pauses such as ‘uhm’ in the response.

Generally, the end of the question was identified as the end of speech activity (coinciding by the end of periodicity distinguishable from noise) by the questioner. However, as discussed in section 2.3, a functional operationalisation of TTT, as the moment speech is initiated relative to the presumed target, was used here, rather than an acoustic operationalisation of TTT as the pure duration of silence and overlap at turn switches. Cases where the responder aims to initiate their response at another Transition Relevant Place (TRP) than the end of the question are not always straightforward to identify, however, particularly because it requires speculating about the speaker’s intentions. This study only considers another point than the end of speech activity by the question as the turn-timing target, when the questioner asks two questions in a row, and the response aligns with the first and not with the second question. A response aligns with a question when it has lexical or propositional content in common with the question. To illustrate, in example 3), the response relates to the first question rather than the second. In example 4) however, it is not clear to which question the response relates.

3) Q: Are you going out tonight? Did you see your parents yesterday by the way?

R: Yes, I’m going to see a movie.

[schematic example]

4) Q: Did you go out yesterday? Want an apple?

R: No.

[schematic example]

In other cases, a TRP may appear before the end of the questioner’s speech activity as well, but there is no clear contrast in alignment of the response to the first TRP or the second. For example, in 5), in the absence of disambiguating prosody, the responder may have intended to launch their response before “or”, as “Did you like that movie?” is potentially a complete question, but they might also interpret *or* as a question tag identifying the end of the question.

- 5) Q: Did you like that movie, or?
R: Not really.
[schematic example]

These cases are likely to be common causes of short overlaps (up to 37%; Levinson & Torreira, 2015; see also Sacks et al., 1978) and the earlier TRP is a likely candidate for being the true target of that speaker's speech initiation. However, in the absence of a systematic way of identifying the speech initiation target independently from hypotheses about expected TTTs, in those cases the end of the questioner speech activity is considered to delimit TTT. Consequently, only when the questioner asks two questions in a row and the responder responds to the first question, the speech signals may be overlapped, but TTT is still positive.

Question Speech Rate The speech rate of the question (expressed as 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 (for gaps) or to the start of the TTT span (for overlaps), divided by the duration of that stretch of speech. This operationalisation of speech rate is used, rather than counting lexically or 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 first language speaker of Dutch fluent in English.

Vowel identification, along with duration measurements, were carried out manually using Praat (Boersma & Weenink, 2021) with 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 most vowels were voiced. Consecutive heterosyllabic vowels were separated when a change in formant structure, a dip in the amplitude envelope, and/or glottalization could be identified. Nasalized vowels or syllabic nasals/approximants were counted as a syllabic unit in this speech rate measurement. One English speaker often whispered. In those cases, changes in the formant structure, supported by auditory impressions, were used to decide on the presence of vowels. When laughter occurred during the question or response, it was included in the syllable count when it contained vowel-like components. If it did not, for example when it contained only voiceless frication, it was excluded.

Unless the question contained a pause of over 1000 ms, pauses were included in the question duration used to define speech rate. Short pauses were included because the speech rate appears to predict patterns of neural entrainment better than articulation rate (with regard to neural entrainment to theta-band fluctuations in the amplitude envelope of speech, see Kayser et al., 2015; for non-speech stimuli see also Bosker, 2017a). In overlapped question-response sequences, the final vowel that started before the overlap was considered the nucleus of the final syllable (see ‘Final Rhyme Duration’ below), so the speech rate was based on the number of syllables up to and including the syllable before it.

This operationalisation of speech rate, as realised vowels per second, was chosen over lexically defined syllable rate to ensure it corresponds to the acoustic input that the responder actually heard, when syllabic units were deleted in connected speech. This is a more direct measure of speech rate than the one used in some previous tests of the relationship between speech rate and TTT (e.g., Roberts et al. 2015, who used normalized deviations from the expected average phoneme duration). Additional support for this choice rests on the assumption that the relevant oscillatory activity that is predicted to entrain the perception-action link by coupled oscillator models is relatively directly affected by auditory input, rather than further downstream in the language processing system where it may be combined with a higher-level representation of the linguistic input. This assumption is supported indirectly by observations suggesting that neural auditory entrainment is induced even when hearing unfamiliar languages (Song, 2018).

Question Length The length of the question was defined over the same domain as speech rate (see above). Question length was expressed 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. Syllables were identified as acoustically realised vowels, using the same criteria as described above for speech rate.

Response Speech Rate The speech rate of the response was measured as the number of syllables, divided by the duration of the response in seconds. The domain of the response was identified as described in Section 4.2.3. Syllables were identified using the same criteria as described above for Question Speech Rate above.

Response Length The length of the response was measured as the number of syllables in the response. The domain of the response was identified as described in section 4.2.3. Syllables were identified using the same criteria as described above for question speech rate.

Final Rhyme Duration When there was a gap between question offset and response onset (i.e., TTT was positive), the final syllable rhyme duration was measured as the span from the final vowel onset to the offset of speech 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. These two criteria typically aligned, but priority was 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 the speech rate measurement, nasalized vowels or syllabic nasals/approximants were counted as a syllabic unit.

Question Type Questions were classified as polar or not, where polar questions are those that can be appropriately responded to with ‘yes’ or ‘no’ (*ja* or *nee* in Dutch).

Response Type Questions were classified as “Answered” or “Not answered”, according to the annotator’s judgement about the relevance of the response to the preceding question. Examples of questions from the English dataset classified as answered and not answered are shown below in 1) and 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?

R: Not very set.

2) Not answered

Q: Well [wu]d wha’s your what are you doing your M.A. in?

R: Eh MSc thank
you very much.

Language The language of each question-response sequence was included as a variable in the analysis as well. As discussed previously, the language of the question-response sequence corresponds directly to the corpus from which it was sampled. For this reason, it is not possible to 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. For example, as discussed in section 4.2.1, the interlocutors in the Dutch corpus had pre-existing familiarity with each other, but the degree of familiarity was inconsistent in the English corpus, which can affect turn-timing (Templeton et al., 2022).

4.2.5 Statistical analyses

The following analysis procedure was applied to the resulting dataset. In a preliminary analysis, the automatically extracted TTT distributions of all speech activity switches and the manually measured TTT distributions of question response sequences from two corpora were described and compared using visualisations.

As turn-timing distributions often are not normal (see Chapter 3), and Wilson and Wilson (2005) predict that they should show periodic clusters, the automatically extracted TTT distributions were explored using gaussian mixture modelling (Scrucca et al., 2016). Assuming that the components of a turn-timing distribution can be modelled using normal distributions, ten mixture models were fit to the data from each language separately, each with an increasing number of normally distributed components from 1 to 10. The mixture models were compared using the Bayesian Information Criterion (BIC). The mixture model with the lowest BIC value was considered the best fit to the data.

Within the question-response TTT distributions, the correlations between predictor variables and potential differences between languages/corpora were assessed, primarily using visualisations. The TTT distributions of the question-response sequences were compared to the automatically extracted TTT distributions using Kolmogorov–Smirnov tests, for each language separately, to determine if the automatically extracted question-response TTT distributions are a representative subset of the automatically extracted TTT data from the full conversations.

Next, the following filtering procedure was applied to the TTT datasets. Question-response sequences with questions shorter than four syllables and longer than 25 syllables – before the final syllable in the case of positive TTT, and before the interrupted syllable in the case of negative TTT – were excluded from the analyses. Question-response sequences with questions that contained a silent intra-speaker pause over 1000 ms were also excluded. One reason for both these exclusion criteria was to eliminate speech rate anomalies (such as excessively low speech rate values), that are due to very short utterances or to the presence of extended pauses. An additional reason to exclude short questions and questions that contain long pauses is that listener entrainment to heard speech takes time to build up (Luo & 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 & Ye, 2015). After excluding very long and very short question-response sequences, all sequences with TTT above or below 2.5 standard deviations from the mean, across the full remaining dataset, were also excluded.

After these exclusion criteria were applied, a model comparison procedure using mixed-effect linear regression models with maximum-likelihood estimation was used to determine the predictors of TTT in the question-response sequences. This procedure was applied to the full dataset, containing both overlapped turns and gaps (i.e., negative and positive TTT, referred to as the Full dataset), as well as separately for a Gaps and an Overlaps dataset. The Overlaps dataset contained observations where $TTT < 0$ ms and the Gaps dataset contained observations where $TTT \geq 0$ ms. A practical reason for separating gaps and overlaps was to include final rhyme duration as a predictor for Gaps, which cannot affect turn timing when it is overlapped.

For each dataset – Full, Gaps, Overlaps – the inferential analysis procedure consisted of two phases. Phase 1 addressed the main experimental research questions regarding the relationship between TTT and Question Speech Rate. For Gaps, the influence of Final Rhyme Duration and the interaction of Question Speech Rate $\times$ Final Rhyme Duration was also tested in Phase 1. In Phase 2, the final regression models from Phase 1 were used as a starting point for a further exploratory analysis of the additional contributions of the dialogue act factors Question Type and Response Type, Language, as well as the contribution of the three further timing variables, namely Question Length, Response

Length, and Response Speech Rate, and all interactions. The contribution of a quadratic term for Question Speech Rate to the model was also tested.

The contributions of all predictors and interactions were assessed using model comparisons with likelihood ratio tests (LRTs). At each step a model with and without the relevant factor was compared. A random intercept for Speaker was included in all models. More complex random effect structures were not included because more complex models did not converge. All categorical variables were sum coded.

For the hypothesis-testing Phase 1 analyses, a backward stepwise approach was used. 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 (two-way interactions) to the least complex (main effects). Effects which did not at least show a trend towards statistical significance ($p > .1$) were removed.

The 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 predictor was added in turn to the baseline model from Phase 1. At each step, of any predictors that significantly improved the model fit, the one that produced the largest reduction in the Akaike Information Criterion value (AIC; Akaike 1973) was retained in the model. Further models were then constructed to test for the additional contribution of the other exploratory predictors, including a quadratic term for Question Speech Rate, and then the two-way interactions between the significant main effects. This procedure was repeated until no significant predictors remained.

A further exploratory analysis aimed to describe the relative contribution of each predictor to the final exploratory model and assess the size of the effect of the variables of interest, Speech Rate and (for gaps) Final Rhyme Duration and Speech Rate $\times$ Final Rhyme Duration, compared to the well-established exploratory variables. However, there is no generally accepted optimal way to assess effect sizes or variable importance in complex regression models. As a measure of effect size, simple or main predictor coefficients are sometimes used, but their interpretation is affected by the predictor's scale, the presence of other predictors in the model, especially in the case of continuous predictor variables in observational studies, domain-specific predictions about the

absolute size of the effect (Winter, 2019). A higher coefficient can therefore not be interpreted straightforwardly as a greater effect size. Similarly, there is no agreement on a single best way to determine variable importance in a multiple regression model (Grömping, 2015).

To assess the importance of the variables of interest here, and particularly to assess how speech timing variables compare to discourse variables, the following heuristics are used in this chapter.

First, the significance of the contribution to the model fit of each predictor was retested after selecting a model using the iterative procedure described above, again using LRTs – 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 next section reflect the significance of the contribution of the relevant predictor at the time it was retained or dropped from the model in the model selection procedure. The LRTs reported in the tables reflect the contribution of each predictor to the final models, after the stepwise elimination of all non-significant predictors.

Second, the AIC of models with each predictor removed were compared. The AIC is a measure that described a trade-off between the goodness of fit of a model and its simplicity (Akaike, 1973). Predictors that most decrease the AIC are considered more important.

Third, partial regression coefficients were calculated for each predictor that is included in the final model. The partial regression coefficient of a particular predictor of interests represents the effect of that predictor when controlling for the effect of all other predictors in the model – i.e., not keeping them constant but accounting for their effect. It is the slope of a regression line between a) the residuals of a regression model of that predictor of interest against all other predictors in the model and b) the residuals of a regression model of the outcome variable against all variables except the predictor variable of interest (Allan, 1966). Partial regression coefficients are visualised in partial regression plots, which are often used to assess whether the assumptions of linear regression hold, and to qualitatively assess the relative contribution of each predictor of interest. The steeper the slope of a partial regression, the greater the contribution of that predictor.

Throughout this chapter, the conventional threshold for significance at $\alpha = .05$ is used, and effects under .1 are considered marginal (i.e., ‘trends’). All analyses were performed using the statistical program R (R Core Team 2022). Visualisations in this chapter were created using ggplot2 (Wickham, 2016). All linear models presented in the next section were build using the lme4 package (Bates et al. 2015), and the mixture models were fit and compared using the mclust package (Scrucca et al., 2023). The e1071 package (D. Meyer et al., 2023) was used to measure skewedness of distributions. The rPraat package (Bořil & Skarnitzl, 2016) was used to import and manipulate Praat TextGrids in R.

4.3 Results

This section presents the results of the analysis procedure described in Section 4.2.5. Sections 4.3.1 to 4.3.6 present preliminary analyses that discuss the nature of the turn-timing distributions. Sections 4.3.7 to 4.3.11 present the results of the experimental and exploratory analyses of the Full dataset, Gaps and Overlaps respectively.

4.3.1 Descriptive statistics of full corpus

In the two corpora, 5361 speech activity switches were automatically identified by comparing the on- and offset of speech as represented by pre-existing alignments. Figure 4.1a and b depict the distributions of TTTs of all automatically extracted speech-activity switches the English and Dutch corpus. They are not exactly normal: they both have relatively wide tails, a large positive skew (Dutch at .87 and English at 1.19), and the usual high peaks at around 200ms. The Dutch histogram has a second peak, at around -300 ms. Table 4.3 lists descriptive statistics of the two distributions.

Corpus	Mean (ms)	Highest bin (ms)	Median (ms)	Skew
Dutch	123.28	60.12	78.41	0.87
English	164.6	126.17	116.8	1.19

Table 4.3 Descriptive statistics of turn switch distributions in the full corpora. The central point of the highest bin is listed, with bin size = 100 observations.

To identify how many gaussian components best fit the distributions, 10 mixture models with an increasing number of gaussian components from 1 to 10 were fitted to the data and compared using BIC. For both languages, the optimal number of component distributions was 2. The parameters of the optimal models are listed in Table 4.4, along with their mixing proportion – i.e., the relative contribution of each component to explaining the data. Figure 4.1a and b illustrate the fit of these models to the data with

overlaid component distributions, scaled by their mixing proportions, on a histogram of the data for each language. Contrary to predictions by Wilson and Wilson (2005) the two components do not peak on either side of 0ms TTT. Instead, the components have similar points of highest density, but one is much wider than the other.

Language	Nr components	Means (ms)	Variance	Mixture proportions
Dutch	2	19.87 276.24	130536.3	0.60
			731698.9	0.40
English	2	80.63 475.20	147243.7	0.79
			1528455.0	0.21

Table 4.4 Parameters of the best fitting mixture model in each language.

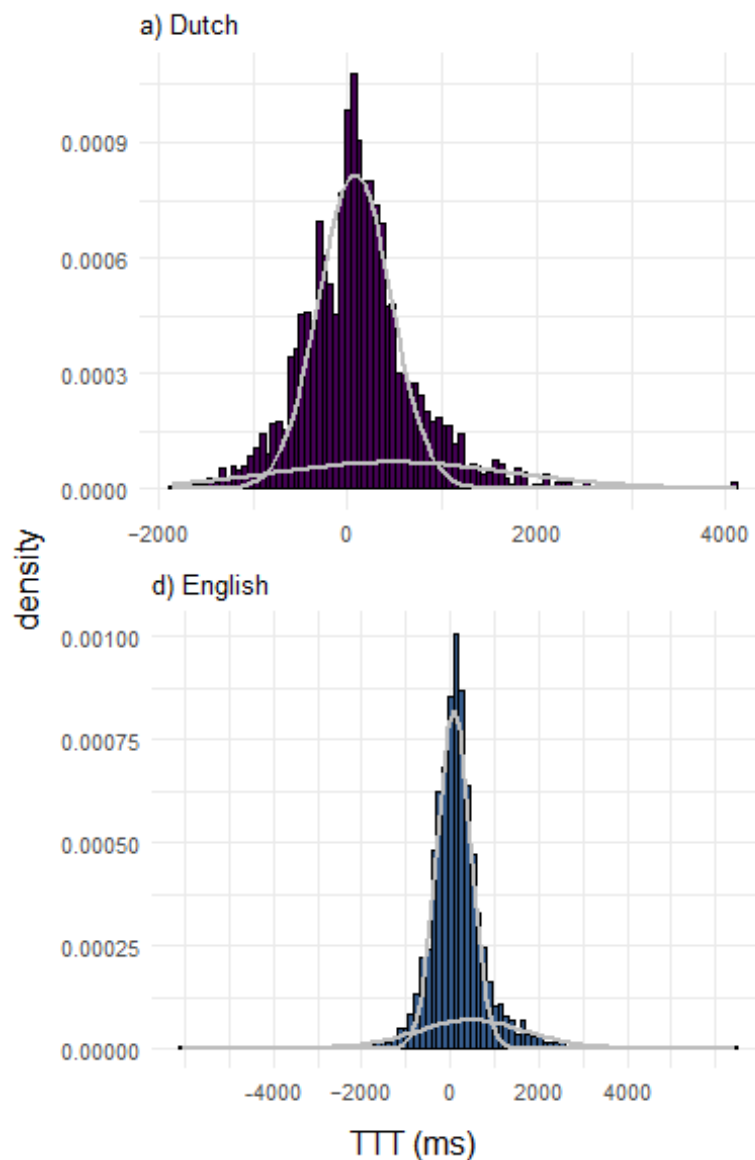


Figure 4.1 Histograms of the distributions of TTTs of all speech activity switches automatically extracted from a) the Dutch corpus, and b) the English corpus. Bin size = 100 observations.

4.3.2 Descriptive statistics of question-response sequences

In total, 568 turn switches were manually annotated as question-response sequences (10.6% of the total number of automatically extracted speech activity switches); 326 from the Dutch corpus, and 242 from the English corpus. After applying the exclusion criteria described in Section 4.2.5 to the question-response sequence sample, removing observations with 1) questions shorter than four and longer than 25 syllables, 2) questions that contain a pause over 1 second, and 3) TTT outliers, 413 observations remain. In both languages, gaps are more common than overlaps. Table 4.5 lists the number of observations in each subset of the data, by language and along the gaps/overlaps split.

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

Table 4.5 Numbers of question-response sequences observed per language and per transition type (Gaps, Overlaps and both) before applying exclusion criteria (pre) and after (post).

4.3.3 TTT distributions

TTT distributions of the question-response sequences from the Dutch and English corpora are visualised in Figure 4.1, and descriptive statistics are listed in Table 4.6 and Table 4.7 for the data both before and after applying the exclusion criteria. The exclusion criteria did not affect the measures of central tendency to a large degree, except that they decrease the means slightly and changed positive and small negatives skew to moderate negative skew (Table 4.6).

Corpus	Mean TTT (ms)		Highest bin (ms)		Median (ms)		Skew	
	post	pre	post	pre	post	pre	post	pre
Both languages	252	218	100	100	244	241	-.46	.05
Dutch	244	199	100	0	170	181	-.32	.17
English	264	242	150	350	285	291	-.67	-.14

Table 4.6 Measures of TTT central tendency per language before applying exclusion criteria (pre) and after (post). The central point of the highest bin is listed with bin size = 50.

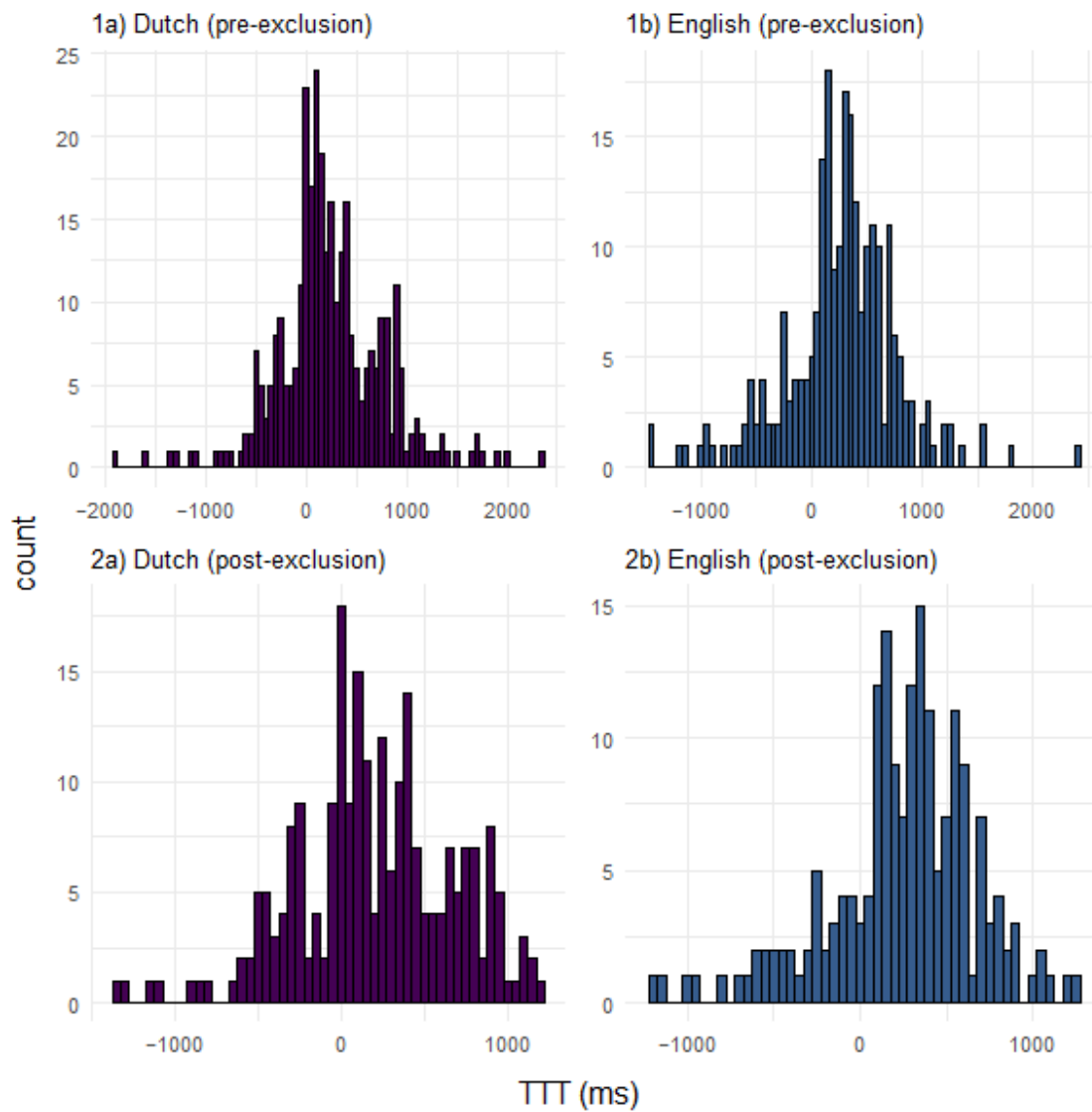


Figure 4.2 Histograms of TTT distributions in Dutch and English, before and after applying exclusion criteria. Bin size = 50 observations.

Corpus	Question		Final Rhyme	Duration (ms)	Question		Response Length	
	Speech	Rate			Length	Length	(syllables)	(syllables)
	pre	post	pre	post	pre	post	pre	post
Both languages	5.00	5.29	200	196	8.28	9.46	11.03	11.65
Dutch	5.21	5.61	190	183	8.44	9.82	10.27	11.03
English	4.72	4.89	212	212	8.07	8.99	12.06	12.44

Table 4.7 Mean speech timing variables per language before applying exclusion criteria (pre) and after (post).

The TTT distributions of the question-response sample were compared to the TTT distribution of the full corpus for each language separately using a two-sided Kolmogorov-Smirnov test, to assess whether the question-response samples can be thought of as representative subsets of the full corpus samples with regard to TTT, though note that they were extracted using different methods (i.e., by performing manual measurements rather than automatic comparisons of pre-existing utterance-level alignments), so their comparability is necessarily limited in this regard. The Dutch question-response TTT distribution was significantly different from the Dutch full conversation TTT distribution, ($D = .16, p < .001$), as was the case for the English distributions ($D = .20, p < .001$), indicating that the manually extracted question-response sequence measurements are not statistically representative of the automatically extracted TTT distributions.

4.3.4 Differences between corpora

Means for the speech timing variables are listed in Table 4.7, along with mean Question Length and mean Response Length. The mean Question Speech Rate is higher in the Dutch corpus by .72 syllables/s, $t(393.94) = 6.08, p < .001$. Accordingly, the mean Final Rhyme Duration is shorter in the Dutch corpus by 394.95 ms, $t(323.47) = -2.15, p = .032$, and the mean Response Speech Rate higher at .56 syllables/s, $t(409.11) = 4.16, p < .001$. In short, the Dutch sample overall had a higher speech rate. The mean differences between Question- and Response Length in the two language corpora are not significant, Question Length: $t(405.99) = 1.8, p = .073$; Response Length: $t(407.34) = -1.52, p = .13$.

Question Type and Response Type counts per language corpus are listed in Table 4.8 and Table 4.9. The Dutch and English corpora do not significantly differ in how many responses are answers, $\chi^2(1) = 3.04, p = .08$, but they do significantly differ in how many

	post			pre		
	non-polar questions	polar questions	total	non-polar questions	polar questions	total
Dutch	53	177	230	80	246	326
English	61	122	183	90	152	242
total	114	299	413	170	398	568

Table 4.8 Proportion of polar and non-polar questions per language before applying exclusion criteria (pre) and after (post).

questions are polar questions, $\chi^2(1) = 10$, $p = .002$; the proportion of polar questions is higher in the Dutch corpus (75.5%) than in the English corpus (62.8%).

	post			pre		
	answer responses	not-answer responses	total	answer responses	not-answer responses	total
Dutch	48	182	230	55	271	326
English	17	166	183	22	220	242
total	65	348	413	77	491	568

Table 4.9 Proportion of answer and non-answer responses per language before applying exclusion criteria (pre) and after (post).

4.3.5 Correlations between predictor variables

Figure 4.3 illustrates correlations between the continuous speech timing variables in the full dataset as well as for gaps and overlaps separately. All correlations are weak; the

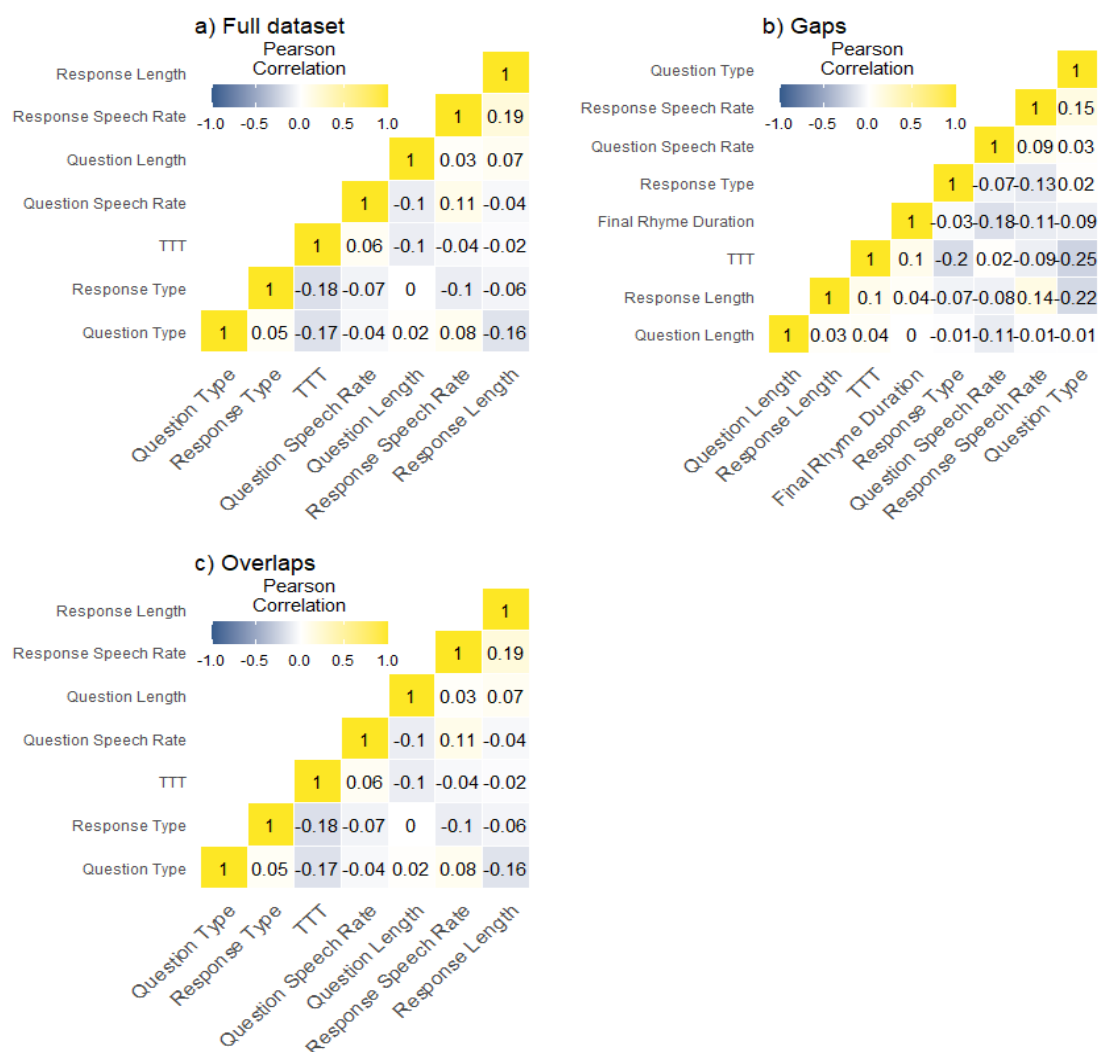


Figure 4.3 Correlation heatmap displaying Pearson correlation between all variables, after applying exclusion criteria, for a) All, b) Gaps and c) Overlaps.

strongest correlation in the full dataset is between Response Length and Response Speech Rate at $r = .189$, such that longer responses are uttered faster, in line with reports that longer utterances tend to be faster (Quené, 2008; Tanner et al., 2024). In the Gaps dataset, the strongest correlation is between Question Type and TTT at $r = -.25$, such that polar questions are associated with shorter TTT. In the Overlaps dataset, the strongest correlation is also between Response Length and Response Speech Rate at $r = .302$. As for questions, longer responses are uttered faster.

4.3.6 Number of questions per speaker and conversation

Before applying exclusion criteria, the mean number of question-response sequences per speaker is 16.71, ranging from 2 to 42. Per conversation, the mean number of question-response sequences is 33.41, ranging from 10 to 70. Speaker pairs appear to approximately match the number of questions they ask; they both ask relatively few or relatively many questions, as illustrated in Figure 4.4, and the biggest difference between two speakers is 18 questions. The number of questions asked by two speakers in a

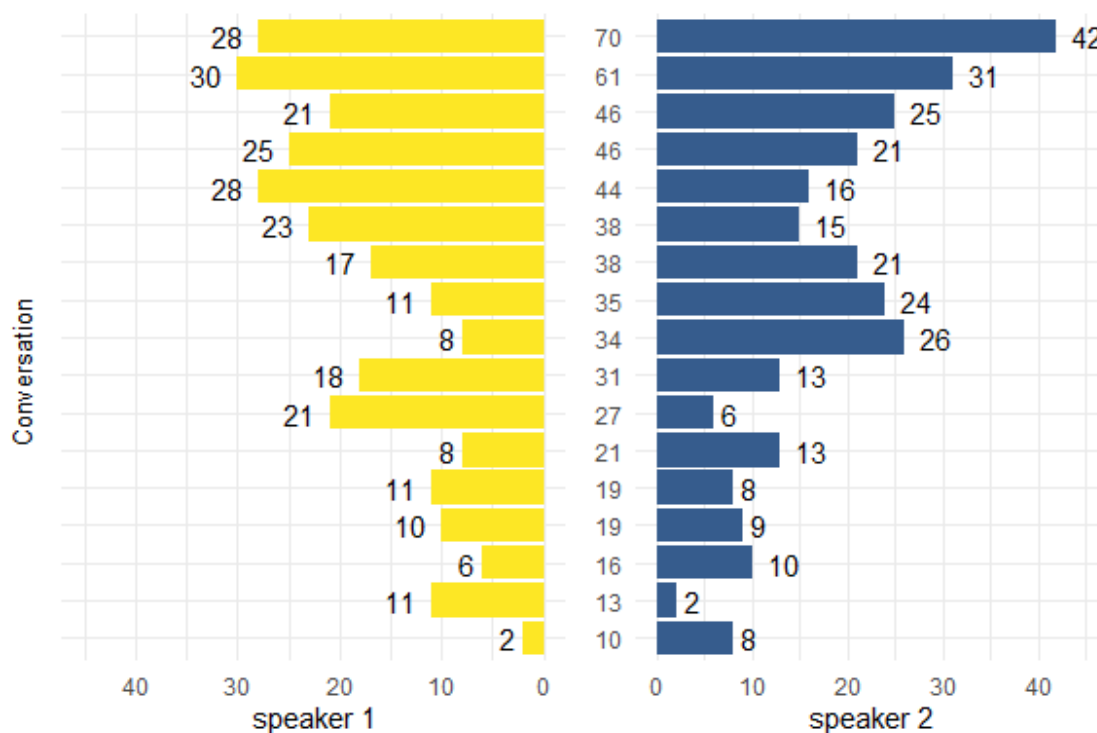


Figure 4.4 Number of questions asked by each of the two speakers in each conversation, before applying exclusion criteria. Each row corresponds to a conversation. Numbers on the middle axis indicate the number of questions asked in that conversation in total, while the bars and numbers on each side indicate the number of questions asked by each speaker. The conversations are ordered by number of questions asked in total, from most (top) to least (bottom). Note that the speakers in each conversation were randomly assigned to be speaker 1 or speaker 2.

conversation is moderately correlated, Spearman's $\rho = .48$, $p = .0497$. This correlation is not significant after applying the exclusion criteria outlined in Section 4.2.4, Spearman's $\rho = .38$, $p = .138$. When normalizing the number of questions asked by each speaker by the total number of turns by that speaker, Spearman's $\rho = .48$, $p = .054$, there is a trend to the same effect.

The next few sections describe the results of the experimental and exploratory analysis that primarily aimed to test the perception-action entrainment hypothesis that speech rate has a significant negative association with TTT.

4.3.7 Hypothesis testing – Full dataset

As described in Section 4.2.5, a linear regression model predicting TTT from Question Speech Rate was first compared to an intercept only model (both with a random intercept for Speaker) using an LRT to test the significance of the contribution of Question Speech Rate to the model fit. Question Speech Rate significantly contributed to the model fit, $\chi^2(1) = 3.91$, $p = .048$. Contrary to the predictions stated in Section 2.3.3, it was positively associated with TTT, i.e., faster questions were associated with longer TTT.

4.3.8 Exploratory analysis – Full dataset

The regression model with a main effect for Question Speech Rate and random intercept for Speaker was then used as the starting point of an exploratory analysis, testing the effects of Question Type, Response Type, Language, Question Length, Response Length, Response Speech Rate, and a quadratic term for Question Speech Rate. The predictors were added in a forward stepwise manner, where at each step the significant variable that most decreases or least increased the AIC of the model was added, and then all other predictors were retested until no more predictors were significant. Subsequently all two-way interactions between each of the significant predictors were tested.

In step 1, Question Type and Response Type both significantly contributed to the model fit, and Response Type most decreased the model's AIC (by 8.93 versus 7.72), so Response Type was first added to the model, $\chi^2(1) = 10.93$, $p = .001$. In step 2, Question Type was also added, $\chi^2(1) = 9.47$, $p = .002$. Responses that did not answer the question were timed later on average than those that did (difference in mean = 227 ms), and polar questions were responded to earlier on average than non-polar questions (difference in

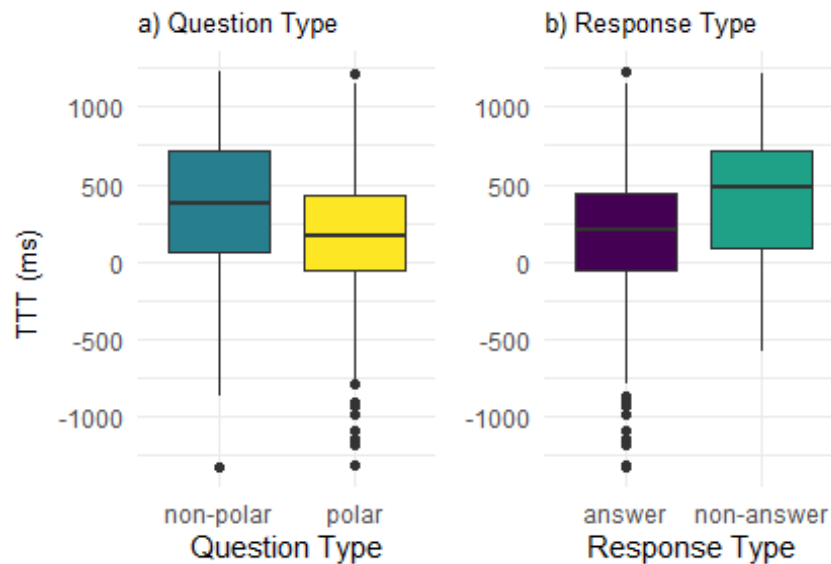


Figure 4.5 Boxplots illustrating the difference in TTT between a) polar and non-polar questions and b) answer responses (answers) and non-answer responses (non-answers).

mean = 172 ms). The effects of Question Type and Response Type are shown in Figure 4.5.

Finally, in step 3, Question Length was marginally significant, $\chi^2(1) = 3, p = .083$, i.e., there was a trend towards longer questions being responded to earlier than shorter questions. Response Length, Response Speech Rate, the quadratic term for Question Speech Rate, and Language were not significant, so they were not included in the model.

Next, all two-way interactions between the main effects (Question Speech Rate, and the exploratory variables Question type, Response Type and Question Length) were tested. Only the interaction between Question Speech Rate and Question Type significantly contributed to the model fit, $\chi^2(1) = 7.14, p = .008$. Faster polar questions were associated with longer TTT, while faster non-polar questions were associated with shorter TTT.

The final linear regression model for the full dataset includes the predictors Question Speech Rate, Question Type, Response Type, Question Length (marginal), and the Question Type $\times$ Question Speech Rate interaction. The final model for the Full dataset is summarised in Table 4.10. After adding the exploratory predictors, all predictors were retested by comparing the final model to a model with each of these predictors removed in turn. Table 4.11 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, as well as partial regression coefficients for each predictor. Figure 4.6 displays partial regression plots for each main effect.

Predictors	Estimate	Std. Error
Intercept	401.01	126.02
Question Speech Rate	-.04	19.99
Response Type	-101.95	29.67
Question Type	-353.59	106.05
Question Length	-8.45	4.58
Question Speech Rate × Question Type	51.72	19.25

Table 4.10 Model estimates for the final model, full dataset. Categorical variables were sum coded such that polar = 1, non-polar = -1, answered = 1, not-answered = -1, so their coefficients represent main effects.

Model	AIC	AIC Difference with Final Model	χ^2	p-value	Description of effect	Partial regression coefficient
Final	6203.97	-	-	-	-	-
- Question Speech Rate	6201.97	-2	0	.998	not significant	14.05
- Response Type	6213.45	9.49	11.49	.001	Answered responses are associated with shorter TTT	-95.67
- Question Type	6212.93	8.96	10.96	.001	Polar questions are associated with shorter TTT	-357.34
- Question Length	6205.33	1.36	3.36	.067	Longer questions are associated with shorter TTT	-7.73
- Question Type × Question Speech Rate	6209.1	5.14	7.14	.008	Polar questions: faster questions are associated with longer TTT Non-polar questions: faster questions are associated with shorter TTT	51.36

Table 4.11 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.

Note that the effect of Question Speech Rate is no longer significant when retested, $\chi^2(1) = .00, p = .998$. Also, Response Type most reduced the AIC, closely followed by Question Type, whereas the slope of the partial regression plot for Question Type is steeper than the plot for Response Type. Question Type and Response Type both have steeper partial regression coefficients and reduce the AIC more than Question Speech Rate, suggesting they explain more variance.

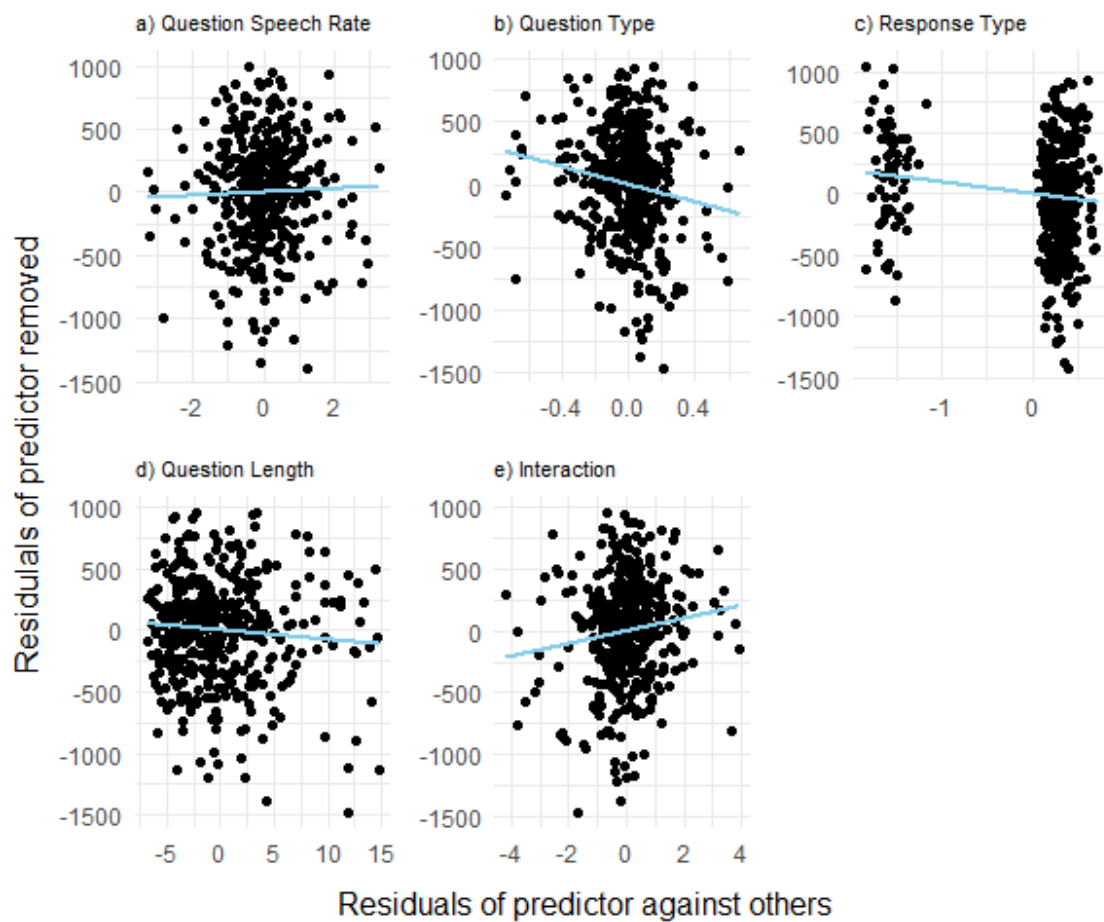


Figure 4.6 Partial regression plots for each predictor in the final model for the Full dataset: a) Question Speech Rate, b) Question type, c) Response Type, d) Question Length, e) Question Type $\times$ Question Speech Rate.

4.3.9 Hypothesis testing – Gaps

In Phase 1 for the Gaps dataset, Question Speech Rate, Final Rhyme Duration, and their interaction, were tested first to assess the contribution of each predictor to the model fit. As discussed in Section 4.3.5, Question Speech Rate and Final Rhyme Duration are weakly but significantly correlated in the Gaps dataset, as expected, Pearson's $r = -.18$, $t(302) = -3.26, p = .001$. However, a linear regression model predicting TTT from these two

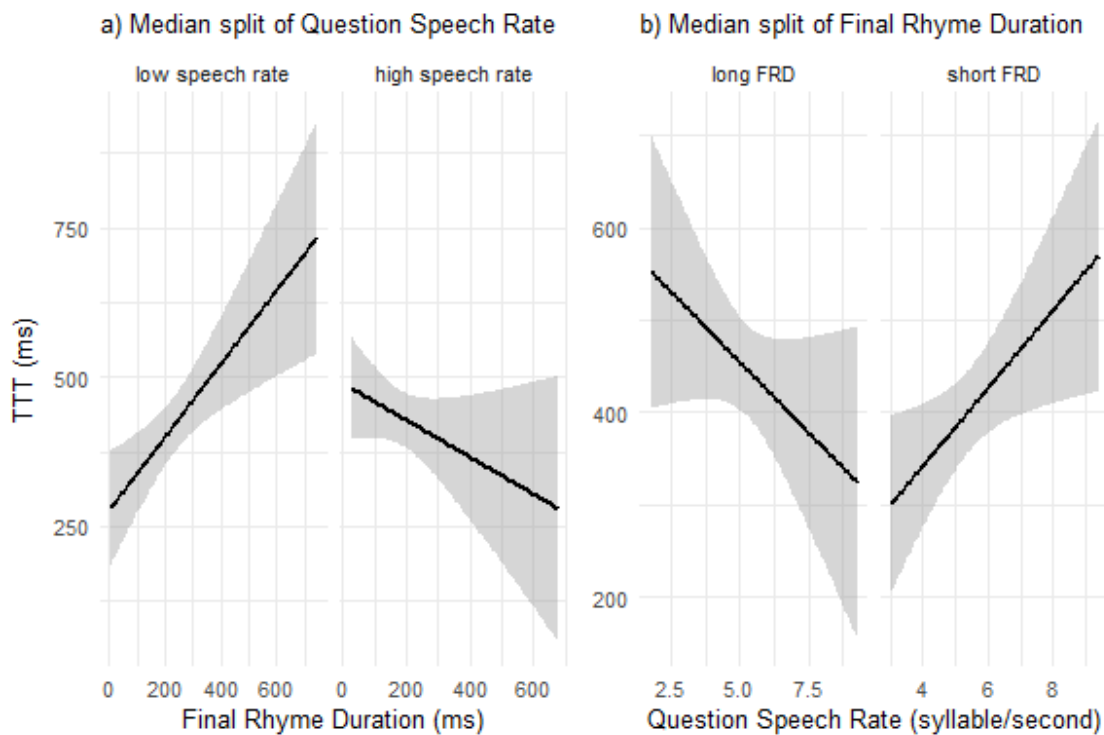


Figure 4.7 Interaction between Speech Rate and Final Rhyme Duration by median split of Question Speech Rate (a) and median split of Final Rhyme Duration (b). The statistical interaction derives from continuous data and the median splits are depicted to illustrate

predictors has a low Variance Inflation Factor, so no adjustments were made (VIF: 1.02; a model with the interaction term included was not assessed because interaction terms necessarily increase the VIF).

In step 1 of Phase 1, the interaction between Question Speech Rate and Final Rhyme Duration made a marginally significant contribution to the model fit, $\chi^2(1) = 3.42, p = .064$, so it is retained. The main effect of Question Speech Rate is significant, $\chi^2(1) = 3.97, p = .046$. As for the full dataset, the main effect of Question Speech Rate is positive, meaning that higher speech rates are associated with longer TTT. The main effect of Final Rhyme Duration is also significant, $\chi^2(1) = 4.84, p = .028$. Overall, longer final rhymes are associated with longer TTT.

Figure 4.7 illustrates the interaction between Question Speech Rate and Final Rhyme Duration with median split plots for both timing predictors – although note that the interaction effects reported here were based on continuous data. From the perspective of a Question Speech Rate median split (figure 4.8 a), 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 4.8b, 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 are associated with shorter TTT. This effect is discussed further in Section 4.4.3.

4.3.10 Exploratory analysis - Gaps

In Phase 2, the model predicting TTT from Question Speech Rate, Final Rhyme Duration and their interaction was used as a baseline. The contribution of the exploratory variables Response Speech Rate, Question Type, Response Type, Question Length, Response length, Language and the quadratic term for Question Speech Rate were tested in a forward stepwise manner. Question Type was added to the baseline model in step, $\chi^2(1) = 16.83, p < .001$. As in the full dataset, polar questions are responded to more quickly than non-polar questions (mean difference = 163.93 ms) as illustrated in Figure 4.8.

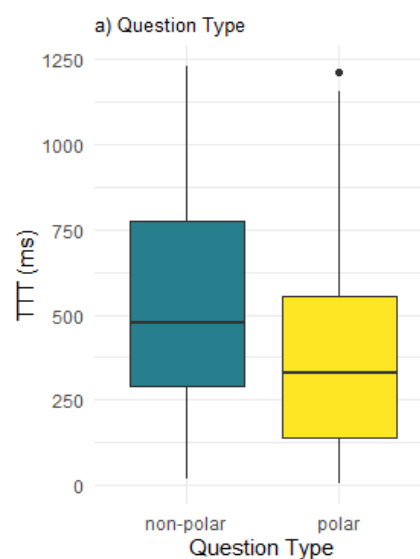


Figure 4.8 Boxplot illustrating the difference in TTT between polar and non-polar questions.

Adding Response Type to the model led to a singular fit, indicating potential overfitting, so it was not included. Response Speech Rate, Question Length, Response Length, Language and the quadratic term for Question Speech Rate did not contribute significantly to model fit at any stage, so they were not included either. Further LRTs indicated no significant interactions between Question Type and any of the other significant main effects.

The final model for the Gaps dataset therefore included Question Speech Rate, Final Rhyme Duration, Question Speech Rate × Final Rhyme Duration, and Question Type. It is summarised in Table 4.12. Each of the predictors was retested in a backwards manner, by comparing the final model against a model with that predictor removed. As indicated

Predictors	Estimate	Std. Error
Intercept	188.88	146.79
Question Speech Rate	43.83	27.01
Final Rhyme Duration	.98	.57
Question Type	-75.39	18.04
Question Speech Rate × Final Rhyme Duration	-.16	.11

Table 4.12 Model estimates for the final model, Gaps dataset. Categorical variables were sum coded such that polar = 1, non-polar = -1, answered = 1, not-answered = -1, so their coefficients represent main effects.

Model	AIC	AIC Difference with Final Model	χ^2	p-value	Description of effect	Partial regression coefficient
Final	4306.82					
- Question Speech Rate	4307.42	.6	2.6	.107	not significant	46.49
- Final Rhyme Duration	4307.7	.88	2.88	.09	Longer final rhymes are associated with shorter TTT	.96
- Question Type	4321.65	14.83	16.83	<.001	Polar questions are associated with shorter TTT	-73.17
- Final Rhyme Duration × Question Speech Rate	4306.76	-.06	1.94	0.164	not significant	-.15

Table 4.13 LRTs comparing the final model for the Gaps dataset and models with each of the predictors removed.

in Table 4.13 (previous page), the Phase 1 predictors no longer significantly improved the final model fit compared to models with that predictor removed (Question Speech Rate: $\chi^2(1) = 2.6$, $p = .107$, Final Rhyme Duration: $\chi^2(1) = 2.88$, $p = .107$, Question Speech Rate $\times$ Final Rhyme Duration: $\chi^2(1) = 1.94$, $p = .164$. Question Type most decreases the model AIC, and its partial regression plot has the steepest slope.

In Figure 4.9, partial regression plots are presented for each of the predictors included in the final model for gaps. Notably, the partial regression plots for a) Question Rate, c) Final Rhyme Duration, and d) the Question rate $\times$ Final Rhyme Duration interaction, all show an outlier. Since these are partial regression plots, it is an outlier in terms of how well *derived* models – so not the final model used to test relationships between the predictor variables and TTT, but models used to calculate partial regression coefficients – predict their respective outcome variable. Partial regression plots are sometimes used to identify such outliers that decrease the fit of the model to build models that make more accurate

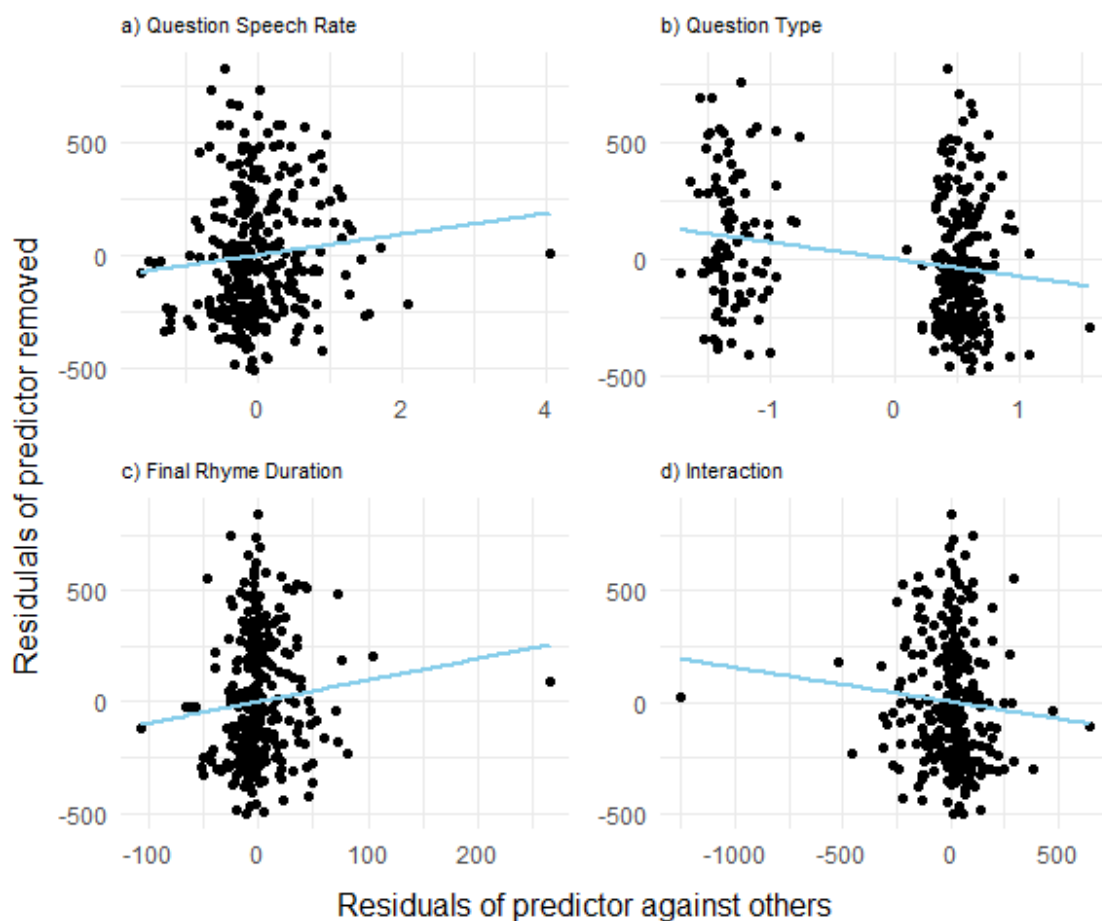


Figure 4.9 Partial regression plots for each predictor in the final model for the Gaps dataset: a) Question Speech Rate, b) Final Rhyme Duration, c) Question type, d) Question Speech Rate $\times$ Final Rhyme Duration.

predictions. However, since the purpose here is hypothesis testing, not building predictive models, doing so would essentially bias the dataset in favour of the hypothesis, so the outlier is not removed.

4.3.11 Hypothesis testing and exploratory analysis – Overlaps

In the overlaps dataset, there is no main effect of Question Speech Rate, $\chi^2(1) = .77, p = .381$. None of the exploratory variables significantly improve the model fit compared to an intercept-only model (Response Speech rate: $\chi^2(1) = .05, p = .831$, Question Type: $\chi^2(1) = .03, p = .865$, Response Type: $\chi^2(1) = .67, p = .412$, Response Length: $\chi^2(1) = 1.22, p = .269$, Language: $\chi^2(1) = .6, p = .437$ and the quadratic term for Question Speech Rate: $\chi^2(1) = .81, p = .367$). There is only a marginal effect of Question Length, $\chi^2(1) = 3.59, p = .058$, which also slightly decreased the model AIC (1557.65 compared to 1559.24). This effect indicates a trend towards longer questions being responded to earlier, as in the full dataset. In short, there is no strong evidence that any of these predictors particularly affects TTT. The final model for Overlaps, including only Question Length and a random intercept for Speaker, is summarized in Table 4.14

Predictors	Estimate	Std. Error
Intercept	-249.43	64.05
Question Length	-10.30	5.38

Table 4.14 Model estimates for the final model, Overlaps dataset.

4.4 Discussion and Conclusion

The main aim of this study was to test the perception-action entrainment hypothesis that higher speech rates are associated with shorter turn-transition time (TTT), specifically in spontaneous question-response sequences. This negative relationship is predicted by coupled oscillator models of turn timing by Wilson and Wilson (2005) and Garrod and Pickering (2015). It is supported by the results of an experimental study by Corps et al. (2020), who used a question-responding paradigm, but has only weak, circumscribed support from a corpus study by Beňuš (2009) and is contradicted by a descriptive observation in the corpus study by Roberts et al. (2015). The main finding of this study was in line with those of Roberts et al. (2015). In the subsets of data where there was any relationship between speech rate and TTT, faster speech in questions was by and large associated with longer TTT before responses.

This study also tested the effect on TTT of another speech timing factor, final syllable rhyme duration and how final rhyme duration interacts with speech rate. Additional exploratory analyses tested the contribution of two dialogue act variables—the type of question (polar or not polar) and the type of response (answered or not answered)—as well as the length of the question and of the response (in syllables), the speech rate of the response, and the language spoken (Dutch or English). Finally, this study inquired whether the relationship between speech rate and TTT is best understood as linear or quadratic, testing whether very slow as well as very fast speech rates are associated with slower TTT.

As the next sections review, the impact of including of some of the additional dialogue factors suggests that the influence of speech rate variation on TTT is subsidiary to the influence of dialogue act-related effects.

4.4.1 *Speech rate as predictor of TTT*

Question Speech Rate had a positive relationship with TTT both in the Full dataset, which includes overlaps and gaps, and in the Gaps dataset. In both cases, faster speech in questions was associated with longer TTT before responses. 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 they found support for the hypothesised relationship only in selected cases. This finding is 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. It also aligns with the observation by Torreira and Bögels (2022) that rate positively predicted response time in their experiment where speakers responded to meaningless strings of syllables when the string contained cues similar to turn-ending cues. Additionally, there was no significant effect of Response Speech Rate in any of the subsets of the data, also contrary to perception-action entrainment hypothesis.

This negative effect of the question's speech rate on TTT indicates that listener entrainment to speech rate in a turn is not the primary determiner of their turn timing in spontaneous conversation. It lends support to Levinson and Torreira's (2015) interpretation of Roberts et al.'s (2015) findings that faster questions have a higher processing cost for the listener and therefore delay their own turn. An alternative

explanation is that turn ending cues are recognised earlier in slower turns, allowing the responder more time to prepare and initiate theirs (Torreira & Bögels, 2022).

A third possibility is that the speech rate effect in this data is itself an artifact of another phenomenon, namely that not just final lengthening (Brusco et al., 2020; Duncan, 1972; Gravano & Hirschberg, 2011), but also more gradual slowing down, is a turn-ending cue (Rühlemann & Gries, 2020). If so, the questions that were relatively slow may have primarily been slow on average because they slowed down, allowing the responder more advance notice of the turn end. In general, however, this effect of speech rate should not be over-interpreted, since the influence of Question Speech Rate on TTT is no longer significant when the dialogue act factors are also considered as predictors (see below).

4.4.2 *Final syllable rhyme duration as predictor of TTT*

There was also an effect of Final Rhyme Duration in the Gaps dataset: longer final rhymes were associated with longer TTT. This effect is contrary to the overall effect of speech rate, and contrary to the expectation that a relatively short final rhyme compared to the phase of the oscillator system should be associated with relatively long TTTs because the oscillator system has not yet reset its phase in response to it. It also contradicts a number of findings by previous studies. First, Corps et al. (2020) found, when response times were measured from final word offset, similarly to how TTT was measured in this study, that a relatively short final (monosyllabic) word was associated with longer response times. Second, Torreira and Bögels (2022; see also Barthel, 2020) concluded that TTT should be shorter if turn-taking cues—such as final lengthening—can be recognized earlier. Finally, it is inconsistent with a recent finding by Franich (2023), who observed that final word duration was negatively related to TTT.

We consider several interpretations of this effect. First, it is possible that the phase of the hypothesized entrained coupled oscillator system is reset very locally, specifically in response to the final syllable, and that this effect largely overrides the frequency that the coupled oscillator system has at the start of the final syllable. Corps et al. (2020) suggest this explanation for their observation that response times were faster after speeded final words if the response time was measured from the start of the final word. In that case, longer final rhymes being associated with longer TTT would be in line with coupled oscillator models, although the coupled oscillator system would then not be robust to local variation in speech rate, as Wilson and Wilson (2005) speculated it should be.

However, there is evidence that the onsets of syllables (or other entraining auditory stimuli), not their durations, are responsible for phase resetting (Bosker, 2017a; Oganian & Chang, 2019). If so, the lengthening of the final syllable rhyme occurs too late to reset the 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 interaction between speech rate and final rhyme duration in the Gaps dataset, as discussed in Section 4.4.3.

4.4.3 Question Speech Rate $\times$ Final Rhyme Duration interaction and TTT

There was a marginal interaction between Question Speech Rate and Final Rhyme Duration in the Gaps dataset. As illustrated in Figure 4.7, this interaction was negative: 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. Although this interaction only made a marginal contribution to the model fit, and should be interpreted with caution, this section discusses a number of possible, potentially complementary, interpretations.

First, one aspect of the interaction offers at least superficial support for the coupled oscillator models (Wilson and Wilson 2005; Garrod and Pickering 2015), particularly when visualized in the median-split graphs for the speech timing predictors in Figure 4.7. Specifically, the relationship between Question Speech Rate and TTT depicted in Figure 4.7b (left panel) is negative, meaning that a faster speech rate is associated with a shorter TTT, as predicted. However, the median split is for illustration and does not reflect a qualitative split in the data; the interaction itself arises between two continuous predictors. Moreover, it is not clear what might cause speech rate entrainment to affect TTT specifically only in this subset of the data. As this negative Question Speech Rate $\sim$ TTT relationship is contrary to the overall positive influence of Question Speech Rate on TTT, it is more parsimonious to assume that something other than intermittent entrainment causes this pattern.

An alternative view of the interaction is that a greater contrast between Final Rhyme Duration and the duration of preceding syllables promotes shorter TTT; specifically, faster speech followed by longer final rhymes make final lengthening more salient. As Section 2.4 discussed, speech-rate normalisation can impact the interpretation of local durational variation, such that segments are perceived as longer when the preceding rate

is faster (Baese-Berk et al., 2014; Morrill et al., 2015; Reinisch et al., 2011; White et al., 2022). Such effects have been associated with predictive timing mechanisms, which may be underpinned by entrainment to the foregoing flow of syllables (Baese-Berk et al., 2014; White et al., 2022). This interpretation is consistent with the view that turn-timing behaviour is the result of a combination of speech processing mechanisms, including predictive and reactive behaviours, rather than as a direct result of perception-motor entrainment, in line with the suggestion by Torreira and Bögels (2022) that perception-action entrainment is overshadowed in more complex contexts. It is also worth noting that where final rhymes are relatively short (Figure 4.7b, right panel), TTT increases with rate. In other words, 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 these considerations is that the crossover interaction can be seen as being underpinned by a change in the relationship between Question Speech Rate and TTT, depending on Final Rhyme Duration, and as a change in the relationship between Final Rhyme Duration and TTT, depending on Question Speech Rate. Importantly, however, the interaction was only marginally significant and should not be over-interpreted. Ideally this effect should be replicated, with evidence from both experimental and corpus studies.

4.4.4 Question Length, Response Length and TTT

In the Full dataset, longer questions (in terms of number of syllables) were associated with shorter TTT, and there was a similar trend ($p = .058$) in the Overlaps dataset. These effects are in line with question length effects found by Roberts et al. (2015). There was no effect Question Length in the Gaps dataset. With regards to the Full dataset and Overlaps dataset, this effect may indicate that transition-relevant information, such as prosodic and syntactic turn-ending cues (Torreira and Bögels, 2022) can occur earlier in longer questions, which may facilitate the listeners' ability both to predict turn endings and to plan their responses prior to the transition. However, it is not clear why the same effect was not detectable in the Gaps dataset.

Additionally, there was no effect of Response Length in any of the analyses, so this study provided no support for previous experimental observations that longer responses are timed later (Meyer et al., 2018). A possible reason is that response length is less clearly definable in spontaneous conversation than experimental paradigms, since responders

can continue to plan their unfolding response while already speaking. Roberts et al. (2015) also found that Response Length was an important variable in explaining TTT in their corpus study of spontaneous conversations, but they suggest that Response Length was a stand-in for dialogue act, which are more constrained in this study.

4.4.5 Question Type, Response Type and TTT

In the Full dataset and the Gaps dataset, polar questions were responded to earlier than non-polar questions. This is unsurprising, given that TTT increases with response complexity (Roberts et al. 2015) and polar questions typically solicit responses that are less complex than open questions because the possible responses are, at least pragmatically, a closed class (Knudsen et al., 2020). Additionally, in the Full dataset, responses that were judged to respond directly to the question are associated with shorter TTT than non-answers, in line with findings by Stivers et al. (2009).

The dialogue act variables appear to be more important predictors of TTT than the speech timing variables Question Speech Rate and Final Rhyme Duration, judging by the heuristics used in this study to assess predictor importance. When they are included in the models, they explained more TTT variance than Question Speech Rate or Final Rhyme Duration (see AIC drops and partial regression coefficients in Table 4.11 and Table 4.13 and plots in Figure 4.6 and Figure 4.9). The significant Question Speech Rate effects also disappear when Question Type, Response Type, Question Length, and the Question Type $\times$ Question Speech Rate interaction are added to the regression model for the Full dataset. Similarly, neither Question Speech Rate nor Final Rhyme Duration are retested as significant TTT predictors in the Gaps dataset when Question Type is added.

This change does not appear to happen because Question Type covertly encodes variations in Question Speech Rate or Final Rhyme Duration; polar questions and non-polar questions do not significantly differ in Question Speech Rate, $t(204.75) = .83$, $p = .407$ (mean difference = .1 syllables/s), or in Final Rhyme Duration, $t(214.24) = .89$, $p = .376$ (mean difference = 12ms). Likewise, answered questions do not differ from non-answered questions in Question Speech Rate, $t(90.15) = -1.36$, $p = .178$ (mean difference = .2 syllables/s), nor in Final Rhyme Duration, $t(112.76) = -.5$, $p = .621$ (mean difference = 7ms).

To conclude, while speech timing factors are predictive of TTT in some circumstances, the effect of Question Speech Rate is contrary to the predictions of the coupled oscillator models (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005). Also, the influence of speech timing appears unimportant compared to dialogue act.

4.4.6 *Question Speech Rate × Question Type interaction*

In the exploratory analysis of the full dataset, there was an interaction between Question Speech Rate and Question Type. For polar questions, faster speech was associated with later responses, while for non-polar questions faster speech was associated with earlier responses. This is not in line 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 (see also Franich, 2023), and contrary to our expectation that spontaneous polar question-response sequences are most likely to pattern with experimental paradigms.

One possible reason might be that the typical information structure of the two types of questions in Dutch and English, and/or the discourse context associated with polar and non-polar question interact differently with the effects of speech rate on processing demands discussed above. Specifically, we might expect that responses to polar questions might be easier to plan, and therefore more likely to be initiated relative to the question's speech rate, in line with the suggestion by Torreira and Bögels (2022) that there is an entrainment mechanism that contributes to turn timing, but it is overshadowed by other behavioural effects in complex contexts. However, the idea that polar questions provide a less complex context in spontaneous conversation is an assumption – they might less frequently be responded to with turn-initial filled pauses than non-polar questions, which are even easier to plan and time, for example. A qualitative analysis of the question-responses sequences that addresses the way responses to open questions are initiated could provide some initial insight into this issue.

Also note that, since polar questions constitute the larger proportion of the full dataset observations (299 polar and 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.

4.4.7 *Language and TTT*

Language (Dutch or English) did not predict TTT in any of the analyses presented here. 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). As the two languages differed in mean Question Speech Rates, any variance related to language itself may have been in part explained by Question Speech Rate as well. A variation in TTT could, however, arise from many factors, including dialogue task, conversational context, social relationships of interlocutors, and measurement and annotation methods. Although Chapter 3 found no consistent pattern related to any of these factors, these factors were relatively consistent in our data. A larger study of similarly balanced corpora across a wider sample of languages could determine the extent to which language and speech rate predictors stand in for each other.

4.4.8 *Observational vs. experimental studies*

The contrast between our finding, and the findings of experimental studies and studies with more constrained interactional contexts (Corps et al., 2020; Franich, 2023; Vranjes & Oben, 2022) supports the speculation that perception-action entrainment is a possible mechanism for speech initiation, but is not active or determinative in spontaneous conversational contexts. Experimental contexts may evoke this mechanism due to specific demand characteristics, which may essentially induce the same behaviour as a speeded response task. Along these lines, Corps et al. note 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) that affect TTT, but it is also possible that there are strategic approaches to turn-taking, engaged by speakers at different times to different degrees (Heldner and Edlund 2010; Campbell 2008). Torreira and Bögels (2022) suggest an idea along these lines, to explain their observation that their participants' response time when initiating a meaningless syllable was affected by the rate of the prime, but only when no other turn-ending-like cues were present. Torreira and Bögels speculate that there is an entrainment-based speech initiation mechanism, but that it is

overshadowed in spontaneous conversation by other speech planning and articulation mechanisms.

Under experimental conditions such as a question-response paradigm, 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, creating a context like in Torreira and Bögels’s (2022) no-turn-ending-cues condition. This task-oriented strategy may promote overall speech rate entrainment more than spontaneous conversation. In other words, an underlying entrainment mechanism may be engaged to a larger degree in experimental studies.

4.4.9 *Timing of gaps vs. overlaps*

A practical reason for distinguishing gaps and overlaps in this analysis was that 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: it is not always simply the case that the listener’s predictions are wrong and therefor lead to accidental interruption. Particularly, longer overlaps may correspond to competition for the conversational floor. In such cases, we might predict weaker or absent relationships between (negative) TTTs and the speech timing of questions. Along these lines, there was no effect of either Question Speech Rate or Response Speech Rate on TTT for the Overlaps dataset. 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 (assuming that speakers aim to start speaking at the end of the previous turn) about turn endings from those that are pragmatically judged as interruptions (cf. Beňuš, 2009).

4.4.10 *Turn-timing distributions*

This study also examined the turn-timing distributions of both the automatically extracted speech-activity switches in the full corpora, and the manually measured turn switches in question-response sequences. Using mixture modelling, the optimal number of component normal distributions of automatically extracted distributions was two in both Dutch and English. Although the Dutch distribution gave an impression of bimodality, the component distributions in neither language suggested that the data can best be modelled as bimodal (see Figure 4.2). Instead, the data best fits a combination of a normal distribution with small variance and one with high variance

It is not straightforward to interpret what the shape of the turn-timing distributions presented here (specifically in the full corpora, which are very similar to many turn-timing distributions presented in the literature, see Chapter 3), tells us about the behaviour that produced those distributions. One possibility is to assume that turn-timing behaviour generally results in normally distributed observations, but that there are two classes of turn-timing behaviour that differ in variance. Difference between these two classes might be the presence of TRP ambiguity, which meant that the annotator could not determine what TRP the responder targeted their response to in a subset of cases, or another source of measurement error. Other possibilities could be the effects of competition for the floor, ambiguous turn-ending cues, the use of meaningful pauses before turns, or some turns being ‘mistimed’ (see below and the next chapter). Finally, a difference in behavioural strategy – such as the response-oriented and non-response oriented discussed above – could distinguish different subsets of turn-timing distributions. A qualitative analysis of the question-response sequences could help to distinguish these cases. However, note that this interpretation rests on the assumption that turn-timing should result in normally distributed measurements.

Section 4.3.3 also tested whether the question-response sequences can be thought of as representative of the full corpus. Stivers et al. (2009) similarly tested whether their manually extracted turn-timing measurements of question-response sequences were representative by comparing their distribution to an automatically extracted corpus of turn-timing in telephone conversations in Dutch. From a non-significant t-test, they concluded that their question-response corpus was representative of turn-timing in general. As the turn-timing distributions in the present study are not normally distributed, this study presented non-parametric tests of goodness of fit (Kolmogorov-Smirnov tests) and found that the question-response sequences in Dutch and English respectively are not representative of turn-timing in general. This difference between question-response sequences and the full corpora may in part be because the former was manually measured, while the latter was automatically extracted, but it may also be due to the difference in the dialogue acts included. These differences suggest that findings regarding question-response sequences cannot be generalized to other types of turns without caution.

4.4.11 Conclusion

The main purpose of this study was to test whether there is a relationship between the speech rate of questions and the timing of responses in spontaneous conversation. Contrary to predictions by the coupled oscillator models proposed by Wilson and Wilson (2005) and Garrod and Pickering (2015), there was an overall positive association between the speech rate of the question and TTT: higher speech rate in questions was associated with later responses. No relationship was observed in overlaps, and the speech rate of the response also did not significantly relate to TTT, contrary to the coupled oscillator models.

For sequences with gaps, there was a marginal interaction between the speech rate of the question and the duration of the final syllable rhyme: when final syllable rhymes were relatively long compared to the foregoing syllables, TTTs were shorter. This effect might reflect a variation in the salience of local timing cues to utterance boundaries but should not be overinterpreted. In the gaps dataset, there was also an interaction between speech rate and the type of question asked; most of the questions in these corpora were polar (yes-no questions), for which higher speech rate was associated with longer TTT. This pattern may be due to the higher processing demands associated with responding to relatively fast questions. In non-polar questions, there was a negative relationship between speech rate and TTT as predicted by coupled oscillator models.

These results suggest that many aspects of question-response sequences affect turn-transition timing, with only a small role for foregoing speech rate, which is moreover generally contrary to the predictions of the coupled oscillator models. The aspects that affected turn-timing in this study include discourse-related features, but interpersonal factors, processing mechanisms and variations in behavioural strategies are also likely to affect it. In short, the results of this study align more closely with the prediction-reaction models of turn timing, than with the coupled oscillator models.

One aspect of the timing of responses that this chapter did not account for is variation that is due to speakers diverging from ‘normal’ turn-timing behaviour, either because they mistimed their turns or intended to convey meaning through their timing. Operationalising the degree to which turns are ‘well-timed’ may shed light on the features of turn switches that determine the appropriate time to respond (Beňuš, 2009),

and further clarify the relationship between speech timing and turn-timing. The next chapter addresses this issue by presenting a rating study of turn timeliness.

Chapter 5. Speech Rate and Turn-Timing Perception: A Rating Study of Perceived Response Timeliness

5.1 Introduction

Not all turns are initiated ‘on time’; variations in turn timing are associated with impressions of disfluency (Van Os et al., 2020) and delayed turn timing can carry pragmatic meaning (Bögels et al., 2020; F. Roberts et al., 2011; F. Roberts & Francis, 2013). The previous chapter found an overall *positive* association between the speech rate of questions and the turn timing of responses in spontaneous conversation, contrary to the perception-action entrainment hypothesis by coupled oscillator models of turn timing (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005) – faster questions are responded to slower, not faster as expected. However, we did not account for meaningful variations in timeliness. This chapter presents two studies that do so by operationalising the ‘timeliness’ of turn-taking as a perceptual variable. We collected listeners’ ratings of the timing of responses to questions. The question-response stimuli are spontaneous, unmanipulated fragments of dialogues, taken from corpora of conversations in British English and Dutch respectively.

The resulting expanded corpus data were then used to test whether the relationship between speech rate and turn timing is modulated by the perceived timeliness of the turn. As in the previous chapter, we focus on the effects of the speech rate and the final rhyme duration of questions on the timing of responses. In addition, we conducted a secondary analysis to investigate whether turn timing is perceived relative to the speech rate of the foregoing speech, as expected from observations that listeners’ perceptions of durational phenomena in speech are subject to speech-rate normalisation (Diehl et al., 1980).

The remainder of Section 5.1 first briefly reviews the motivation for this study (Section 5.1.1). Section 5.1.2 states the research questions addressed in this study and discusses its methodology.

5.1.1 Motivation

Turn timing production is often presented as ‘precise’, and psycholinguistic models of turn timing aim to account for that precision. However, turn timing is also variable, as turn timing distributions have wide spreads (cf. Heldner & Edlund, 2010) and are affected by

many factors (cf. S. G. Roberts et al., 2015). Some of that variation in turn timing is salient to listeners and interlocutors, as dysfluency (Van Os et al., 2020), or as conveying specific pragmatic meaning, such as reluctance (F. Roberts & Francis, 2013).

Psycholinguistic models of turn timing might therefore aim to account for ‘typical’ turn timing, against which saliently different cases are interpreted. For example, Levinson and Torreira present their model as accounting specifically for typical cases of turn timing (Levinson & Torreira, 2015, p. 14, “The model therefore captures the most typical turn transition values observed in conversational corpora.”). Similarly, when coupled oscillator models predict a negative relationship between turn timing and speech rate, we might expect that this relationship holds for turns that are ‘on time’ and is systematically modulated by the degree to which turns are perceived as earlier or later.

For these reasons, it is useful to operationalise turn ‘timeliness’ (cf. Beňuš, 2009) independently from the predictions of the model, to distinguish between turns that are ‘on time’ and turns that are early or late. One way to do that is through listeners’ interpretations of timing. Some studies have collected listeners’ judgements of a construct hypothesised to relate to turn timing, such as turn timing fluency (Van Os et al., 2020), naturalness (Itoh et al., 2009), or a more specific pragmatic interpretations (F. Roberts & Francis, 2013). However, these judgements also depend on other aspects of the context (cf. Itoh et al., 2009; F. Roberts et al., 2011; Van Os et al., 2020). Although the aforementioned studies controlled for those aspects experimentally, that is not an option when operationalising timeliness of spontaneous data is the goal, as in the present study.

Another approach is to explicitly ask for listeners’ judgements of the timeliness of turns. Stivers et al. (2009) took this approach when they asked annotators to judge whether responses to polar questions were on time or delayed. They found that the threshold for delay appeared to differ in ten languages. This approach is more likely to avoid confounding impressions of fluency or other constructs. However, by asking few annotators rather than a large participant pool, the thresholds they observed may also be due to annotator idiosyncrasies rather than intrinsic to the language or data. To generalise those observations would require the assumption that a single speaker can adequately access a shared knowledge of appropriate timing.

In the present study, we expanded the method by Stivers et al. (2009) by explicitly asking for participants' judgements of turn timeliness in turn sequences from corpora of spontaneous conversations. Our study asked for judgements on a more fine-grained scale (7 points from early to late rather than a binary choice) and included a larger group of participants (N=144).

We then used the rating data to create a 'timeliness' variable associated with the spontaneous corpus data. This study's primary aim is to test whether the relationship between speech rate and turn timing is different for timely turns than for early or late turns. As in the previous chapter, we consider both the global speech rate of the question (in syllables per second) and the duration of the final syllable rhyme as a local rate measure.

In addition, we conducted a secondary analysis to explore based on what features of the turn switches participants formed their interpretation of timeliness. We expected to see an effect of speech rate normalisation on turn timing perception: speech rate normalisation refers to the phenomenon that durational distinctions in speech are interpreted relative to the surrounding speech rate. For example, the threshold for interpreting a phoneme with a particular voice-onset time as voiced is earlier when the context speech rate is fast than when it is slow (Diehl et al., 1980).

Speech rate normalisation affects perception of discrete as well as continuous contrasts in meaning (cf. White et al., 2022), it is also found in the relation between speech by two different speakers (cf. Bosker, 2017b; Newman & Sawusch, 2009), and it holds in relation to spontaneous speech-rate variation (cf. Nagao & De Jong, 2007). As turn timing is also a continuous contrast in meaning that is interpreted in spontaneous speech between multiple speakers, we hypothesized that turn timing perception is also subject to speech rate normalisation. In addition, we consider the effect of the length of the question and of the response on participants' ratings.

5.1.2 Research questions and methodology

To summarise, this chapter addresses the following research questions:

3. What is the relationship between speech rate, turn timing, and perceived turn 'timeliness'?

Experimental research question:

3.1 Is the relationship between speech rate and turn-transition timing in question-response sequences modulated by participant ratings of turn timeliness? In other words, is there a consistent relationship between speech rate and turn timing specifically when turns are perceived as ‘on time’?

Secondary research question:

3.2 Do speech timing features (particularly speech rate) of the question (and/or responses) affect participants’ ratings of response timing?

In a preliminary analysis, we also explored whether participants considered the timing of responses to be a salient and variable feature of the spontaneous question-response sequences and whether ratings were consistent between participants.

To address these questions, this chapter expands on Stivers et al.’s (2009) approach of relying on listener perceptions of timing to supplement corpus measurements by collecting more fine-grained ratings – on a 7-point scale rather than binary – from a much larger sample of listeners – 45-50 participants per stimulus rather than one annotator, who are not informed about the study purpose. The participants rated stimuli from spontaneous conversations in which, unlike in previous perception studies (F. Roberts et al., 2011; F. Roberts & Francis, 2013; Van Os et al., 2020), the features of interest were not manipulated.

Like all studies discussed in this section, the studies presented in this chapter make use of an overhearer paradigm: the stimuli consist of a subset of the question-response sequences analysed in the previous chapter, rather than conversational turn switches in which the participants were or are themselves involved. The overhearer paradigm potentially introduces the issue that participants’ ratings diverge from interlocutors’ online judgements of timeliness to an unknown extent, as various aspects of the conversational context are inaccessible to the rating study participants – such as visual cues, or a common ground shared between the original interlocutors. Conversely, the rating study participants are artificially required to listen to the whole response before rating its timing, while interlocutors have the continuous opportunity to act on their judgements of timeliness before they know the responder’s complete response. However, asking for judgements from the speakers themselves can likewise introduce

issues. For example, doing so during the conversation alters the conversational context, while doing so afterwards essentially turns the speakers themselves into overhearers whose retrospective access to the relevant aspects of the conversational context cannot be controlled. For these reasons, the studies in this chapter follow previous work by opting for an overhearer design with many participants.

Also following previous studies, the studies presented here ask participants for an off-line rating to operationalise timeliness, rather than an implicit measure based on reaction time. This choice was made because the responses in the stimuli varied in length, and the purpose of this chapter was to observe participants' ratings of timeliness in stimuli that were as close to the original spontaneous behaviour as possible, without manipulating the length of the responses.

Unlike many previous studies, participants are asked to rate turn-timing directly, rather than a construct that is hypothesised to be related to turn-timing, such as perceived fluency, naturalness, or enthusiasm. This approach risks directing the participants' attention to an interpretation of the turn switch that they may not explicitly pay attention to in spontaneous conversation, namely its timeliness, but it avoids accessing intuitions about other aspects of the turn switches that are part of constructs such as fluency, including within-speaker pausing behaviour, repetitions, or intonation.

Stivers et al.'s (2009) approach implies that turn timing can be evaluated relative to a shared expectation of timing by a competent speaker of the language. Our approach does not eliminate that assumption, but in a preliminary analysis of questionnaire responses we investigate whether there is a degree of agreement among participants in their ratings, and whether the notion of timeliness was meaningful to them. If so, the resulting rating will be used as an intersubjective measure of timeliness to address the questions whether perceived timeliness modulates the relationship between speech rate and turn-timing (research question 3.1).

Finally, this chapter deploys a subset of the spontaneous question-response sequences in Dutch and English as stimuli, to supplement the corpus study from the previous chapter, and to study effects on the perceived timeliness of responses in these two different languages.

In the rest of this chapter, Section 5.2 lays out the study methods and Section 5.3 presents the results. Sections 5.3.1 and 5.3.2 discuss a preliminary analysis to assess the validity of this design using responses to a debriefing questionnaire and mostly descriptive analyses of the rating data. Sections 5.3.3 and 5.3.4 test whether the relationship between speech timing features of the question-response sequence and TTT is modulated by participants' ratings of response timeliness. Sections 5.3.5 and 5.3.6 address the question whether participants' perception of turn-timing is subject to speech rate normalisation.

5.2 Methods

We conducted two rating studies, one with stimuli in British English and speakers of British English as study participants, and subsequently one involving stimuli in Dutch and speakers of Dutch as study participants. Both studies followed the same procedure, aside from the language of presentation, the corpora used to collect stimuli, and to some extent the way in which stimuli were selected from the corpora. The next sections describe the design of both studies, indicating where elements differed between the two languages.

5.2.1 Participants

All participants (total $N = 144$) provided their informed consent to participate in the study, submit their data for analysis and allow an anonymised version of their data to be published. Ethical approval for this study was granted by the Newcastle University Ethics Committee (reference: 9401/2020) and the Newcastle University School of Education, Communication and Language Sciences Ethics Committee.

5.2.2 English participants

100 speakers of English completed the rating task in this study. Of the initial 100, six were excluded, leaving a total of 94 participants (mean age = 22.04 years, $SD = 6.84$ years, range = 18-51 years, 83 female, 8 male, 3 other).

The participants were recruited through the Newcastle University Speech and Language Sciences undergraduate participant pool, the Newcastle University Speech and Language Sciences paid participant pool, the Newcastle University Psychology undergraduate participant pool and through a sign-up form distributed in department newsletters and student social media at Newcastle University. The form screened

participants for whether they were 18 years old or older and first language speakers of English. Participants recruited through the Newcastle University student participant pools could not be screened for whether they were first-language speakers of English, so they were screened through a demographic questionnaire presented at the end of the study (see Section 5.2.4). Student participants received course credits for their participation according to their study programme's requirements. Non-student participants received a £6 voucher.

Of the 100 participants, one was excluded because they indicated on a debriefing questionnaire that they were not first-language speakers of English – i.e., they responded 'no' to the question "Do you consider yourself a native speaker of English?". For a subset of participants who were recruited through the Newcastle Speech and Language Science students (see below) that question failed to display. For that set, the two participants who indicated they were not native speakers of SSBE and did say they spoke another language, without specifying that they started learning it as adults, were excluded. Two participants were excluded because they did not pass the attention check. One participant was excluded because they reported having a listening processing issue for which they normally require support.

5.2.3 Dutch participants

51 first-language speakers of Dutch completed the rating task in this study. Two additional participants took part but did not complete the whole study and left before completing the questionnaire. Their data were excluded. Of the initial 51, one was excluded, leaving 50 participants (mean age = 25.28 years, SD = 16.12 years, range = 18 to 66 years, 29 female, 20 male, 1 other). One participant was excluded because they did not pass the attention test.

The participants were recruited through the participant recruitment platform Prolific and through a sign-up form distributed in department newsletters and communications from the department secretary at higher education institutions in the Netherlands, as well as through social media. The form screened participants for whether they were 18 years old or older and first-language speakers of Dutch. Participants received a €7 voucher or €7 for their participation.

5.2.4 Procedure

All participants took part in the study online. The study stimuli were presented using Labvanced presentation software (Finger et al., 2017). The study was presented on a grey background using black text, as depicted in Figure 5.1.

After clicking the study link, participants were presented with the following tasks. They first read the participant information sheet and consent form and were asked to tick a box to indicate their consent before moving on. The information sheet and consent form are reproduced in Appendix A6. Next, they completed a headphone test to confirm that they were using working headphones. The headphone test consisted of a calibration test, where participants pressed a button and checked if they heard a sound, and a standard headphone task from the Labvanced open library (Finger et al., 2017) in which participants heard three tones at different volumes and indicated which tone was the

A) English stimuli presentation screen

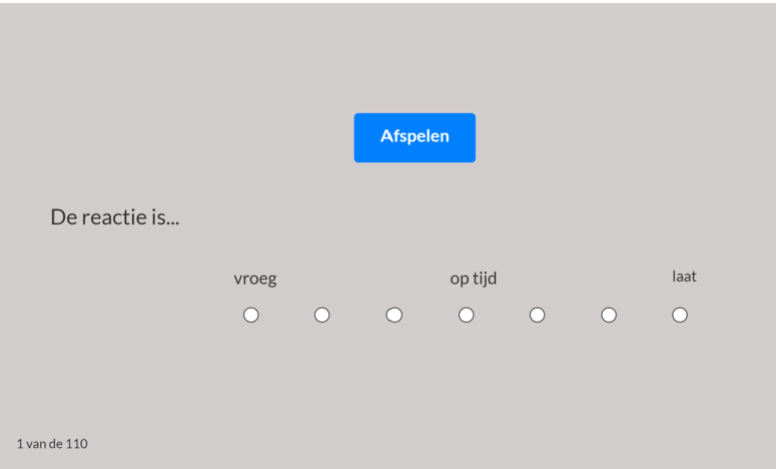


The reaction is...

early on time late

1 of 103

B) Dutch stimuli presentation screen



De reactie is...

vroeg op tijd laat

1 van de 110

Figure 5.1 Stimulus presentation screen in English (A) and Dutch (B).

quietest, for a total of 12 times. If they made more than one mistake, they failed the headphone test and were asked to retry the study after checking their hardware.

After the headphone tests, participants moved on to the main task, which was a self-paced rating task. In each trial, participants heard a stimulus that consisted of a question-response sequence taken from spontaneous conversations (see Sections 5.2.5 to 5.2.84 for details). They were asked to rate the timing of the response to the question on a 7-point scale from ‘Early’ (left) to ‘On Time’ (middle) to ‘Late’ (right). Only the leftmost, rightmost, and middle points were labelled on the screen (see Figure 5.1). At the start of each trial an empty grey screen was shown for 500 ms, followed by a screen showing a ‘play’ button, the rating scale, a counter showing their progress, and a ‘continue’ button. Participants could only enter their response after clicking ‘play’; they could only listen to each stimulus once, and they could only move on to the next trial after listening to the whole stimulus and recording their response.

In addition to the experimental trials, there were 8 attention-check trials. Attention-check trials looked identical to the experimental trials, but instead of a question-response sequence, the ‘play’ button played a recording of the sentence “please press W” in English and “*druk spatie om door te gaan*” (press space to continue) in Dutch. Participants were instructed beforehand to press the W-key or the space key instead of the ‘next trial’ button to move on in those trials, to ensure that they were listening to the stimuli. The ‘w’ and ‘space’ keys were chosen because they are commonly referred to with terms that are easy to distinguish from other keys. Data from participants who failed to recognise one or more attention check trials were excluded from the analysis.

Before the start of the main task, there were four practice trials, one of which was an example attention check trial. Participants could try the practice trials as many times as they wanted. Instructions (reproduced in Appendix A7) were displayed on the screen before the practice trials and again before the main study.

After completing the rating tasks, participants completed a demographic and a debriefing questionnaire. The main aim of the demographic questionnaire was to identify participants who do not meet the study criteria – for example, due to having hearing difficulties or not being a first-language speaker of the target language. For the Dutch participants, the debriefing questionnaire was shortened, because participants of the

English study had previously indicated that they found the questionnaire too long and felt unmotivated towards the end. No comprehensive analysis of the questionnaire responses is presented here, but Sections 5.3.1 and 5.3.2 consider responses to the debriefing question “To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?”.

5.2.5 Stimuli

The stimuli were audio fragments of a conversation that contained question-response sequences taken from the free conversation section of the Leverhulme corpus (White et al., 2012) for English and the IFADV corpus (Van Son et al., 2008) for Dutch. The stimuli were taken from the sample of question-response sequences analysed in Chapter 4, which was selected using the procedure described in Section 4.3, briefly repeated in the following paragraphs.

An exhaustive set of question-response sequences was identified from the conversations according to their syntactic and/or prosodic structure, or the question’s apparent response-evoking purpose within the context of the conversation. Each sequence was annotated for a set of variables described further in Section 5.2.8, including the length of the question and of the response in number of syllables. A subset was then selected according to, in part, the same criteria as outlined in Section 4.2: only question-response sequences with questions longer than five syllables and shorter than 26 syllables were included. Note that the analyses in Chapter 4 additionally excluded all sequences with a TTT above or below 2.5 times the standard deviation from the mean from analysis, calculated from the pooled Dutch and English datasets, but this last exclusion step was not applied here.

The stimuli were then prepared for presentation using Praat (Boersma & Weenink, 2021). First, the audio fragments containing the included question-response sequences were cut from the full conversation recording from the corpora using a time-window of two seconds before the start of the question to 500ms after the end of the response. This margin was used to ensure that participants had some time to focus their attention on the stimulus at the start of each trial, and that the stimulus audio did not end abruptly. This second consideration was based on an observation made during pilot testing that stimuli with responses that sounded incomplete (e.g., because the speaker was interrupted) distracted from the target turn switch.

Each resulting fragment was amplitude-normalized using the `normalize` function from the Praat Vocal Toolkit (Corrette, 2012). Additionally, the 500 ms offset of each fragment was attenuated using the `fade` function from the Praat Vocal toolkit, which decreased the amplitude of the speech gradually to zero. Some stimuli were additionally edited for clarity – although never in a way that affected the TTT or speech rate – as described in Sections 5.2.6 and 5.2.7 for English and Dutch separately.

Of the question-response sequences that were not included in the stimuli set, three were chosen for each language to serve as practice stimuli and processed according to the same procedure as described above.

5.2.6 *English stimuli*

The English stimuli were chosen from the Leverhulme corpus (White et al., 2012). All speakers in the corpus were first-language speakers of English and had Standard Southern British English accents.

Of the 242 question-response sequences identified in the corpus, 190 were included as stimuli in the study. Four of the resulting set of 190 stimuli were additionally edited to increase the salience of the turn transition. One was edited because a previous question and response were audible in the two-second onset window; here, the first question and response were removed, and some microphone silence from later in the fragment was added to the first section. Two were edited because the response was also a question, and the response to that was overlapped and still audible in the fragment; in this case, the final 500ms offset fade-out was removed and the final 500ms of the response was faded. Finally, in one fragment, the question was relatively short, and the response was very long. To reduce the possibility that participants were distracted from the target turn switch during the long response, the end of the response was removed at the clearest syntactic and prosodic boundary within the response, and a 500ms fade-out was added to the end of resulting fragment.

Because the number of stimuli was high, participants were assigned to one of two participant groups. Each group rated only half of the full set of stimuli, to keep the total runtime of each participant's rating session under 30 minutes so that they were less likely to become fatigued and lose focus. Each stimulus was randomly assigned to one of the two participant groups, but no more than 75% of stimuli taken from each of the eight

conversations was assigned to one group – except for one conversation (indexed as fr-s11s12 in the corpus), from which only one question-response sequence met the inclusion criteria described above. Within each study session, stimuli were presented in random order. Randomisation was done by the experimental presentation software Labvanced built-in randomisation setting (Finger et al., 2017). Participants who took part through Sona (see below) were randomly assigned to one of the two participant groups by Labvanced. Participants who were recruited through the sign-up form were assigned alternatingly to group one or two based on the order in which they signed up.

Of the 94 participants whose data are included, 45 took part in group one and 49 took part in group two. This discrepancy is due to unequal participant retention between groups one and two.

5.2.7 Dutch stimuli

The stimuli for the Dutch study were question-response sequences selected from conversations DVA3E to DVA12S from the IFDVA corpus (Van Son et al., 2008). Initially, 244 question-response sequences were selected, using the same criteria as outlined above, and processed according to the same procedure. However, pilot testing by the researcher and two participants unaware of the purpose of the study indicated that in many of the stimuli, the target turn switch to be rated was not salient, making the overall task confusing. This was the case, for example, when the question was overlapped by frequent backchannels that could be confused for an interruptive response.

As most participants of the English study did not report finding their task confusing (only two out of the 100 participants indicated that they were sometimes unsure to what aspect of the stimulus to pay attention), it is unclear what caused this discrepancy, but it may be due to differences in the style of speaking between the corpora; the Dutch speakers sounded more informal and were more familiar with each other (see Section 4.2 for a discussion). Interlocutors who were more familiar with each other may have used more idiosyncratic speech that is more difficult to understand for outsiders, or more overlapping backchannels. As pointed out in Section 4.3, the Dutch question-response sequences also had an overall faster speech rate.

To avoid confusing participants, a subset of these stimuli was selected in the following way: One pilot participant volunteered to categorize the stimuli into ‘clear’ and ‘unclear’.

They were asked to determine, for each stimulus, if it was clear to them to which turn switch to pay attention. The following three selection criteria were established by identifying commonalities between the stimuli categorised as ‘unclear’:

- 1) Stimuli were excluded from the original set of 244 if there was more than one turn switch audible before the start of the question, which could not be removed because they overlapped with the question. If the additional turn switch could be removed, it would be replaced with microphone silence to make sure the target question still always starts 2 seconds after the start of the stimulus.
- 2) Stimuli were also excluded if there was any overlap of speech other than the target question or response with the target question or answer, including backchannels.
- 3) Finally, stimuli were excluded if there was any utterance audible after the response that could not be removed because it overlapped with the target response.

If the additional turn switches – utterances by the responding speaker before the question, or utterances by the questioning speaker after the response – could be removed without affecting the question or response, they were replaced with microphone silence from elsewhere in the same conversation.

Of the 244 initially selected stimuli, 102 were selected after pilot testing. Of those, 46 were edited according to the principles described above. Because the total time required to listen to all stimuli was shorter in the Dutch study than in the English study (approximately 20 minutes in total, compared to 40 minutes in the English study), the Dutch participants were not assigned to two participant groups, but all participants rated all stimuli.

For stimuli that were edited, the time-window of 2 seconds before and 500 ms after the question-response sequence was preserved, and the principle that the question-response sequences themselves must be as they occurred in the corpus, was applied to increase comparability with the English study. However, in addition to the fact that the corpus was recorded for different purposes and with different procedures and methods - see Section 4.2 for a discussion – this more stringent selection of stimuli does limit comparability of the two studies, because the selections of stimuli represent the corpora to different degrees.

5.2.8 Variables

Because the stimuli were selected from the question-response sequences analysed in Chapter 4, the speech-timing related variables of interest in this study had already been extracted. The variables TTT, Question Speech Rate, and Final Rhyme Duration were used to address research questions 3.1 and 3.2. The remaining variables Question Length and Response Length were used in further exploratory analyses that aimed to test whether other timing related features of the stimuli affected participant ratings. All speech-based variables were calculated from manual measurements in Praat (Boersma & Weenink, 2021) using default spectrogram settings (a view range of 0–5kHz, window length of 5ms, dynamic range of 70dB with a visual inspection window between 450 and 600ms). All annotations were performed by one annotator (the author of this thesis). Section 4.2.4 described in more detail how each variable was measured, and those used in this study are listed in Table 5.1 (next page), along with summary definitions for each variable.

5.2.9 Analysis

We first performed preliminary analyses to assess the rating task’s validity. The consistency between participants’ ratings was assessed using visualisations of the raw ratings and the standard deviation of ratings per stimulus. If the timeliness of turns was salient to participants, and they shared their expectations of timing to a large degree, we would expect a small standard deviation of ratings per stimulus. The effect of presentation order on ratings was also assessed using visualisations. We also present a qualitative discussion of participants’ responses to the debriefing question “To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?”.

From the raw ratings, we then constructed an aggregate variable, Mean Standardised Rating, which is used to represent the ‘timeliness’ of the response in each stimulus. Mean Standardised Rating is calculated as the mean of ratings per stimulus, standardised per participant. The distribution of Mean Standardised Rating and its relationship to TTT and Question Speech Rate are discussed using visualisations.

Experimental research question

3.1 Is the relationship between speech rate and turn-transition timing in question-response sequences modulated by participant ratings?

Variable	Definition
Turn-Transition Time (TTT)	The duration of the interval between the end of the question responded to, and the start of the response (in milliseconds).
Question Speech Rate	Speech rate of the question in syllables per second, derived from the number of acoustically realized vowels between the start of the question to the start of the final-syllable rhyme (for gaps) or the start of the final vowel that starts before the overlap (for overlaps), divided by the duration of that stretch of speech in seconds.
Response Speech Rate	Speech rate of the response in syllables per second, derived from the number of acoustically realized vowels in the response, divided by the duration of the response in seconds.
Question Length	The number of syllables in the question, up to but not including the final syllable for gaps, and for overlaps, the final syllable whose vowel started before the overlap.
Response Length	The number of syllables in the response.
Final Rhyme Duration (for gaps)	Time from the onset of the final vowel to the offset of speech.

Table 5.1 Variables associated with each question-response sequence.

Mean Standardised Rating was used to split the question-response sequences into three categories, ‘Early’, ‘Timely’, and ‘Late’, corresponding to the lowest, middle and highest third of Mean Standardised Ratings. Multiple linear regression models were built for each subset of the data separately, predicting TTT from Question Speech Rate. The significance of the contribution of Question Speech Rate to the model fit compared to an intercept-only model was assessed using analyses of variance (ANOVA), with $\alpha = .05$ as the conventional threshold for significance. For the subset of question-response sequences that had positive TTT – i.e., Gaps, we also tested Final Rhyme Duration and the interaction between Question Speech Rate and Final Rhyme Duration. Predictors with $p \leq .1$ and $p > .05$ were considered marginally significant and retained in the model.

A random intercept for Speaker (i.e., the speaker who uttered the question in the question-response sequence) was initially included in these models but led to a singular fit with many of the models. This singular fit may be due to a sparsity of data in some levels of the Speaker variable, since for some of the Dutch speakers, only one question-response sequence was included in the stimuli set. For this reason, the models in this analysis were built without a random effect structure.

In Chapter 4, the data were filtered to exclude extreme TTT values (± 2.5 times the SD). However, if Mean Standardised Rating adequately captures listeners' impressions of timeliness, we might expect that very high TTT values would tend to be associated with high Mean Standardised Rating values, and very low TTT values would tend to be associated with low Mean Standardised Rating values. Mean Standardised Rating should therefore account for differences in the relationship between Question Speech Rate and TTT across extreme TTT values and more central values. For this reason, the data were not filtered in this analysis.

An analysis of Mean Standardised Rating as a continuous variable was also conducted but did not produce interpretable results. A summary of this analysis is included in Appendix A8 for completeness.

Secondary research question

3.2 Do speech timing features of the question affect participants' ratings of response lateness?

In a secondary analysis, we explored what features were predictive of participants' un-averaged ratings of timeliness. First, to test whether participants' ratings or turn timeliness were affected by speech rate normalisation, a backward linear regression model comparison procedure was applied using LRTs to test the significance of the contribution of Speech Rate, Final Rhyme Duration, TTT and their interaction to a model predicting (non-standardised and non-averaged) Ratings. TTT was included because participants are expected to pay attention to that duration to make their judgements about response timeliness, whether in relation to preceding speech rate or not. Final Rhyme Duration was included to test whether there was a local as well as a global rate effect, since local variation in segment duration has also been observed to induce rate normalisation effects (Summerfield, 1981). Because Final Rhyme Duration was included,

this procedure was applied only to the subset of data that has $TTT \geq 0$ (Gaps). A quadratic term for Question Speech Rate was also tested, based on visual inspection of scatterplots (see Sections 5.3.5 and 0). All models included random intercepts for Participant, Trial Number and Stimulus. The model selected using this procedure is referred to as the experimental model.

Next, a forward model selection procedure using LRTs and the Akaike Information Criterion (AIC, Akaike, 1973) was used to test the significance of the additional contribution of the remaining speech rate and duration related variables Question Length, Response Length and Response Speech Rate, to the experimental model. The model selected using this procedure is referred to as the exploratory model. Response Speech Rate was added to test if there is a retrospective normalising effect by the speech rate of the response. The speech rate of the response is perhaps less likely to affect the perceptions of turn timing of interlocutors, who form their interpretations of turn timing online (Bögels et al., 2020), but could affect the perceptions of participants, who also heard the response before rating the stimuli. Question Length and Response Length were added to explore if aspects of the duration of stimuli affected participants' impressions of timeliness, which could indicate an effect of the task demands.

In previous analyses, the rating data were used to create an aggregate, supplementary variable to a (subset of) the corpus of Dutch and English question-response sequences analysed in the previous chapter. As this secondary analysis concerns unaggregated rating data, we analysed the two studies that collected these data separately.

Note that although the outcome variable in this study, Rating, is ordinal, this procedure used linear regression models. This means that Rating was treated as an interval variable. The reason for this choice was that ordinal mixed effect regression models predicting Rating that included any random intercepts for Participant, Stimulus and/or Trial Number were rank-deficient. In this case, the rank deficiency may be the result of an insufficient number of instances of observations in each combination of the levels of these variables, perhaps because the mid-range of the rating scale was used more than the tails (see Sections 5.3.1 and 5.3.2 below). The predictor variables of interest were repeated because multiple participants rated every stimulus, hence, linear regression was used to allow random intercepts to be specified, rather than for the mathematically more appropriate ordinal regression model that does not account for repeated instances of the

variables of interest – the latter option could inflate the significance of the predictor variables of interest.

An additional consideration in favour of using linear regression models, rather than ordinal regression models without random effect structure, was that the distances between categories on the rating scale are more likely to represent equidistant points of a continuous concept than ordinal categories often do (i.e., on a scale from ‘primary school’ to ‘university’, ‘high school’ does not represent the mean in any sense, but category 4 ‘on time’ is more likely to represent a mean). In other words, perceptions of timeliness can be measured on a continuous scale as well as a discrete scale, since a perceived timeliness of 4.6 is not any less or more meaningful than a perceived timeliness of 4 or 5. Although this expectation is not validated, Rating is used as an interval variable in this analysis, and mixed-effect linear models are chosen in favour of sacrificing the random effects structure to treat the outcome variable as ordinal.

All visualisations and statistical analyses were implemented in the statistical software R (R Core Team, 2022) using the packages lme4 (Bates et al., 2015), ggplot2 (Wickham, 2016), and car (Fox & Weisberg, 2019). Density plots were drawn using ggplot2’s default Gaussian kernel for density estimation and lines of best fit were drawn using ggplot2’s default linear and polynomial functions. Questionnaire responses were processed using Excel (*Microsoft Excel*, 2018).

5.3 Results

This section presents the results of the analysis procedures described in Section 5.2.9. First, Sections 5.3.1 and 5.3.2 present preliminary analyses of the English and Dutch studies separately. Using mainly descriptive statistics of the rating data and questionnaire responses, these two sections aim to characterize the distribution of responses and the degree of inter-rater agreement. These sections also address the effect of presentation order and consider participant responses to the debriefing questionnaire.

Sections 5.3.3 and 5.3.4 then address research question 3.1, and Sections 5.3.5 and 5.3.6 address research question 3.2. Although Sections 5.3.1 and 5.3.2 present descriptive statistics of the rating data in general, additional descriptive statistics and visualisations are presented where they are relevant to subsequent analyses.

5.3.1 Preliminary analysis of the English study

The distribution of ratings from the English study across the seven levels of the rating scale is displayed in Figure 5.2. The categories ran from 1 to 7; rating level ‘1’ corresponded to ‘early’, rating level 4 to ‘on time’ and rating 7 to ‘late’, while the other levels were not labelled. Participants were asked to use the whole scale, but they chose level 4 ‘on time’ most often, although they did also use the bottom and top ends of the scale. Level 3 was chosen second most often, and level 7 ‘late’ least often. Most participants (63 out of 94) used the whole scale. Of those that did not, 15 did not use level 7, five participants did not use level 1, and eleven participants used a smaller range than six out of seven levels.

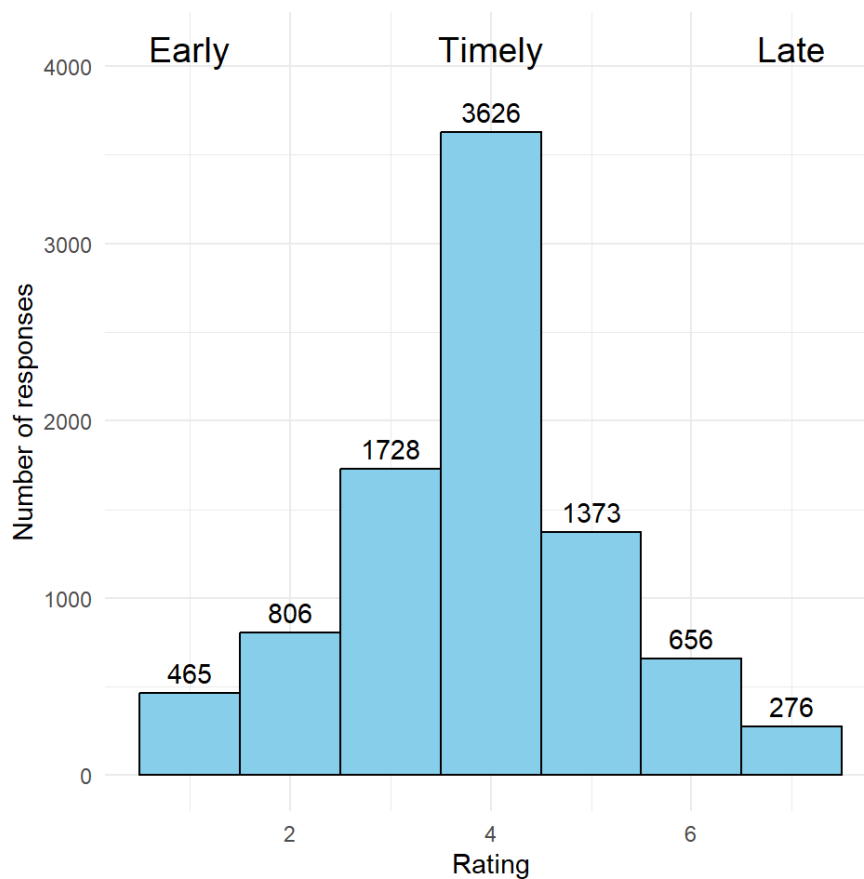


Figure 5.2 Distribution of ratings by English-speaking participants.

There is some degree of consistency between participants’ ratings per stimulus. The mean standard deviation per item is relatively small, at 1.00 points, which corresponds to one level of the range from 1 to 7, or 17% when treated as an interval scale. The mean range, however, is wide, at 4.3 out of 7 (covering 61% of the absolute range), indicating that there is a degree of variability in ratings per stimulus between participants.

The distribution of standard deviations is depicted in Figure 5.3a. As TTT is expected to closely relate to participants' ratings of the perceived timeliness of responses, Figure 5.3b depicts the standard deviation by TTT, to assess whether the degree of agreement between participants depends on the actual duration of TTT. There appears to be no clear pattern indicating that the degree of disagreement between participants, expressed as the standard deviation, is strongly related to TTT, except that those stimuli with a small standard deviation in ratings tended to have a TTT between 0 and 500ms.

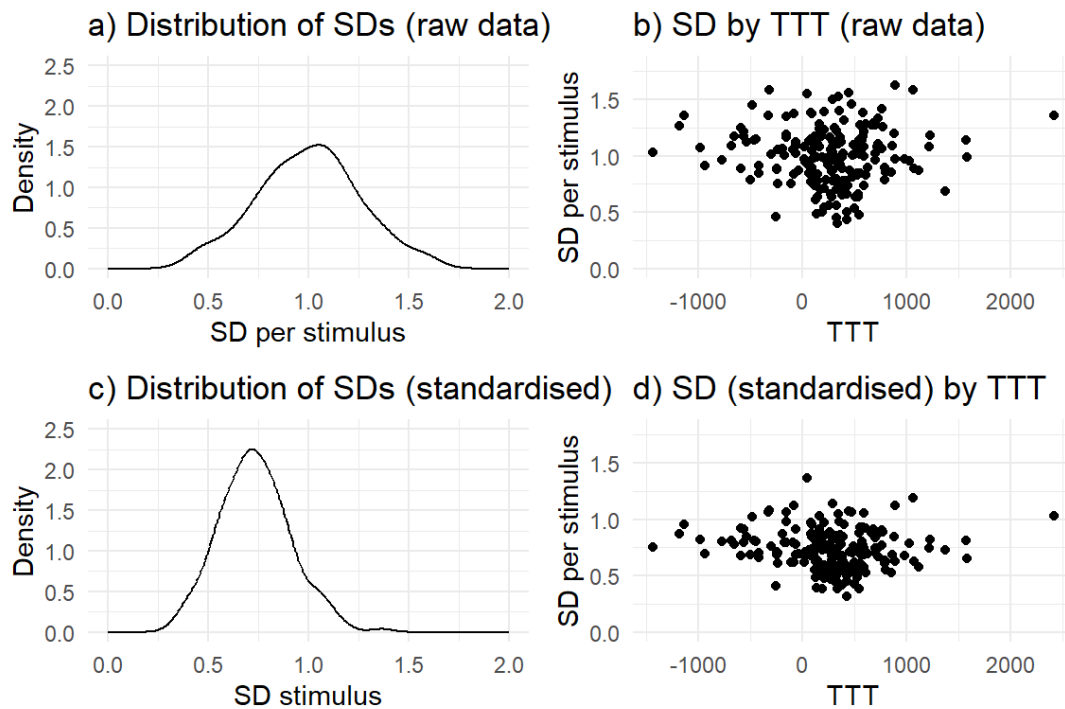


Figure 5.3 Illustrations of the spread of ratings per stimulus for the English study for raw and scaled data.

A potential difference between participants in how they use the rating scale somewhat accounts for the variability in ratings per stimulus: when the ratings are standardised per participant, the mean standard deviation per stimulus is 0.73 on a scale from -3.79 to 3.72 (10% of the range of ratings standardised per participant) and the mean range is 3.32 (covering 44% of the total range). Note, however, that the maximum range of the data when they were standardised per participant is only approximately meaningful, as the maximum and minimum values of this transformed rating scale were not explicitly available for participants to choose during the study, and do not correspond to a true observable minimum or maximum.

Figure 5.3c depicts the distribution of standard deviations per stimulus of ratings standardised by participant, and Figure 5.3d depicts the standard deviation per stimulus of ratings standardised by participants, plotted against TTT. The pattern that those stimuli

Category	Count (of 94)	Description	Examples
yes	15	Participant described timing as noticeable or very noticeable	<i>"It was definitely noticeable. Timing completely sets the tone of the interaction."</i>
A little	12	Participant described timing as a little noticeable	<i>"not to a large extent"</i> <i>"Somewhat noticeable, but I couldn't help focusing more on the content of their answers."</i>
sometimes	33	Participant described timing as noticeable in some cases, such as when it is very early, very late, or both	<i>"the later ones were more noticeable"</i> <i>"very early or very late were clear, the rest sounded like they were an appropriate reaction time even if they were slightly early or later because it depends on context of the question"</i>
interpretation	11	Participant only described their interpretation of timing rather than the degree to which they found it noticeable	<i>"I think it set apart if they were nervous or confident, knew each other or didn't"</i> <i>"Earlier answers sounded more like everyday conversation and therefore it is likely the speakers are friends or already know each other"</i>
no	18	Participant described timing as not noticeable	<i>"Not much"</i> <i>"Barely noticeable at all in most cases."</i>
?	5	Participant gave an answer that did not pertain to the question, expressed they did not understand the question, or filled in a white space.	<i>"Don't understand"</i> <i>"Not sure"</i>

Table 5.2 Categories of responses from the English study to the debriefing question "To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?".

with a small standard deviation in ratings tended to have a TTT between 0 and 500ms is less pronounced for the standardised rating data, but still present.

Participants generally reported finding the timing of responses a salient feature of the stimuli set in general, although many said they found only specific cases of very early or very late responses obviously salient. Table 5.2 (previous page) summarises participant responses to the debriefing question “To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?” This question was included to confirm that uninformed participants thought they were aware of the phenomenon they were asked to rate, and to invite them to comment on it. Because this question required an open-form response, the participant responses were coded into six categories. Table 5.2 describes each category and the frequency of participants whose answer fell into this category. Most participants reported finding timing as a feature of responses noticeable, a little noticeable, or noticeable in a subset of cases (60 out of 94). 18 participants reported not finding the timing of responses a noticeable feature of responses.

To investigate whether presentation order affected participants’ responses, the effect of trial number on ratings was assessed. There is a significant positive correlation between Trial Number and Rating: Spearman’s $\rho(8928) = .07, p < .001$ (the p-value is approximate as there were ties in the ranked data). Trials that appeared later in the study were rated as ‘late’ to a larger degree than earlier trials. There was a similar correlation when the data were standardised per participant: Spearman’s $\rho(8928) = .07, p < .001$. Figure 5.4 (next page) illustrates this relationship between Trial Number and Rating, plotting Trial Number against mean (unstandardised) Rating by Trial Number. The difference between the trial number with the highest and lowest mean rating is not large, at 1.04 points, but does span more than one rating category. Figure 5.4a depicts a narrow range on the y axis to visualise the confidence interval of a line of best fit and Figure 5.4b shows the full range of ratings on the y axis to illustrate the gradient of that effect.

Although this correlation between Trial Number and Rating is significant, it is extremely weak. However, the order of presentation of each stimulus was randomized for each of

the 94 participants, meaning that this weak effect is unlikely to have arisen because of a feature of the stimuli themselves. Also, the confidence interval in the visualisation in Figure 5.4a is narrow and suggests a true effect.

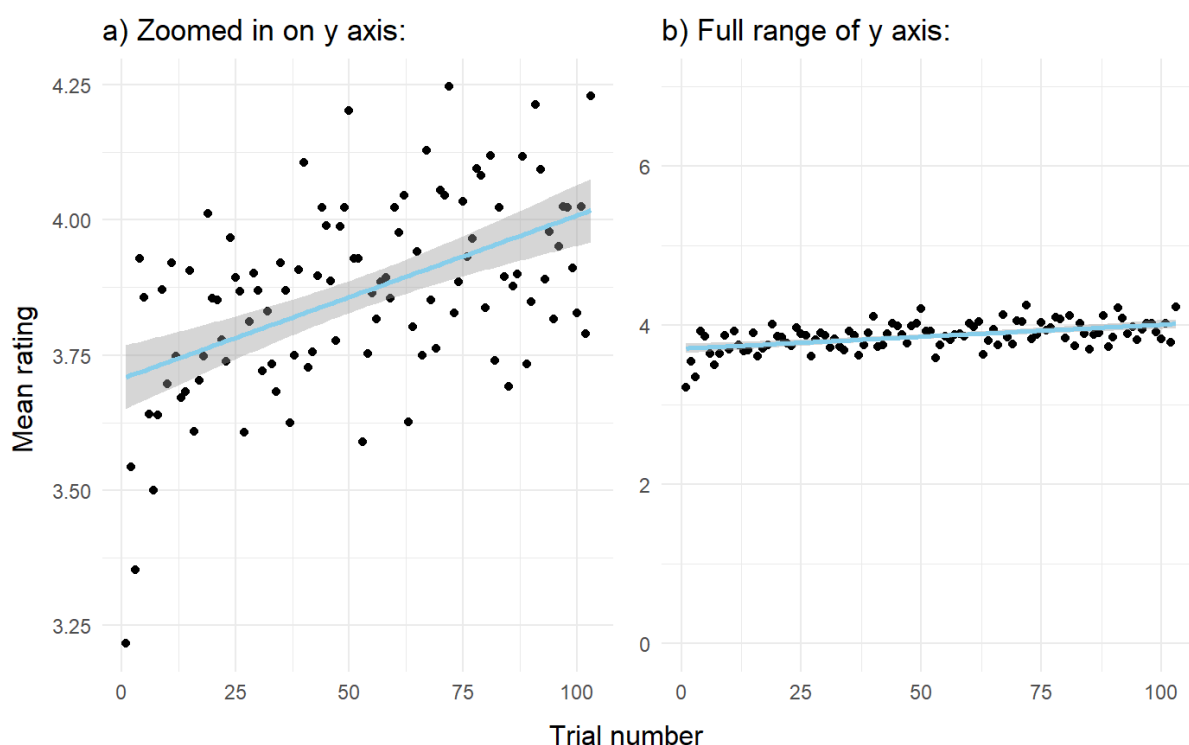


Figure 5.4 Illustration of the effect of presentation order in the English data as Trial Number against mean Rating by Trial Number

This effect of presentation order may be due to habituation, if participants perceived the responses they heard in stimuli later in the study as later, because they were familiar with the speaker's voice at that point. However, this effect may also simply be due to response bias: later in the study, participants may have been fatigued and more often biased towards the right (i.e., 'late') side of the scale, assuming most participants were right-handed. Since the scale was not flipped between participants, this scenario may result in the pattern observed here. If it is the case, we might expect that stimuli that were heard later, were responded to with ratings in a narrower range than stimuli that were heard earlier. Conversely, participants may have familiarised themselves with the range of turn-timing in the stimuli towards the end of the study and started using a wider range of responses that drew the mean ratings up.

Neither of these expectations is confirmed by the distribution of ratings of stimuli that appeared early versus late in the study. Figure 5.5 illustrates the distribution of ratings in each rating category for the first and last quarter of the study (i.e., first and last quartile of

Trial Number). Figure 5.5 does not reveal a large discrepancy between the distributions of ratings of stimuli that were heard early in the study versus those that were heard late in the study.

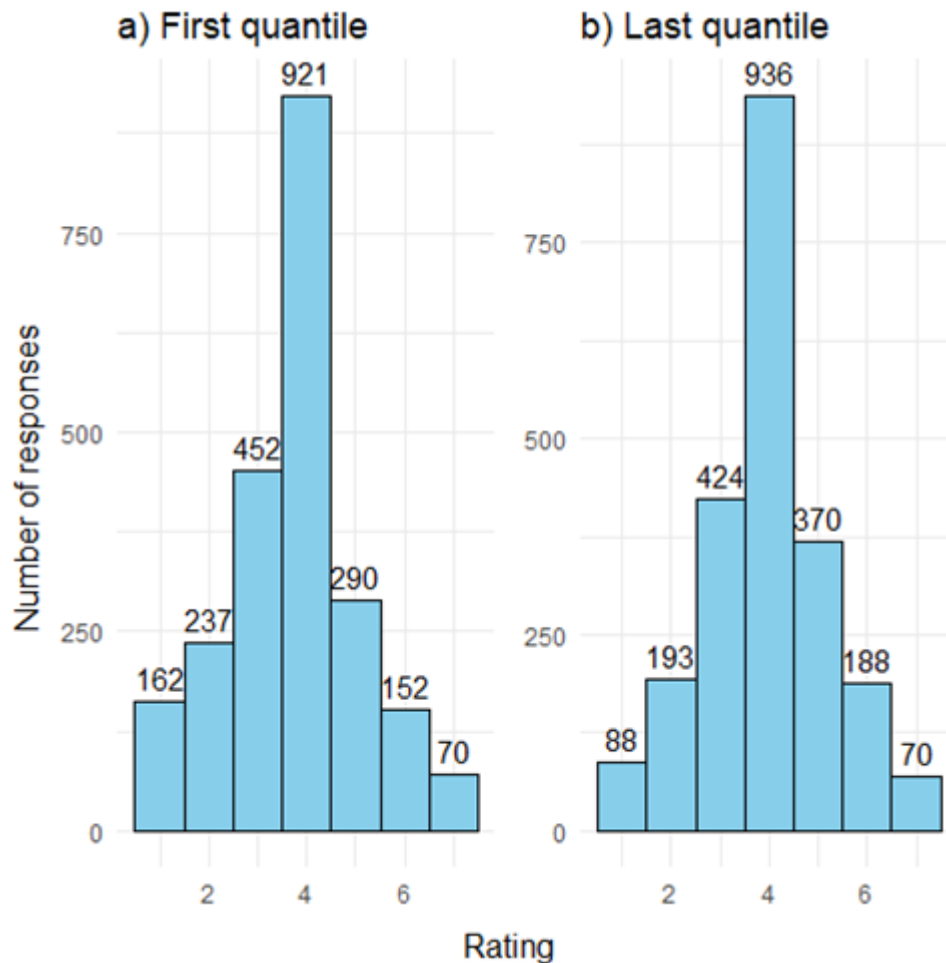


Figure 5.5 Distribution of ratings by English speaking participants in the first and last quarter of the study (i.e., first and last quantile of Trial Number).

5.3.2 Preliminary analysis of the Dutch study

The distribution of ratings across the seven levels for the Dutch study, illustrated in Figure 5.6, looks similar to the English study. As before, 1 to 7; rating '1' corresponded to 'early', rating 4 to 'on time' and rating 7 to 'late'. Other categories were not labelled, and participants were asked to use the whole scale. The Dutch participants also chose level 4 'on time' most often, and Figure 5.6 shows a similar distribution of response frequencies towards the ends of the scale as Figure 5.2 does for the English study, with level 4 chosen second most often and level 7 least often. Most participants (35 out of 52) used the whole scale, ten participants did not use level 7, two participants did not use level 1, and five participants used a smaller range than 6 out of 7 levels.

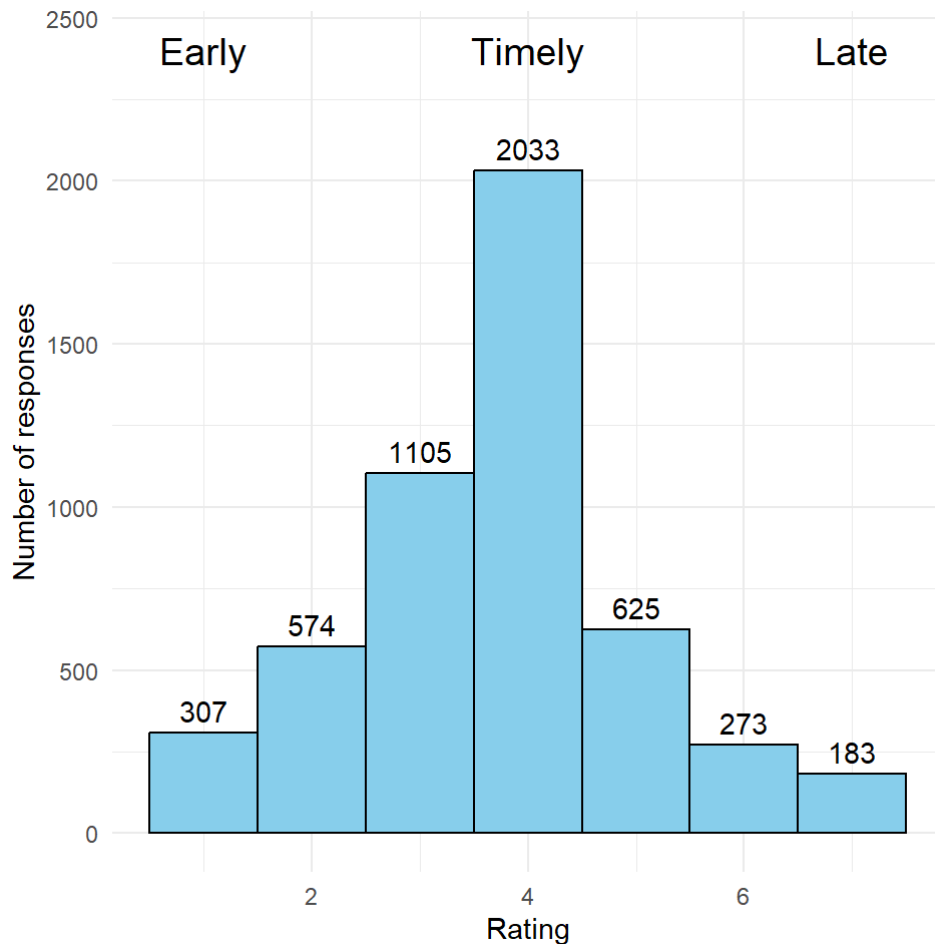


Figure 5.6 Distribution of ratings by Dutch-speaking participants.

The spread of ratings per stimulus looks similar to that in the English study as well. The mean standard deviation of ratings per stimulus is 1.0 (17% of the total range from 1 to 7). The mean range of ratings per stimulus is 4.5 of 7 points (65% of the total range). When the rating data are standardised per participant, the mean standard deviation of standardised ratings per participant per stimulus is 0.7, 9% of the total range of data standardised per participant, -3.7 to 4.0. The mean range of standardised ratings per stimulus is 3.3 – 43% of the total range of standardised ratings.

The distributions of standard deviations are depicted in Figure 5.7a and Figure 5.7c for the raw and standardised data respectively, and Figure 5.7b and Figure 5.7d depict standard deviation plotted against TTT for the raw and standardised data respectively. The distributions of standard deviation are skewed, with the highest density of standard deviations towards the low end. The pattern in the English data, that those stimuli with a

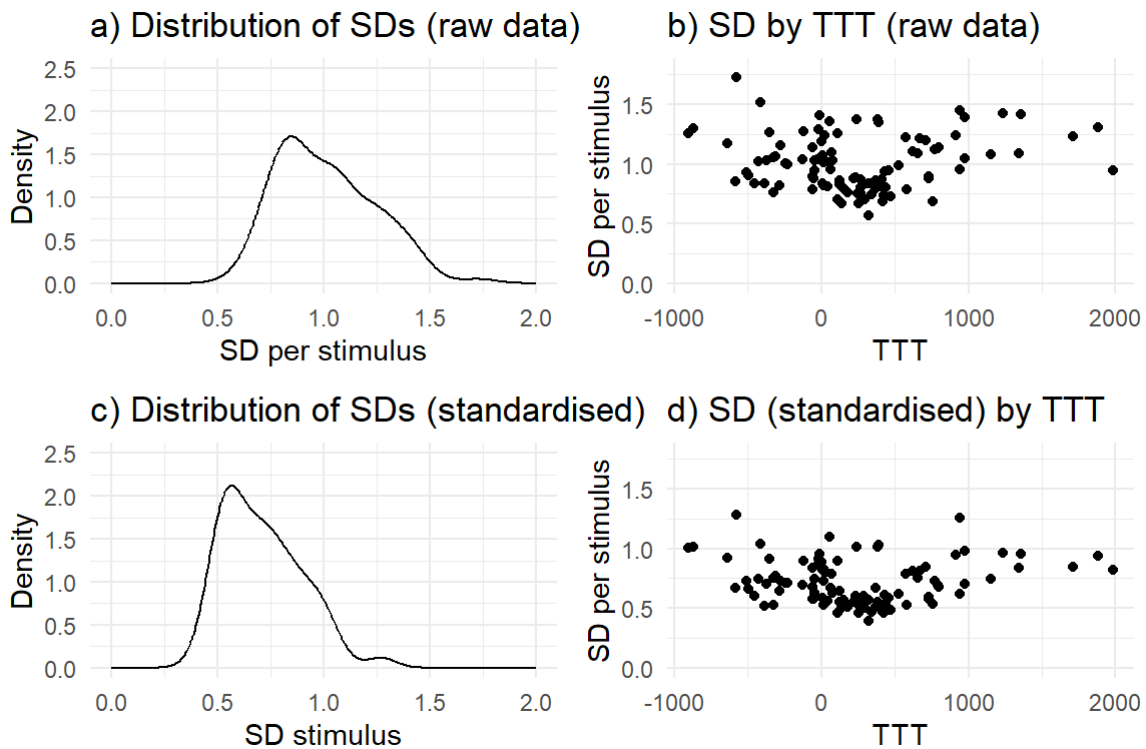


Figure 5.7 Illustrations of the spread of ratings per stimulus for the Dutch study.

small standard deviation in ratings tended to have a TTT between 0 and 500ms, is not visible in the Dutch data.

The Dutch participants' responses to debriefing questionnaire question "To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?"⁸ were also coded, and the frequency of responses in each category are summarised in Table 5.3. Again, most participants reported finding the timing of the response a noticeable, or sometimes noticeable, feature of the stimuli (42 out of 50). Others reported that it was not noticeable, or that the stimuli seemed very similar in response timing.

Unlike in the English study, there is no significant correlation between Trial Number and Rating: Spearman's $\rho(5212) = .02$, $p = .227$ (the p-value is approximate as there were ties in the ranked data), although there is a trend in the same direction, and the difference between the lowest and highest rated Trial number is larger than in the English study, at

⁸ Dutch: "In hoeverre vond je de timing van antwoorden een duidelijk merkbaar fenomeen? Waren de fragmenten duidelijk verschillend in de timing van de antwoorden?"

1.21. This relationship is similar when the data are standardised per participant: Spearman's $\rho(5212) = .01, p = .353$.

Category	Count (of 50)	Description	Examples and translation
yes	13	Participant described timing as noticeable or very noticeable	<i>"Best wel"</i> – Quite <i>"Ja dit was wel merkbaar"</i> – Yes this was quite noticeable
A little	4	Participant described timing as a little noticeable	<i>"Niet heel duidelijk. In de meeste gevallen wilde ik 'op tijd' antwoorden, maar door subtiele verschillen antwoordde ik vaak iets anders."</i> – Not very noticeable. In most cases I wanted to respond 'on time', but because of subtle differences I often responded something else.
sometimes	25	Participant described timing as noticeable in some cases, such as when it is very early, very late, or both	<i>"het was soms duidelijk hoorbaar hoe de timing is."</i> – It was sometimes clearly audible what the timing was like.
interpretation	2	Participant only described their interpretation of timing rather than the degree to which they found it noticeable	<i>"bij de informele gesprekken waren de antwoorden vaker vroeg"</i> – In the informal conversations the responses were more often early.
no	5	Participant described timing as not noticeable	<i>"Nee, als ik die gesprekken in de trein zou horen zou mij niks opvallen"</i> - No, if I heard these conversations in the train I wouldn't notice anything.
?	1	Participant gave an answer that did not pertain to the question, expressed they did not understand the question, or filled in a white space	" " - [white space]

Table 5.3 Categories of responses from the Dutch study to the debriefing question "To what extent do you think the timing of the answers was a noticeable phenomenon to you? Did it clearly set apart different fragments?"

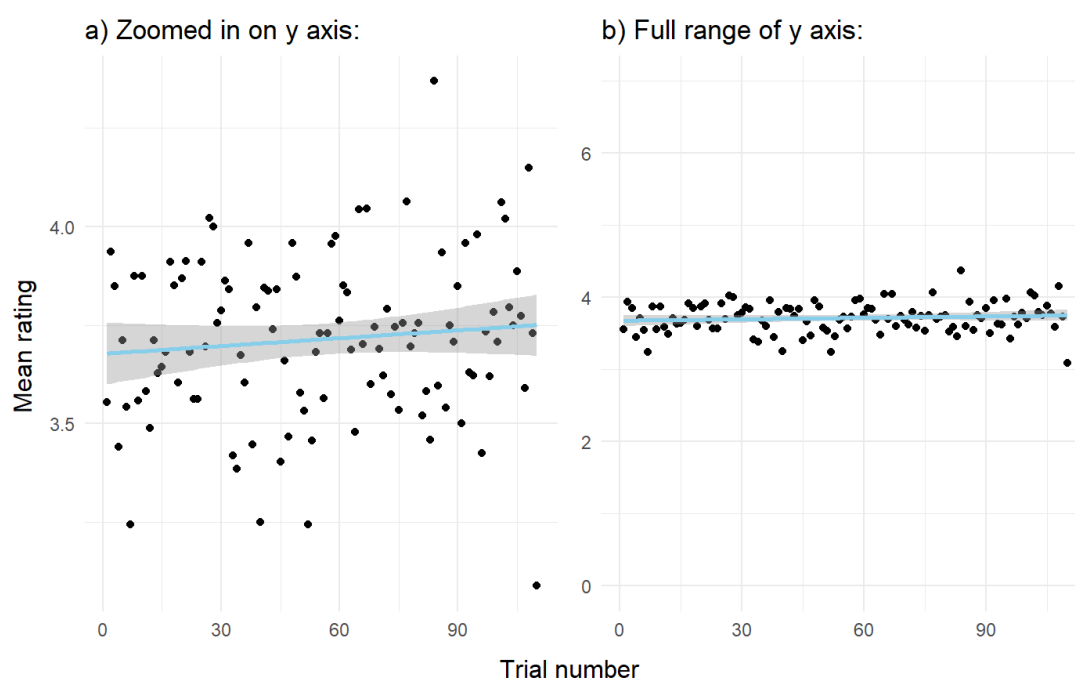


Figure 5.8 Illustration of the effect of presentation order in the English data as Trial Number against mean Rating by Trial Number.

Figure 5.8 illustrates this non-significant trend in the relationship between Trial Number and Mean Rating per Trial Number. Note that the confidence interval of the linear regression line in Figure 5.8a is wider than in Figure 5.4a.

In conclusion, there is a moderate degree of inter-rater agreement, and participants generally found the task meaningful, but mostly when extremely early or late responses were concerned.

5.3.3 Mean Standardised Rating

The primary aim of this chapter is to test whether the relationship between speech rate and turn timing is modulated by whether the turn sounds ‘on time’. To do so, we express the perceived timeliness of responses in each observation as the mean of ratings per stimulus after they are standardised per participant. This value is denoted as Mean Standardised Rating, and is used to distinguish timely turns from early and late turns. These analyses concern the pooled data for English and Dutch, following the same procedure as in Chapter 4.

First, Figure 5.9 depicts the distribution of Mean Standardised Ratings per stimulus for both languages together. The range of unaveraged ratings standardised per participant is -3.8 to 4.0. The distribution of Mean Standardised Ratings falls in a relatively narrow range on that scale, from -1.6 to 2.3, due to inter-rater variability. The black vertical lines in Figure 5.9 indicate the point before which one-third of observations fall ('Early'), and the point below which two-thirds of observations fall ('Late'). In Section 5.3.4, these three bins were used to categorise the data.

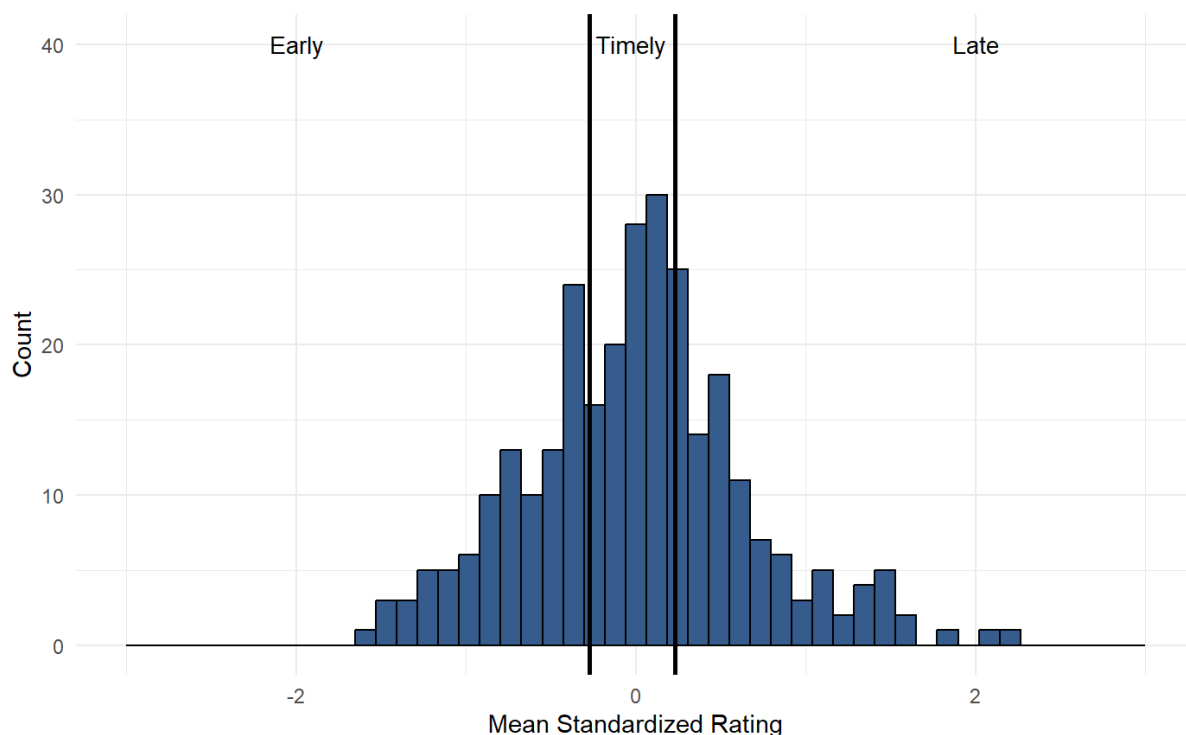


Figure 5.9 The distribution of Mean Standardised Ratings.

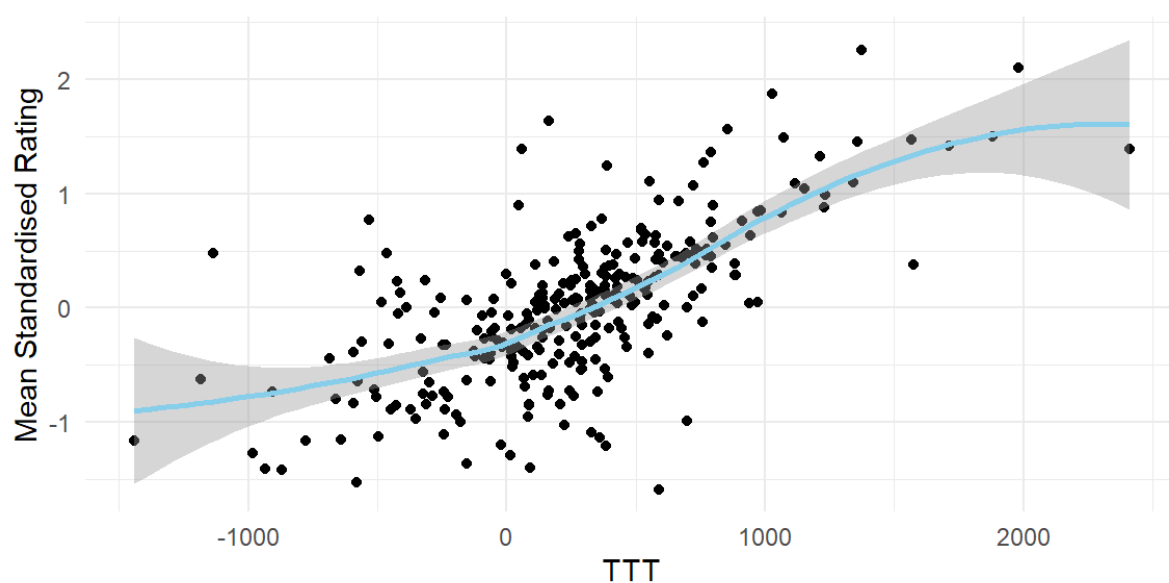


Figure 5.10 Scatterplot of TTT against Mean Standardised Rating per stimulus.

Mean Standardised Rating seems to be correlated with TTT, as depicted in Figure 5.10 (previous page). This relationship between Mean Standardised Rating and TTT is expected, since participants were expected to at least also pay attention to TTT when rating turn-timing, even if they may also use other cues or be affected by speech rate normalisation.

The coupled oscillator models of turn-timing (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005) predict that for a particular value of TTT, responses to a faster question should be rated as later than responses to a slower question. However, a 3D scatterplot of TTT against Mean Standard Rating against Question in Figure 5.11 does not reveal a clear modulating effect of Mean Standard Rating on the relationship between Question Speech Rate and TTT. The regression plane is almost flat and almost parallel to the Question Speech Rate axis, and the distribution of prediction errors is uniform along the Question Speech Rate axis. If Mean Standardised Rating clearly modulated the relationship between Question Speech Rate and TTT as predicted, then the bottom right corner of the regression plane as depicted in Figure 5.11b would be curled backward, i.e., for a particular TTT, the faster the speech rate, the higher the Mean Standardised Rating is expected to be. This expectation is not supported by Figure 5.11.

5.3.4 Speech rate and turn timing in Early, Timely and Late sequences

To test whether the relationship between Question Speech Rate and TTT is different when turns are perceived as timely, early, or late, we first used Mean Standardised Rating to categorically distinguish between timely turns, early turns, and late turns. To do so, we split the Mean Standardised Rating distribution into three equal subsets: question-response sequences with a Mean Standardised Rating in the lowest third of the distribution are referred to as Early, the middle third are Timely, and the highest third are Late. The black vertical lines on the histogram of Mean Standardised Ratings in Figure 5.9 indicate this split. A breakdown of the number of observations in each subset is listed in Table 5.4 (page 177).

For each subset, a linear model predicting TTT from Question Speech Rate, was compared to an intercept-only model with an ANOVA. Question Speech Rate only made a significant contribution to the model fit in the Timely subset, $F(1) = 6.68, p = .011$. It does

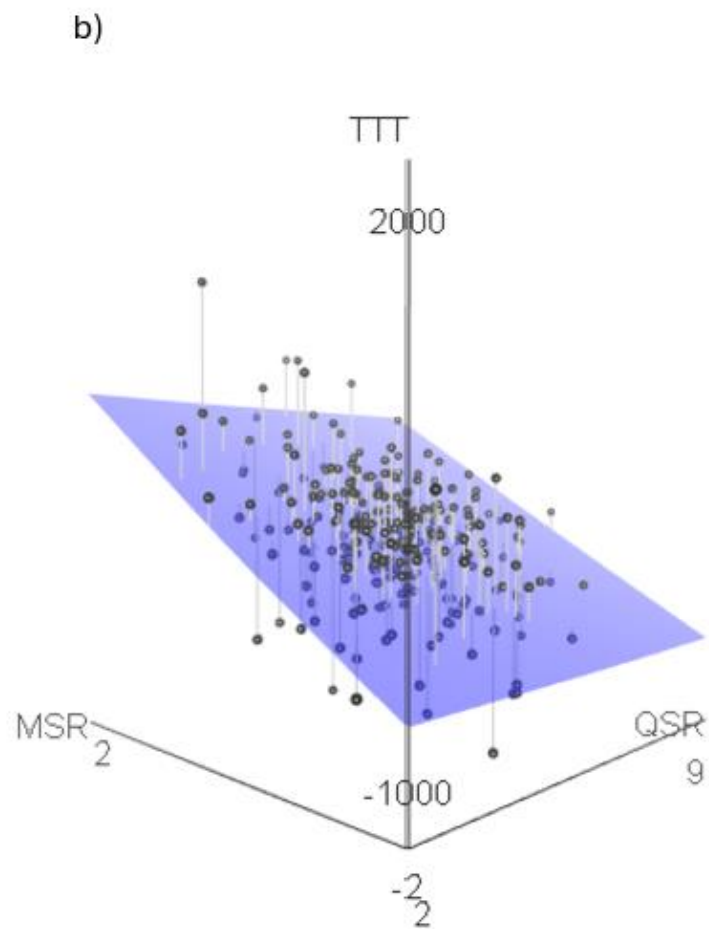
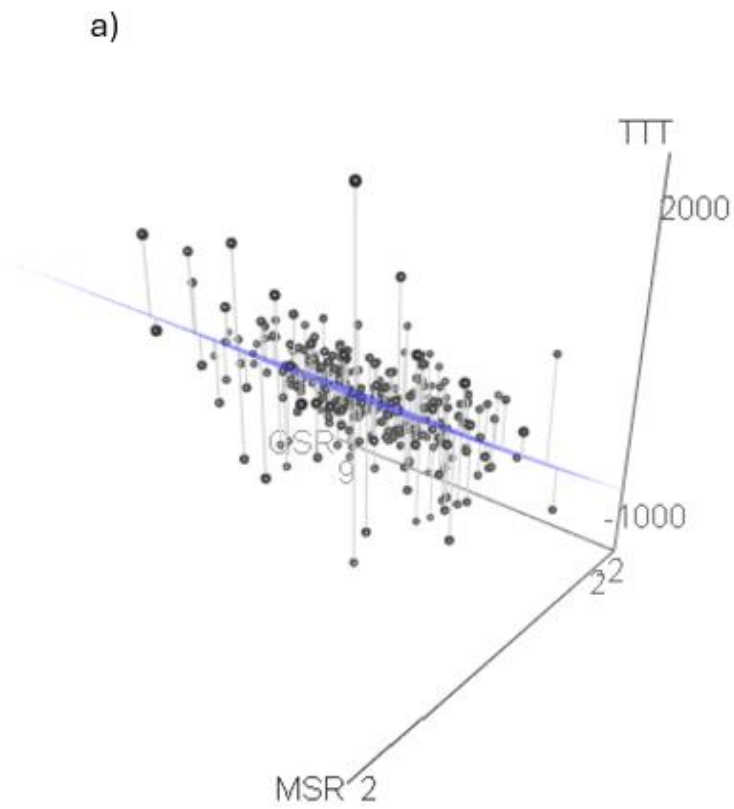


Figure 5.11 3D scatterplot of TTT against Mean Standardised Rating (MSR) against Question Speech Rate (QSR), from two different angles in a) and b).

not in the Early subset, $F(1) = .14$, $p = .710$, or Late subset, $F(1) = .17$, $p = .681$. Table 5.5 lists the coefficients for the Question Speech Rate effect in the models for each of the subsets, as well as model comparison test statistics.

For the Timely subset, the coefficient for Question Speech Rate is positive, meaning that faster questions are associated with later turn timing. This effect is in the same direction as in Chapter 4, and contrary to predictions of the coupled oscillator models (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005). Figure 5.12 illustrates the relationship between Question Speech Rate and TTT in each subset, as well as the distribution of TTT in those subsets. In line with the results of the inferential procedure, the regression line in the Timely subset is positive, and its confidence interval is narrower than those in the Early and Late subsets.

Subset	Full dataset	Gaps
Early	97	48
Late	98	82
Timely	97	90
<i>Total</i>	292	220

Table 5.4 Overview of number of observations per subset.

Subset	Question Speech Rate coefficient	<i>F</i>	p-value
Early	-12.57	.14	.710
Timely	57.38	6.68	.011
Late	-18.03	.17	.681

Table 5.5 Coefficients for Question Speech Rate for models predicting TTT from Question Speech Rate in the Early, Timely and Late subsets.

In parallel to the analysis in the previous chapter, the contribution of Question Speech Rate, Final Rhyme Duration and Question Speech Rate $\times$ Final Rhyme Duration in models predicting TTT for only Gaps (TTT ≥ 0) in each subset was also tested.

For Gaps in the Timely subset, the interaction does not significantly improve the model fit, $F(1) = 1.86, p = .177$, and neither does Final Rhyme Duration, $F(1) = 3.27, p = .177$.

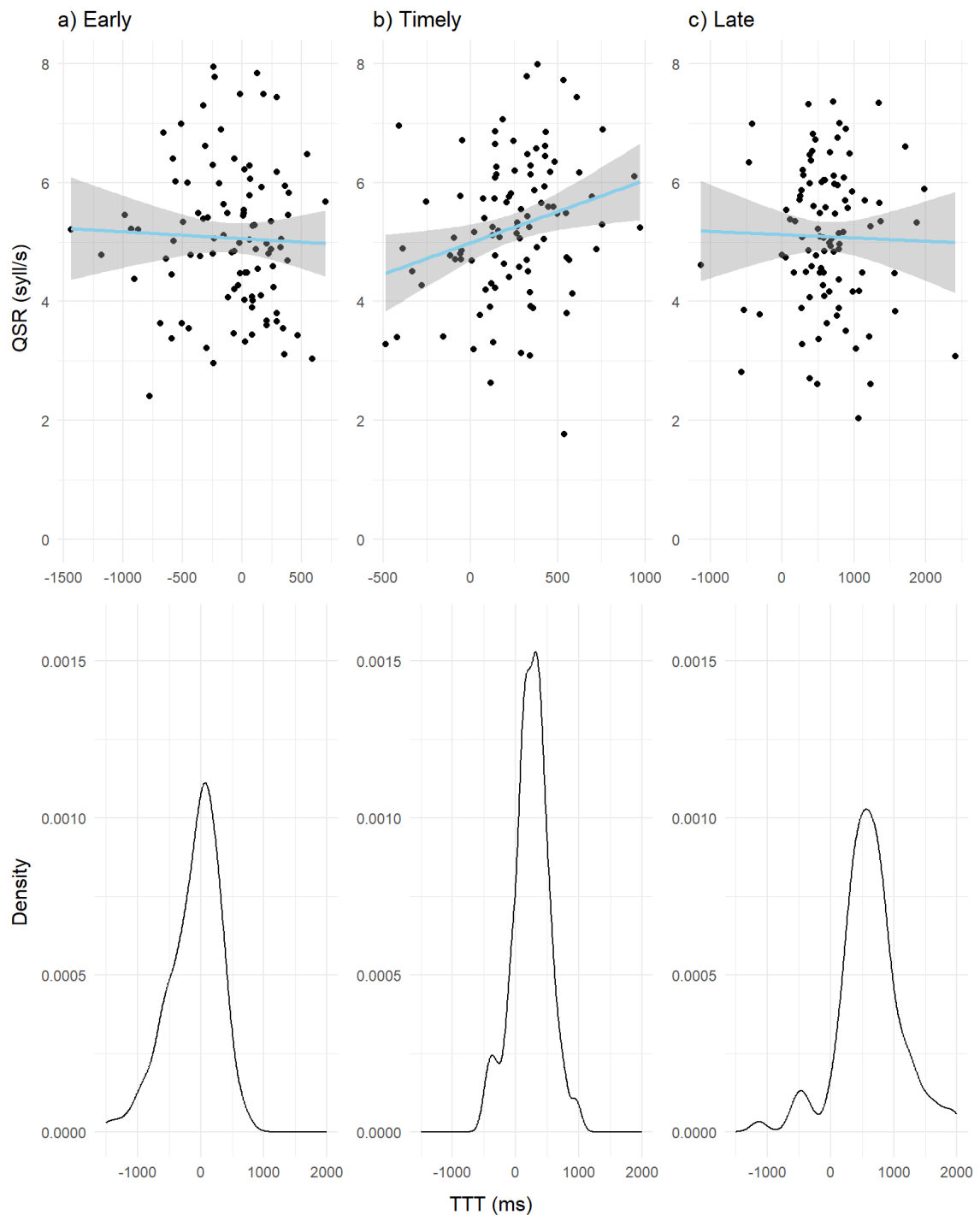


Figure 5.12 Scatterplots of Question Speech Rate (QSR) and TTT and density distributions of TTT for a) Early, b) Timely, and c) Late subsets of the full dataset.

Question Speech Rate is marginally significant, $F(1) = 3.27, p = .074$. As before, faster speech rates are associated with longer TTTs.

In the Early subset, the interaction Question Speech Rate $\times$ Final Rhyme Duration makes a marginally significant improvement to the model fit, $F(1) = 3.74, p = .059$. Question Speech Rate does not significantly improve the model fit, $F(1) = 1.56, p = .218$, and neither does Final Rhyme Duration, $F(1) = 2.67, p = .109$.

In the Late subset, none of the predictors significantly improve the model fit. The interaction does not, $F(1) = 1.74, p = .19$, and neither does the main effect for Question Speech Rate, $F(1) = 2.14, p = .146$, or for Final Rhyme Duration, $F(1) = .69, p = .41$.

Table 5.6 lists, for each subset, the coefficients for included predictors and the F statistic and p-values. The effect of each marginally significant predictor was in the same direction as in the previous chapter: the coefficient for Question Speech Rate for Timely Gaps was positive, contrary to the perception-action entrainment hypothesis, and the interaction between Question Speech Rate and Final Rhyme Duration was marginal and negative.

Subset	Predictor	Predictor coefficient (intercept)	F	p-value
Early	Question Speech Rate $\times$ Final Rhyme Duration	-0.07 (276.84)	3.74	.059
Timely	Question Speech Rate	33.31 (152.61)	3.27	.074
Late	None			

Table 5.6 Coefficients predictors in final models predicting TTT for Gaps in the Early, Timely and Late subsets.

In short, this analysis corroborates the findings of the previous chapter: the relationship between Question Speech Rate and TTT is contrary to the perception-action entrainment hypothesis even when excluding question-response sequences of which the response is perceived as early or late. Moreover, Question Speech Rate only significantly predicts TTT in the Timely subset.

5.3.5 Speech Rate Normalisation in English

Instead of using participants' ratings as an additional variable in the corpus analysis, this section and the next consider the rating from a different perspective. It asks whether

listeners' perception of turn-timing is subject to speech rate normalisation, by testing whether Question Speech Rate, Final Rhyme Duration, and TTT affect raw participant ratings of response timeliness, for English and Dutch respectively. This analysis was performed using the portion of data that contains positive TTT, Gaps, where the response starts after the end of the final syllable of the question.

For English, 150 out of 190 stimuli had positive TTTs. Figure 5.13 illustrates the relationship between Rating and each of the three predictor variables of interest in this subsection, TTT, Question Speech Rate and Final Rhyme Duration. The light blue lines indicate a polynomial line of best fit. Note that Rating is a categorical variable, but it is plotted with horizontal jitter (width = .4) to illustrate the density of observations at a particular value of the variable represented on the y axis.

The relationships between the variables in Figure 5.13 are not very clear. The line of best fit suggests, as expected, that there is a positive relationship between TTT and Rating, where stimuli with longer TTT appear to have been rated as later. This relationship is much stronger for the higher half of the rating scale than for the lower half. Contrary to expectations, Final Rhyme Duration and Rating also appear to stand in a positive relationship, where responses were rated as later in stimuli in which the question had a longer final rhyme. The relationship between Question Speech Rate and Rating does not seem to be linear: responses to faster questions were rated later than slow questions within the 'early' part of the scale, but earlier within the 'late' part of the scale.

Outliers in the TTT, Final Rhyme Duration and Question Speech Rate distributions were identified through visual inspection of the scatterplots. They are marked using yellow, blue, and red respectively in Figure 5.13. As Figure 5.13 illustrates, no stimuli were outliers for all three variables. These outliers were removed from the analysis.

To test the significance of the contribution of each of the predictor variables of interest, a linear regression model was built predicting Rating from TTT, Question Speech Rate and Final Rhyme Duration, and all possible two-way and three-way interactions between TTT, Question Speech Rate, and Final Rhyme Duration, with random intercepts for Participant, Stimulus and Trial Number. A quadratic term for Question Speech Rate was also added, based on the shape of the polynomial line of best fit between Question Speech Rate and

Rating in Figure 5.13. Following the backward model selection procedure described in Section 5.2.9, the significance of the contribution of each predictor variable to the model fit was tested iteratively using LRTs, starting with the quadratic term, then the most complex interactions, while dropping non-significant variables.

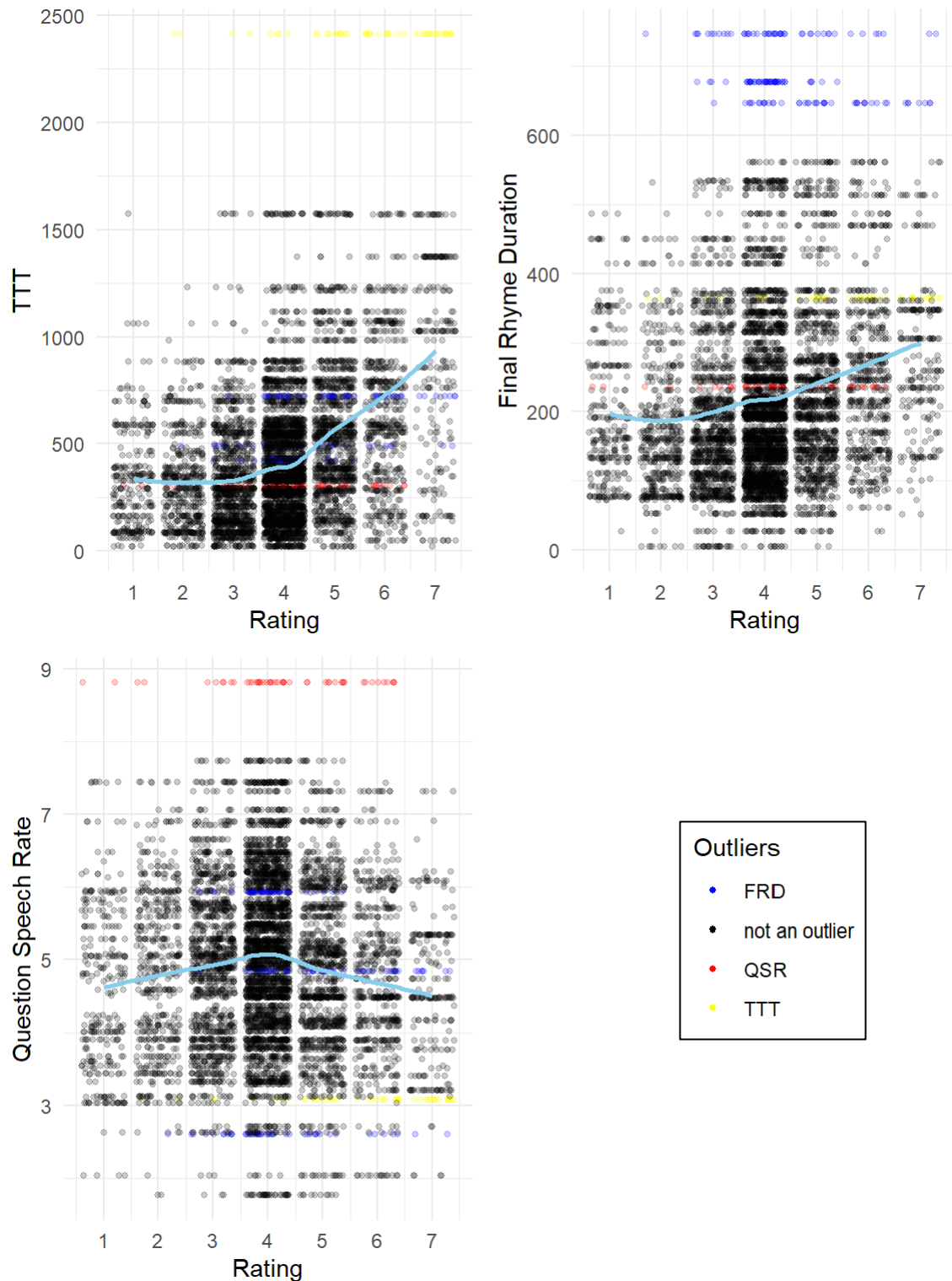


Figure 5.13 Scatterplots of Rating against TTT, Final Rhyme Duration (FRD) and Question Speech Rate (QSR), for English.

First, the quadratic term for Question Speech Rate does not significantly improve the model fit, so it is dropped from the model, $\chi^2(1) = .18, p = .672$. Next, the three-way interaction does not make a significant contribution to the model fit and is also dropped, $\chi^2(1) = .02, p = .879$. The TTT $\times$ Question Speech Rate interaction, $\chi^2(1) = .73, p = .394$, and the Question Speech Rate $\times$ Final Rhyme Duration interaction, $\chi^2(1) = 1, p = .310$, are dropped as well. The TTT $\times$ Final Rhyme Duration interaction significantly improves the model fit, $\chi^2(1) = 5.93, p = .015$, so it is retained.

Next, none of the main effects significantly contribute to the model fit, so Question Speech Rate, $\chi^2(1) = .77, p = .382$, TTT, $\chi^2(1) = 2.59, p = .107$, and Final Rhyme Duration, $\chi^2(1) = .104, p = .747$, are not included in the final experimental model.

In the final experimental model, Rating is predicted from only the TTT $\times$ Final Rhyme Duration interaction. Its coefficient is positive, meaning that as the combination of TTT and Final Rhyme Duration becomes higher, the response tends to be rated as later. As this analysis concerns stimuli with gaps, TTT and Final Rhyme Duration are consecutive segments of the dialogue and cannot have negative values because they never overlap. Therefore, this interaction suggests that when the total span of the final rhyme and TTT increases, responses tend to be rated as later.

Next, the contribution of the exploratory variables Response Speech Rate, Question Length and Response Length were tested in a forward manner. Response Speech Rate significantly contributes to the model fit, $\chi^2(1) = 29.02, p < .001$, so it is added first. Question Length also significantly contributes to the model fit, $\chi^2(1) = 6.96, p = .008$. Response Length does not contribute to the model fit, so it is excluded, $\chi^2(1) = 1.66, p = .198$. The final exploratory model's coefficients and standard errors are listed in Table 5.7 along with LRT statistics for each predictor. The coefficient for Response Speech Rate is negative, indicating that faster responses tended to be rated as earlier for a given TTT. This effect is contrary to a retrospective normalisation effect of the speech rate of the response. The coefficient for Question Length is positive, indicating that responses to longer questions tended to be rated as later.

Predictor	Coefficient	Std. Error	χ^2	p-value
Intercept	4.21	0.21		
TTT × Final Rhyme Duration	.000004	.0000004	5.93	.015
Response Speech Rate	-.21	.04	29.02	< .001
Question Length	.03	.01	6.96	.008

Table 5.7 Model coefficients, standard errors and LRT statistics for each predictor (English).

5.3.6 Speech Rate Normalisation in Dutch

For Dutch, 70 out of 102 stimuli had positive TTTs. Figure 5.14 illustrates the relationship between Rating and each of the three predictor variables of interest in this subsection,

TTT, Question Speech Rate and Final Rhyme Duration. As before, the light blue lines indicate a polynomial line of best fit. None of the relationships are very clear, but TTT does seem to be positively correlated with Ratings. Unlike in the English study, the polynomial line of best fit does suggest a slight negative relationship between Final Rhyme Duration and Rating. No particular pattern is visible in the relationship between Question Speech Rate and Rating.

To test whether each of the variables of interest significantly contributes to a model predicting Rating, the same procedure was applied as for the English data in Section 5.3.5, starting with a model predicting Rating from TTT, Question Speech Rate and Final Rhyme Duration, and all possible two-way and three-way interactions between TTT, Question Speech Rate, and Final Rhyme Duration, with random intercepts for Participant, Stimulus and Trial Number. A quadratic term for Question Speech Rate was added here as well to mirror the procedure applied to the English data, although Figure 5.14 does not suggest there is a quadratic relationship between Question Speech Rate and Rating.

Correspondingly, the quadratic term for Question Speech Rate does not significantly improve the model fit, $\chi^2(1) = .37, p = .541$. Neither does the three-way interaction, $\chi^2(1) = .41, p = .523$, or the TTT × Question Speech Rate interaction, $\chi^2(1) = .36, p = .550$, the TTT × Final Rhyme Duration interaction, $\chi^2(1) = 2.13, p = .145$, or the Question Speech Rate × Final Rhyme Duration interaction, $\chi^2(1) = .63, p = .426$. Of the main effects, Question Speech Rate does not significantly improve the model fit, $\chi^2(1) = .71, p = .399$, but TTT

does, $\chi^2(1) = 42.41$, $p < .001$, and Final Rhyme Duration marginally improves the model fit, $\chi^2(1) = 3.04$, $p = .081$, so they are both retained in the model.

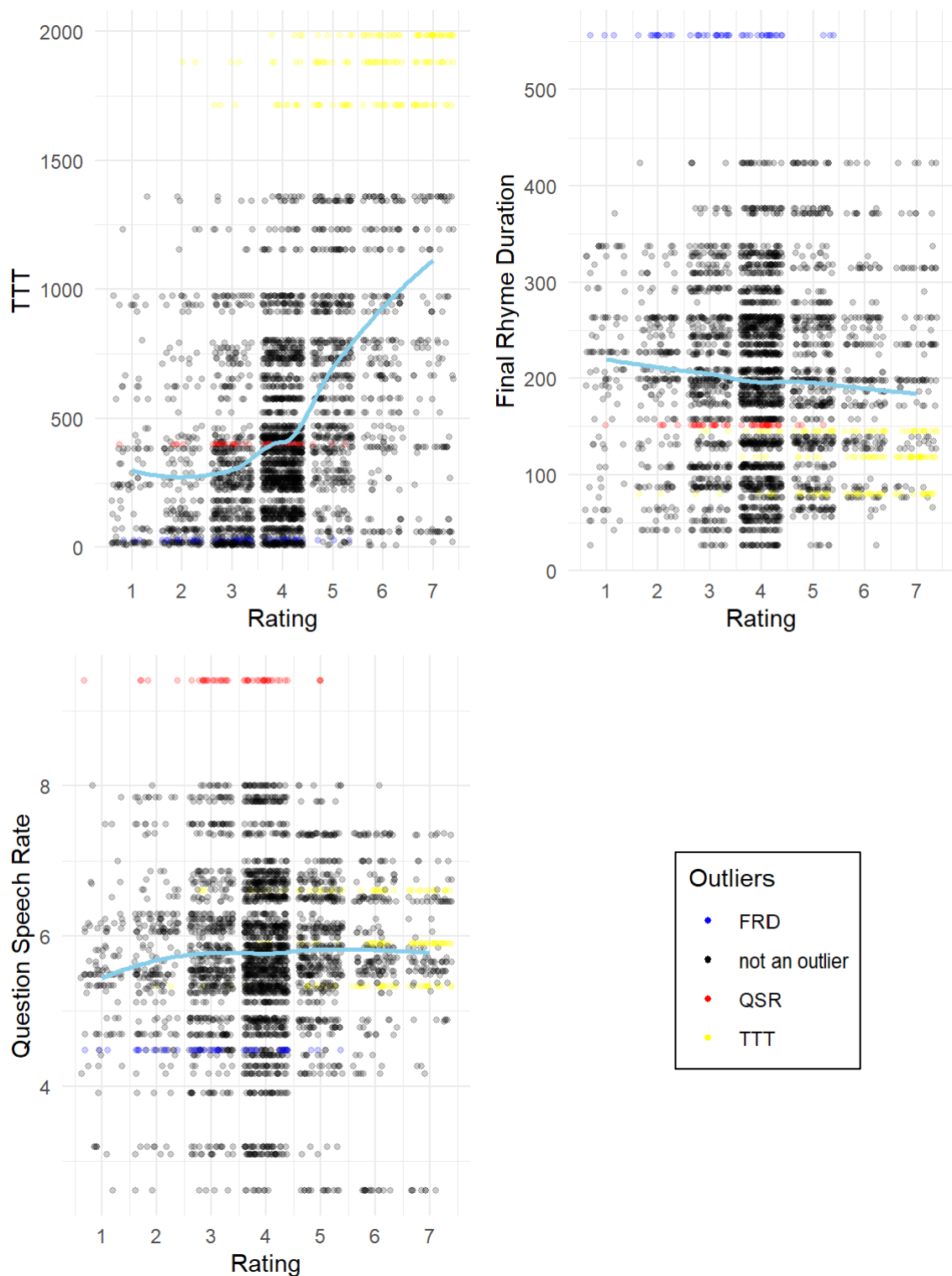


Figure 5.14 Scatterplots of Rating against TTT, Final Rhyme Duration (FRD) and Question Speech Rate (QSR), for Dutch.

The final experimental model for the Dutch gaps data predicts Rating from TTT and Final Rhyme Duration. TTT is positively associated with Rating: unsurprisingly, as TTT is longer, the response tends to be rated as later.

The experimental model was then used as a starting point to test the additional contribution of Response Speech Rate, Question Length and Response Length. Response Speech Rate significantly contributes to the model fit, $\chi^2(1) = 4.93$, $p = .026$, and is added first. Question Length also significantly contributes to the model fit, $\chi^2(1) = 4.97$, $p = .026$, but Response Length does not, $\chi^2(1) = .68$, $p = .408$.

The exploratory model is summarised in Table 5.8. As in the English study, the coefficient for Response Speech Rate is negative, indicating that faster responses tended to be rated as earlier. The coefficient for Question Length is positive, indicating that responses to longer questions tended to be rated as later.

Predictor	Coefficient	Std. Error	χ^2	p-value
Intercept	3.39	.23		
TTT	.001	.0001	42.41	< .001
Response Speech Rate	-.10	.04	4.93	.026
Question Length	.03	.01	4.97	.026
Final Rhyme Duration	.0010	.0006	3.04	.081

Table 5.8 Model coefficient and standard errors and LRT statistics for each predictor (Dutch).

5.4 Discussion and Conclusion

The main objective of this chapter was to investigate whether the relationship between turn-timing and other variables, in this case mainly speech rate, is different when responses to questions are ‘on time’ than when they are early or late. To do so, participants were asked to rate the timeliness of responses in recordings of spontaneous question-response sequences in English and Dutch respectively. The resulting data were used to construct a supplementary perceptual variable for a corpus of question-response sequences. The results corroborated the previous chapter’s conclusion that the overarching relationship between speech rate and turn-timing is *positive* in timely turns, contrary to predictions by the entrainment theories of turn-timing of Wilson and Wilson (2005) and Garrod and Pickering (2015).

A secondary objective of this chapter was to analyse the rating data from a different perspective, and test whether turn-transition time perception is affected by the foregoing speech rate, as is the case for many other durational phenomena in speech, such as VOT or vowel length contrasts (Bosker & Reinisch, 2017; Diehl et al., 1980; Dilley & Pitt, 2010; White et al., 2022). Both the effect of global utterance-level speech rate and local final syllable rhyme duration were tested. Surprisingly, no effect of speech rate normalisation appeared to be present in either the Dutch or the English study. Instead, there seems to be evidence for an opposite effect of the duration of the question-final syllable rhyme: in English, a longer total span of final rhyme and TTT was associated with responses being rated as later.

5.4.1 *Task salience and task demands*

This section first discusses the degree to which the rating task successfully accessed the phenomenon of interest: the perceived ‘timeliness’ of conversational turns. The two perception studies operationalised the timeliness of responses in Dutch and English question-response sequences by asking participants to make an explicit judgement of response timeliness on a 7-point scale. Some previous studies have used paradigms with similar designs, either by asking participants for implicit rather than explicit judgements on a scale (e.g., Roberts et al., 2011; Van Os et al., 2020) or by asking expert annotators for explicit binary judgements of timeliness (Stivers et al., 2009). To discuss whether the combination of these aspects succeeded in accessing perceptions of timeliness that are shared to a reasonable degree between listeners, Sections 5.3.1 and 5.3.2 assessed the mean standard deviation of ratings per stimulus. These indicate that there is a degree of disagreement between participants, partially due to different uses of the scale, though there is also an identifiable central tendency in ratings per stimulus.

In the English study, there also seemed to be an effect of the order of presentation on participants’ ratings of the timeliness of responses in the stimuli. Responses were rated as ‘late’ more often in stimuli that were presented toward the end of the study session. It is not clear whether this association indicates response bias, habituation to the speakers’ speech, or another effect. Although presentation order could be accounted for in the analyses of the raw data in Sections 5.3.5 and 5.3.6 by including random effects in the models, further studies should clarify whether this effect is either intrinsic to the phenomenon of perceiving response timing because it is due to habituation to the

speakers, or the result of the task demands because it is due to response bias. Particularly if it is the latter, the effect should be minimised in the design of the study, when the objective is to use a summary statistic of ratings as a variable in a corpus analysis.

In an exploratory analysis of the raw rating data in Sections 5.3.5 and 0, we also tested if and how the length of the question, and the length and the speech rate of the response, were associated with participants' ratings of the responses, to explore potential task effects. In both Dutch and English, a negative relationship between Response Speech Rate and Rating was observed. This relationship was contrary to a usual speech rate normalisation effect: faster responses tended to be rated as earlier given a particular TTT, not later. This effect could indicate that participants' perception of response timing was affected by a general impression of speed.

In both languages, Question Length had a positive effect on Rating, so responses to longer questions were rated as later. This effect could similarly indicate that participants' perception of response timing was affected by a general impression of lateness. It could be due to the overhearer paradigm: during longer questions, participants had more time to anticipate what the response might be but had to wait longer for the actual response. In an overhearer paradigm with stimuli from spontaneous conversations, such effects cannot be controlled, so further studies could aim to replicate any of the effects found under these conditions with experimentally controlled stimuli.

Sections 5.3.1 and 5.3.2 also discussed participants' reports on how salient the response timing was to them as a feature of the stimuli. Although most participants reported finding the timing of responses at least somewhat noticeable, many also said that they mostly found extremely early or extremely late responses noticeably different from the rest. One way to interpret this level of disagreement and these reports is by concluding that a 7-point scale is too fine-grained. Perhaps a smaller scale, or in fact a binary or ternary judgement of timeliness, more economically captures listeners' discrimination of timely and non-timely turns. If so, a narrower scale can be applied when the objective is to understand any differences in behaviour between timely and early or late turns.

5.4.2 *Speech rate and turn-timing in timely turn transitions*

In line with this interpretation, Section 5.3.4 split the corpus data into three subsets. The three subsets corresponded to those rated as early, timely and late, defined as the lower third, middle third and higher third of the scale of mean of the standardised ratings. The relationship between the speech rate of the question and TTT was analysed in each of these subsets separately. Question Speech Rate only significantly predicted TTT in the timely subset, and its effect was positive: faster questions were responded to more slowly, when the response was timely. This effect is the same as that observed in Chapter 4, and contrary to the predictions of the coupled oscillator models. It appears that the expected negative relationship is not obscured by turns that are early or late due to a dysfluent response, or whose timing carries a marked pragmatic meaning. Instead, these results suggest that the positive relationship between speech rate and turn timing is in fact driven by timely turn switches.

When this analysis was repeated for only those observations with gaps (i.e., positive TTT), in order to test the additional effect of final syllable rhyme duration and its interaction with the speech rate of the question, none of the predictors significantly predicted TTT. Question Speech Rate was marginally significant in the Timely subset, and the interaction between Question Speech Rate and Final Rhyme Duration was marginally significant in the Early subset. No variables were significant in the Late subset. Where present, these marginal effects went in the same direction as observed in Chapter 4: the interaction between Question Speech Rate and Final Rhyme Duration was negative, and the main effect of Question Speech Rate was positive.

5.4.3 *Speech rate normalisation*

Sections 5.3.5 and 5.3.6 analysed the rating data from a different perspective, mainly to test whether there was a speech rate normalisation effect on participants' perception of response timeliness. However, there was no significant effect of Question Speech Rate on ratings in either language. Such an effect was also not supported by the polynomial lines of best fit in Figure 5.13 and Figure 5.14.

5.4.4 *Turn-final segments as part of the turn transition*

The analyses in Sections 5.3.5 and 5.3.6 also tested whether there was a local rate normalisation effect, which would be supported by a negative association between Final Rhyme Duration and Ratings: a given response would be expected to be perceived as

earlier when the preceding final rhyme is longer than when it is shorter. Such an association between Final Rhyme Duration and Ratings was not observed. On the contrary, in the English data there was a positive interaction between TTT and Final Rhyme Duration. When the product of TTT and Final Rhyme Duration increased, responses were rated as later.

In the Dutch data, this effect was not replicated. A positive main effect along the same lines was observed for Final Rhyme Duration, but this effect was only marginally significant ($p=.081$) and not supported by the polynomial line of best fit in Figure 5.14. This difference may be due to a difference between how listeners of the two languages interpret turn timing as a segment of continuous speech, but the Dutch study also had fewer stimuli and may therefore have had lower sensitivity to detect an effect. This apparent difference between the two languages should ideally be replicated.

The observations regarding English suggest that listeners paid attention to the final syllable rhyme duration as well as the TTT when they judge response timeliness. In other words, perhaps turn-timing – at least from the perspective of the listener and in the case of gaps – is not best understood as the silence between speakers, but as the composite span covering at least the final syllable rhyme duration and the silence between speakers.

This unexpected effect fits with some observations regarding within-speaker pauses. Firstly, listeners' perception of within-speaker pauses in speech do not always appear to correspond to the presence of acoustic silences directly, since some silences are not perceived as pauses – plosive closures, for example. Conversely, some pauses are perceived as present even when there is no corresponding acoustic silence (Warren, 1985). Similarly, the results presented in this chapter indicate that listeners' perceptions of *between*-speaker pauses – positive turn-transition times, in other words – also do not correspond to the presence of acoustic silence between speaker turns only.

Warren (1985) also observed that a lengthened final fricative before a silence increases the likelihood of a pause being perceived. Similarly, Carlson et al. (2006) describe the effect of what they term a 'compensatory pattern' of final lengthening and silence on listeners' perception of hesitations in speech: it is the total duration of the pre-silence syllable rhyme and the silence itself that best corresponds to perceptions of hesitation in

their study, rather than the duration of each segment separately. A similar compensatory relationship may hold between final lengthening and between-speaker silences for the perception of between-speaker pauses, i.e. for the perception of turn-timing.

Interestingly, however, Warren (1985) notes that the effect of final lengthening appears to depend on listeners' expectations of the phrasing of the utterance. He speculates that when a pause is expected either due to its default syntactic interpretation or an information-structural interpretation that increases the likelihood of a pause, then there is a compensatory effect of final lengthening, but when the pause is not expected, only silence affects listeners' perception. Warren's conclusions regarding this effect are speculative, since his experimental design used a single utterance as stimulus and did not explicitly control for information-structural expectations. It is also not clear if and how this effect would take hold in the case of between-speaker pauses, and particularly how it would interact with variations in turn timing between preferred and dispreferred responses.

Warren (1985) additionally found that the compensatory relationship of final lengthening on pause perception only holds for relatively short pauses (see also Fant & Kruckenberg, 1996). In his study, a lengthened final fricative increased the likelihood of pause perception compared to a short final fricative, but only when it was followed by a silence of around 150 ms or less. Further studies could ask, first, if a similar threshold holds in turn timing perception, and second, if that threshold corresponds to a change in interpretation of turn timing.

This interpretation of the unexpected positive effect of Final Rhyme Duration as indicating a compensatory relationship between final lengthening and silence in between-speaker pause *perception*, is also in line with recent findings regarding turn-timing *production*. Specifically, Franich (2023) observed that the duration of the final word and the between-speaker pause appear to stand in a proportional relationship in question-response sequences from yes-no guessing games. They conclude that the between-speaker pause tends towards occupying either one-third or two-thirds of the total span of the final word and the between-speaker pause. They point out that a similar effect has been observed for within-speaker pause production (Fant & Kruckenberg, 1996). Chapter 6 returns to potential implications of this connection between production and perception.

Note that it might not strictly be the final syllable rhyme that stands in a compensatory relationship with the between-speaker silence. The results in this chapter would also be observed if the compensatory segment is one that correlates with the duration of the final syllable rhyme. Candidates are the final stress foot or the final word. The analyses in this chapter did not distinguish between those possibilities.

It should also be noted that the studies presented here were essentially observational, since the predictors of interest – Question Speech Rate, Final Rhyme Duration, and TTT – were not manipulated to test their individual effects but presented as they occurred in spontaneous speech. These variables are correlated (see Section 4.3), so these effects could be produced by a different underlying relationship. This correlation could also have inflated the variance explained by each predictor, causing false positives. A perception experiment that systematically varies the between-speaker silence and final lengthening is needed to verify these observations.

5.4.5 Conclusion

The main purpose of the studies presented in this chapter was to test whether the relationship between speech rate and turn-timing depends on the timeliness of turns. These studies operationalised turn timeliness as the mean of standardised participant ratings and used the resulting timeliness variable in a corpus analysis on the relationship between speech rate and turn-timing in spontaneous question-response sequences. The results of this analysis support the conclusion of Chapter 4: the relationship between speech rate and turn-timing in spontaneous turn-taking is generally positive, not negative, contrary to the predictions of the entrainment theories. Faster questions tend to be responded to later. This relationship does not seem to be due to instances of mis-timed turns or turns that express a particular meaning through delayed timing, since the positive relationship holds specifically for timely turns.

This chapter also asked whether turn-timing perception is subject to speech rate normalisation, even if there seems to be no negative relationship between speech rate and turn-timing behaviour. An analysis of the rating data did not provide any evidence for this prediction. To the contrary, there appears to be a compensatory relationship between the duration of the turn-final syllable rhyme and turn-transition time, at least when turns do not overlap: a response after a longer total duration of the final syllable rhyme and the between-speaker pause was perceived as later. In other words, listeners

appear to listen to more than just the between-speaker silence to interpret turn-timing. This effect could indicate that the usual operationalisation of turn-timing as the time between the end of one turn to the start of the next should be re-examined. The next chapter discusses this idea in more detail.

Chapter 6. Discussion and Conclusion

6.1 Introduction

This thesis presented three studies on conversational turn timing that aimed to 1) contribute to statistically characterizing the degree of precision and variation in turn timing production and 2) test predictions on the role of speech rate variation in turn timing production and perception. This chapter first briefly summarizes the background to these studies in Section 6.2 and the main findings and limitations of each study separately in Section 6.3. Section 6.4 then discusses general implications of the results.

6.2 Summary of Context

Conversational turn timing refers to the time speakers take to initiate their turns during conversations. It is often measured as the time between the end of one speaker's turn and the start of the next speaker's turn, the *turn-transition time* (TTT). The interpretation of TTT depends on how exactly it is operationalised. For example, if, on the one hand, it represents the time between the start of a turn and the end of the turn the speaker apparently intended to respond to, we can think of it as an aspect of speech timing. On the other hand, if it is measured as the offset between two stretches of speech activity, we can think of it as an aspect of conversational coordination: how well speakers avoided gaps and overlaps (see Section 2.2 for an example of this difference).

Turn timing is often presented as precise (e.g., Holler et al., 2016; M. Wilson & Wilson, 2005, and many others). However, how the exact extent of precision in turn timing should be characterised is not clear, since there are often small gaps and overlaps between turns, and variability in turn timing relates to a wide range of factors, from physiological to social (e.g., Howes & Lavelle, 2023; S. G. Roberts et al., 2015; Sørensen et al., 2021). Some of that variation is due to intentional differences in pragmatic meaning associated with turn timing (F. Roberts et al., 2011), and variation in turn timing can also be associated with disfluency (Van Os et al., 2020).

Models of turn timing aim to account for how speakers manage to achieve precision timing during conversation. One type of model are the *coupled oscillator models* (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005), which hypothesise that a listener's auditory neural activity entrains to the syllable rate of the speaker and that this oscillatory activity in turn constrains the time speakers initiate their speech through a perception-action

link. As a result, they predict that speech rate and turn timing are related – faster turns are expected to be responded to faster than slower turns (the perception-action entrainment hypothesis). Since some variation in turn timing is salient to interlocutors and related to pragmatic interpretations, we might furthermore expect that this relationship is modulated by how ‘on time’ turns sound to listeners.

6.3 Summary of Findings

6.3.1 Study 1: Turn timing in conversation: a systematic review

The first study aimed to contribute to the discussion on the scope of variation in turn timing. It did so through a qualitative synthesis of 41 turn timing samples from 25 studies collected through a systematic review. We compared reported measures of central tendency and the shape of distributions both across all included studies, and within more homogeneous subsets. We also attempted to determine whether any variation could be attributed to the characteristics of the study or sample, such as the operationalisation of turn timing or aspects of the conversational setting.

The included samples were not normally distributed: they were often right skewed and tended to have a single high peak and thick tails (i.e., they were leptokurtic). Moreover, the distributions tended to have a heavy shoulder, often, but not always, to the right of their peak. Although the distributions did not quite conform to Wilson and Wilson’s (2005) prediction that turn timing distributions should be bimodal with peaks before and after 0ms TTT, this pattern does suggest clustering. Qualitative analysis could clarify whether there are functional features of turn switches that distinguish these clusters.

With regard to turn timing precision, Chapter 3 found that measures of central tendency overall clustered between 0 and 500ms and ranged between -812 and 3300ms. The reported modes were much more narrowly distributed, with an inter-quartile range between 0 and 171ms, close to the often-cited 200ms (Stivers et al., 2009). The ranges were narrower in more homogeneous (but also smaller) subsets of samples.

Most of the variation within these ranges could not be explained by study or sample characteristics. We found no clear predictors of differences between estimates. Samples from the same language did not cluster together and samples with only question-response sequences were not noticeably different from samples of all dialogue acts. The means, but not medians and modes, of task-oriented conversations were distributed

more widely than the means of spontaneous conversations, possibly because they more frequently involve specific pragmatic contexts that allow for particularly early or late timing. However, this pattern only held in the most homogeneous subset, which included only nine samples. It may not hold in less homogeneous subsets because the added variability between samples introduced noise, or because it was spurious.

One implication of this range of measures of central tendency, and the absence of consistent effects, bears on Meyer et al.'s (2018) proposed heuristic for testing the ecological validity of turn taking experiments. As Section 2.3.6 discussed, Meyer et al. (2018) suggest comparing the measures of central tendency of response times in experiments to measures of central tendency from spontaneous conversations (specifically, from the same language and dialogue acts). When these values diverge, the experiment may not have successfully captured ecologically valid turn timing behaviour. The ranges and variability in the study in Chapter 3 suggest that simply comparing mean response times to a measure of central tendency of TTT from a study on spontaneous conversation (e.g., Stivers et al., 2009, as Meyer et al. do) is perhaps too strict.

This study used relatively strict inclusion criteria; we excluded any report that did not explicitly define its operationalisation of turn timing. However, most included studies still did not provide explicit criteria for how they identified turn boundaries, meaning that this criterion resulted in a somewhat arbitrary selection. Future studies of secondary data could use a more permissive application of the same criteria, although that would increase noise in the data, or use citation chaining as an additional search strategy to identify a larger selection of eligible samples. Including a wider range of data types, such as raw data obtained from authors, or proportional data, could support a more detailed analysis of differences between distributions. Meanwhile, future studies of primary data could aim to use operationalisations of turn timing that are specific on a pragmatic, lexical, phonetic and acoustic level.

6.3.2 Study 2: Speech rate and turn timing in spontaneous question-response sequences

The second study tested the perception-action entrainment hypothesis in a corpus of spontaneous question-response sequences from dyadic conversations in Dutch and English, using manual measurements of turn timing and speech rate, and mixed effect linear regression modelling. It tested the effect of the question's speech rate and its final

syllable's rhyme duration on the timing of the response. It also looked at the effect of question- and response type and length, the speech rate of the response, and the language.

We found that the question's speech rate (measured as syllable per second) was *positively* associated with TTT. In the subset of data where TTT was positive, we also found that a longer final syllable rhyme was associated with a later TTT, and that the interaction between speech rate and final rhyme duration was a marginally significant predictor of TTT – a larger contrast between the length of the final syllable and the previous syllables was associated with earlier responses.

Using our heuristics for variable importance – retesting variables, comparing AIC, and partial regression plots – we also found that these effects were small compared to other significant predictors, such as the type of question, the type of response, and the length of the question.

This speech rate effect is furthermore contrary to the perception-action entrainment hypothesis: Not only did we find no evidence for the expected negative relationship between speech rate and TTT, but we found evidence that the contrary relationship holds. Section 6.4.1 reviews potential explanations for this effect and its implications.

This study assumed that an entrainment effect can be detected using linear statistical methods, because most turns were expected to be initiated on the first cycle of the hypothesised coupled oscillator model. Although the opposite direction of the effect of speech rate in this dataset suggests that this assumption did not disguise any effect of entrainment, since rather than no effect, we found the opposite effect than hypothesised, further studies could use non-linear modelling to account for a potential cyclical structure of TTT. The Study 3 also accounted for this possibility to some extent by considering perceptible delays. We also used only one operationalisation of turn timing, and did not look at possible effects of including, for example initial inbreaths or clicks (cf. Kendrick & Torreira, 2015, also, see Section 6.4.2).

6.3.3 Study 3: Speech rate and turn-timing perception: a rating study of perceived response timeliness

The third study distinguished turn switches that are 'on time' from those that are timed earlier or later. In two perception studies, in Dutch and English, participants were asked

to rate the timeliness of responses in a subset of the question-response sequences from the previous Chapter. Their responses were aggregated per stimulus and used as a perceptual variable, Mean Standardised Rating, denoting ‘timeliness’ in a further corpus analysis of the relationship between speech rate and turn timing.

When turn switches that were ‘on time’ were distinguished from early and late turn switches categorically, by splitting the range of ratings in three categorical subsets, we found that the speech rate of the question only significantly predicted turn timing when the responses were perceived as ‘on time’. The relationship between the speech rate of the question and TTT was again positive, contrary to expectations: faster questions were responded to slower, not faster. These results corroborate the previous chapter’s observations, which were contrary to the perception-action entrainment hypothesis. Moreover, they additionally suggest that the inverse effect we observed is not driven by a few turn switches with particularly marked timing but rather holds specifically for those that are on time.

A secondary analysis also tested what features of each question-response sequences are predictive of participants’ ratings of stimuli with gaps (i.e., $TTT \geq 0\text{ms}$). We expected to see an effect of speech rate normalisation (Diehl et al., 1980): a particular TTT was expected to be perceived as longer after a fast question rather than a slow question. This effect was not observed in either Dutch or English. Instead, we found that in English, a longer total span of the question-final syllable rhyme + TTT was associated with later ratings, rather than just TTT. In Dutch, this effect was not strongly present: there was a trend towards a positive effect of the final syllable rhyme’s duration on the ratings, and TTT by itself positively predicted ratings. Unsurprisingly, longer TTTs were rated as later than shorter TTTs.

This unexpected effect of the final syllable rhyme in English may be due to the relationship between final rhyme duration and TTT rather than a true effect on perception because those two variables were correlated in the spontaneous data (see Chapter 4), so an experimental study in which they are systematically varied should be used to test whether this effect holds consistently. If so, it suggests that participants paid attention to more than just TTT when judging timeliness. This observation is in line with previous observations on the role of pre-pause segment lengths for within-speaker pauses (Warren, 1985).

Further studies could also aim to experimentally control some other observed effects, that may have been due to the task demands of the study. Specifically, we found that responses to longer questions, faster responses, and later trials (in English) were rated as later. Additionally, further studies could aim to establish whether it is the turn-final syllable rhyme, or another turn-final segment, that best fits with participants' perceptions of turn timing.

6.4 Implications

This section discusses general implications of the results. Section 6.4.1 discusses consequences and further questions for models of turn timing. Section 6.4.2 reviews the observations on the role of the turn-final syllable rhyme in perception and production of turn timing and connects them to observations from the literature. Section 6.4.3 revisits the topic of how turn timing is measured.

6.4.1 Models of turn timing

Coupled oscillator models of turn timing (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005) predict that faster speech rates should be associated with shorter TTTs. As Section 2.3 reviewed, previous studies found contrary evidence regarding this prediction. Experiments (Corps et al., 2020), and other studies that strictly controlled the content of turns in verbal exchanges (Franich, 2023; Vranjes & Oben, 2022), found evidence in favour: faster turns were responded to earlier. Observational studies of spontaneous conversations, on the other hand, found no evidence (Kendrick & Torreira, 2015), weak evidence (Beňuš, 2009), or evidence to the contrary (S. G. Roberts et al., 2015). Specifically, Roberts et al. (2015) found that speech rate is the fourth most important predictor of turn timing of 32 predictors using a random forest analysis, which does not explicate the relationship between the variables. They also observed that fastest quartile of turns was followed by a TTT that was 100ms longer than the slowest quartile of turns, if they restricted the turn sequence context by excluded backchannels and short turns.

The studies in Chapter 4 and 5 of this thesis tested whether there is a relationship between speech rate and turn timing in question-response sequences, the dialogue act studied in the experimental studies, from spontaneous conversations. In this restricted dialogue act context, speech rate was *positively* related to turn timing, contrary to the hypothesis: faster questions were responded to later, particularly when TTT was positive, and when the response was perceived as 'on time'. These studies therefore align with

Roberts et al. (2015) and suggest that the entrainment observed by the experimental studies does not reflect spontaneous conversational behaviour. This conclusion is further supported by the observation that the positive relationship between speech rate and turn timing specifically held when the questions were polar questions rather than open questions, notwithstanding that the experimental studies used specifically a polar question-response paradigm (Corps et al., 2020; cf. Franich, 2023).

There are several possible explanations for this positive relationship between speech rate and turn timing. First, Levinson and Torreira (2015) suggest, in response to their observations in Roberts, Torreira and Levinson (2015) that the fastest quantile turns were responded to earlier than the slowest quartile, that speakers have less time to prepare their response when responding to a faster question. If so, we might have expected an interaction between the speech rate of a turn and its length in syllables because longer turns also give interlocutors more time to prepare a response. However, Study 2 in Chapter 4 did not observe such an interaction. Relatedly, Roberts et al. (2015) also suggest that faster turns might be responded to later because faster speech increases processing load on the listener.

Second, this effect could also be due to the way a turn's speech rate interacts with cues to turn timing. Specifically, in slower turns, turn ending cues can appear earlier (all else being equal, cf. Torreira & Bögels, 2022), allowing speakers more time to respond. The slower speech rate of turns that were responded to faster could also be an artefact of a cue to turn ends, namely a gradual slowing down towards the end of a turn (Rühlemann & Gries, 2020). If so, overall slower turns may be slower because they slow down, which signals to the interlocutor that a turn end is coming up and gives them more time to prepare their turn.

Further studies could clarify which of these possible explanations is most likely, for example by observing the relative position of turn ending cues and how they relate to speech rate and turn timing in spontaneous conversational data, or by measuring the relative salience of turn ending cues. Another possibility is to systematically vary speech rate against its relationship to turn ending cues in a semi-spontaneous production experiment, in which participants think they are engaged in a spontaneous conversation with no eye-contact, but some target turns of a confederate are pre-recorded and manipulated (for example, as in Bögels, 2020).

Interestingly, experimental studies (Corps et al., 2020; Torreira & Bögels, 2022) demonstrate that a speech-rate entrainment responding mechanism is *possible*, but it does not seem to be the best fit for observed patterns in spontaneous behaviour. The present studies found no evidence that this contrast between experimental studies and spontaneous conversations is due to the restricted set of dialogue acts (polar question-response sequences) included in experimental paradigms or the responding mechanism that such dialogue acts evoke. The contrast might indicate that an entrainment-based responding mechanism may be overshadowed in spontaneous conversation by other effects on speech processing (Torreira & Bögels, 2022) or because it is simply not recruited by speakers during conversation.

Further studies could focus on clarifying in what kind of vocal exchanges, beyond experiments, entrained responding behaviour is active and detectable. A possible avenue of research is to track the relationship between speech rate and turn timing along language development. If entrained responding is active during conversation but overshadowed by other aspects, we might expect that young infants respond earlier to fast turns in adult-infant interactions, but that they lose this entrainment effect during language acquisition.

6.4.2 The role of turn-final segments in perception and production

This section compares evidence on turn final segments, such as the turn final syllable or word, and discusses their potential role in turn timing production and perception.

As reviewed in Section 2.3.6, the duration of turn final segments of the previous turn could have several different effects on the timing of the subsequent turn. First, lengthening of final syllable(s) as a cue to utterance endings is well-established cross-linguistically (Paschen et al., 2022; Price et al., 1991; White, 2014), including turn endings (Brusco et al., 2020; Duncan, 1972). If interlocutors time their turns at least in part by reacting to turn final cues, a more salient (presumably longer) final lengthening cue might be associated with earlier TTT. This implies that turns are generally initiated late enough to be in principle a reaction to the moment final lengthening becomes salient.

Second, coupled oscillator models of turn timing can imply two contradictory predictions about the relationship between turn-final segments and turn timing, depending on how locally the phase of the oscillators can be reset. On the one hand, if the hypothesised

entrained oscillator can be reset locally by the duration of a single syllable, a longer final syllable should be associated with later turn timing (Corps et al., 2020). On the other hand, if the oscillator's phase is reset by a point in the onset of syllables (Bosker, 2017a; Oganian & Chang, 2019), then a lengthened final syllable can no longer influence turn timing through the hypothesised perception-action link. If so, we would expect that relatively lengthened final syllables would be associated with earlier turn timing (relative to the end of the final syllable), because they take up more of the phase of their cycle of the oscillatory system: the final syllable and (positive) TTT should stand in a compensatory relationship.

Some previous studies have observed such a compensatory relationship between the duration of a turn-final segment and turn timing in production. For example, Magyari et al. (2017), in a an experimental question-response paradigm, found that longer final words were associated with shorter response times. Similarly, Corps et al. (2020) also found that a non-speeded final word was associated with shorter response times than a speeded final word, when the response time was measured from the end of the final word (one of their two operationalisations, see Section 2.3.5). In a yes-no-guessing game, Franich (2023) also found that longer turn-final words were associated with earlier turn timing.

To the contrary, the results of Chapter 4 indicated an additive effect of the final syllable rhyme: Longer question-final rhymes were associated with longer TTT in Gaps. Similarly, when measuring response time from the *start* of the final word, Corps et al. (2020) found that speeded final words were associated with shorter response times than non-speeded final words.

It is not clear what could explain this pattern of results, other than local resetting of an entrained oscillatory perception action link, for which this study found no independent evidence in spontaneous conversation. One possible, but very speculative, explanation relies on another observation by Franich (2023). In addition to finding that longer question-final words were associated with longer response times, they observed that between-speaker pauses (i.e., positive TTT) tended to occupy one-third or two-thirds of the total duration of the final word and the between-speaker pause in their data. The question-final syllable tended to occupy three-fourths or half of the total duration of the final syllable and the between-speaker pause. If this observation represents a consistent

pattern, then the contrasting results regarding the relationship between turn-final word or syllable durations and turn timing could be driven by this effect. However, it is not clear how interlocutors manage to coordinate their timing to co-construct this pattern, or due to what speech timing mechanism this pattern would arise. It implies a coupling mechanism between interlocutors during conversation that is similar to, but not identical to, the coupled oscillator models.

Study 3 (Chapter 5) also observed that the final syllable rhyme duration may have a role in turn timing *perception*. Specifically, a longer total span of the final syllable rhyme duration and TTT was perceived as ‘later’ in the English study, and there was a trend to a similar effect in Dutch. This effect is similar to Warren’s (1985) observation on within-speaker pause perception: They found that a lengthened final fricative before an acoustic silence increased the change of a pause being perceived (cf. Carlson et al., 2006), specifically for short and syntactically expected pauses.

The effects observed in Chapter 5 should ideally be replicated. However, they do suggest that, at least from the perspective of a third-party listener (i.e., the participant in the overhearer paradigm of Chapter 5), the time between the end of one turn and the start of the next may not be the communicatively relevant duration, but other speech segments should be taken into account. If so, we could also question whether it is the most useful operationalisation of turn timing from the perspective of production. The next section discusses this issue further.

6.4.3 Measuring turn timing

As Section 2.2 reviewed, turn timing is most commonly measured as the time between the end of one turn and the start of the next. A version of this operationalisation was used by all studies included in the systematic review in Chapter 3, as well as by the studies in Chapters 4 and 5 of this thesis. However, it is an assumption that this duration best captures all relevant aspects of turn timing.

This assumption can be challenged. For example, as Section 2.2 discussed, Kendrick and Torreira (2015) point out that the *start* of a turn is a complex phenomenon. Using a mixed-methods approach, they tested whether preferred responses are timed earlier than dispreferred responses, as intuition and qualitative studies suggest. They demonstrate that it depends on how the start of the turn is identified: When turn-initial clicks, inbreaths

and filled pauses are considered part of the second turn, dispreferred responses are timed earlier, but when these elements are considered part of TTT, they are timed later (see Section 2.2).

With regard to the end of the previous turn, the usual operationalisations of turn timing as TTT implies that the end of the previous turn adequately represents the speaker's timing target. This view is not self-evident, since it is not a priori clear what speakers are aiming for when they initiate their turns – it seems informed by Sacks et al.'s (1978) no-gap-no-overlap principle. However, identifying the true timing target, from a psycholinguistic perspective, is also not straightforward.

To my knowledge, only few studies used operationalisations of turn timing that do not imply that the end of the previous turn is the relevant point: First, Corps et al. (2020) measured turn timing from the start of the turn-final word in a question-response experiment to account for the extra preparation time that longer turn-final words afford (see Section 2.3.5). Second, Harrison et al. (1978) measured it from the moment speakers saw an experimental cue to indicate whether they should lie or not during their upcoming turn. Third, one of two operationalisations used by Beňuš (2009) measured turn timing as the time between energy peaks associated with pitch accents. Finally, to account for their observation that shorter final words were associated with longer response times in a question-response experiment, Magyari et al. (2017) considered a number of possible timing targets, including the start of the final word, stressed vowel, pitch peak, and the moment the final word could be identified. They rejected that any of these alternative measurements are better operationalisations of response time, on the grounds that the duration of the remainder of the question *after* each of those points also still effected response times.

From a theoretical perspective, prediction-reaction models may also imply that the end of the turn is not the most accurate point from which to measure turn. Specifically, models of turn timing that propose that a prepared response is released from a buffer in reaction to turn ending cues (Levinson & Torreira, 2015) or in reaction to reaching a threshold of certainty that the turn is ending (Barthel, 2020) imply that the relevant timing target is the relevant cue or certainty point. However, Levinson and Torreira (2015) and Barthel (2020) still measure turn timing from the end of the previous turn.

A practical reason for doing so may be that identifying an alternative timing target is not straightforward. Especially if we assume that turn timing is reaction time to the end of a cue, there are many potential variable parameters: Turn endings are often associated with multiple turn-ending cues, which may not all be always necessary or sufficient (Gravano & Hirschberg, 2011). Speakers may also differ, even on a case-by-case basis, in which cue they respond to, and there are also many potential factors that could affect their subsequent reaction time (cf. S. G. Roberts et al., 2015). Testing which point best corresponds to the speaker's timing target requires systematically controlling all these unknown factors. However, without testing this assumption, we risk that our measurements of turn timing are noisy when testing hypotheses regarding the factors that affect turn timing.

An alternative approach to identifying how to measure turn timing is through perception studies. Although this approach still does not conclusively tell us when the speaker aimed to start speaking, it allows us to identify what duration conversationalists use to convey turn-timing related meaning. The unexpected results of the English perception study in Chapter 5, for example, suggest the start of the turn-final syllable's rhyme as a potential candidate.

The study in Chapter 5 is not conclusive, and it did not consider any other durations other than TTT and the final syllable rhyme. However, in the next few paragraphs I will consider two implications of measuring turn timing from the start of the turn-final syllable rhyme to the start of the next turn (FR+TTT).

First, if FR+TTT accurately captures turn timing production, because speakers aim to start speaking after the start of the final syllable, the usual argument that turn timing is often faster than minimal vocal reaction times of around 200ms (see Section 2.2.3) loses some of its support. For example, the modal TTT of the Gaps dataset in Chapter 4 is 390ms but the modal FR+TTT is 510ms (with bin size = 30ms, and excluding all cases of overlap).

Second, we might expect that the predicted negative relationship between the speech rate and turn timing holds between speech rate and FR+TTT, because TTT alone would be sensitive to noise caused by final lengthening. Note that FR+TTT corresponds more closely to one of the operationalisations used by Corps et al. (2020; see Section 2.3.6), and it might also better capture turn timing production than TTT if we assume, 1) that the

coupled oscillator models are correct and 2) that local phase-resetting of the oscillator is not possible.

However, this speculation is not supported by a re-analysis of the Gaps data from Chapter 4: A model predicting FR+TTT from Question Speech Rate, with a random intercept for Speaker, did not significantly better fit the data than a null model, $\chi^2(1) = .08$, $p = .773$, so re-analysing turn timing as FR+TTT does not provide support for the perception-action entrainment hypothesis.

In short, the usual operationalisation of turn timing as TTT rests on the assumptions that the end of the previous turn adequately represents the speakers' timing target. Perception studies could challenge or support this assumption by testing whether, and what, alternative measurements capture listeners' interpretations of turn timing.

6.5 Conclusion

This thesis presented three studies on the production and perception of turn timing in spontaneous conversations that considered the influence of both pragmatic and speech-timing factors.

Through corpus analyses (Chapters 4 and 5) of question-response sequences from dyadic, spontaneous conversations in Dutch and English, we tested the perception-action entrainment hypothesis (Garrod & Pickering, 2015; M. Wilson & Wilson, 2005), which predicts that faster questions should be associated with shorter turn-transition times (TTT). We observed the opposite pattern: faster questions were responded to after longer positive TTT. This effect was subsidiary to the influence of question- and response type on TTT but held specifically when turns were perceived as 'on time' by participants of a rating study (Chapter 5), as opposed to early or late turns. Combined with previous experimental studies, these results indicate that while an entrainment-based responding mechanism may be possible in experimental contexts (Corps et al., 2020; Torreira & Bögels, 2022), it is not active or visible in spontaneous conversation behaviour.

In a systematic review of 41 turn timing samples (Chapter 3), we found that TTT distributions tended to be unimodal, leptokurtic, and right-skewed, often with a heavy right shoulder. The measures of central tendency of TTT distributions clustered between 0ms and 500ms. Besides potentially the type of conversation, there were no consistent predictors of variation. Some of this unexplained variation may be due to unknown

differences in the ways turn timing was measured, since most included studies did not specify the acoustic or phonetic criteria by which they identified turn boundaries.

In a secondary analysis of participants ratings of the timeliness of turns (Chapter 5), we also observed that the duration of the turn's final syllable rhyme affected participants' perception of turn timing. Responses were perceived as later after a longer total duration of the final syllable rhyme and TTT. This effect suggests that TTT alone may not best capture turn-timing *perception*. It also suggests that perception studies could provide a method for identifying perceptually validated operationalisations of turn-timing *production*.

In short, these studies found no evidence for the coupled oscillator models but instead suggest that pragmatic factors (such as dialogue act) are more important in determining turn timing than speech rate. They also suggest a need for perceptually validated operationalisations of turn timing.

References

- Ahn, S., Cho, H., Kwon, M., Kim, K., Kwon, H., Kim, B. S., Chang, W. S., Chang, J. W., & Jun, S. C. (2018). Interbrain phase synchronization during turn-taking verbal interaction—A hyperscanning study using simultaneous EEG/MEG. *Human Brain Mapping*, 39(1), 171–188. <https://doi.org/10.1002/hbm.23834>
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory.*, 276–281.
- Allan, D. W. (1966). Statistics of atomic frequency standards. *Proceedings of the IEEE*, 54(2), 221–230. *Proceedings of the IEEE*.
<https://doi.org/10.1109/PROC.1966.4634>
- Alpert, M., Kotsaftis, A., & Pouget, E. R. (1997). Speech Fluency and Schizophrenic Negative Signs. *Schizophrenia Bulletin*, 23(2), 171–177.
<https://doi.org/10.1093/schbul/23.2.171>
- American Psychological Association. (2023). *PsycINFO*.
<https://www.apa.org/pubs/databases/psycinfo>
- Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., & Miller, J. (1991). The HCRC map task corpus. *Language and Speech*, 34(4), 351–366.
- Andrews, R. (2021). *Polyrhythmicity in Conversation and Speech Prosodies* (Vol. 12). Springer. https://doi.org/10.1007/978-981-16-0566-6_5
- Aubanel, V., Cooke, M., Villegas, J., Lecumberri, M. L. G., & Loquendo S.p.A.; Nuance Communications; AT and T Labs - Research; Google Research; Interactive Media S.p.A. (2011). *Conversing in the presence of a competing conversation: Effects on speech production*. 2833–2836.
- Auer, P., Couper-Kuhlen, E., & Muller, F. (1999). *Language in time: The rhythm and tempo of spoken interaction*. Oxford University Press.
- Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science*, 25(8), 1546–1553.
- Barthel, M. (2020). *Speech Planning in Dialogue* [PhD Thesis]. Universiteit Nijmegen.

- 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(MAR).
<https://doi.org/10.3389/fpsyg.2017.00393>
- Basset, P., Amelot, A., Vaissière, J., & Roubeau, B. (2001). Nasal airflow in French spontaneous speech. *Journal of the International Phonetic Association*, 31(1), 87–99.
- Beattie, G. W. (1978). Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology*, 17(1), 7–15.
<https://doi.org/10.1111/j.2044-8260.1978.tb00889.x>
- Becic, E., Dell, G. S., Bock, K., Garnsey, S. M., Kubose, T., & Kramer, A. F. (2010). Driving impairs talking. *Psychonomic Bulletin & Review*, 17(1), 15–21.
<https://doi.org/10.3758/PBR.17.1.15>
- Beňuš, Š. (2009). Are We 'in Sync': Turn-Taking in Collaborative Dialogues. *Tenth Annual Conference of the International Speech Communication Association*, 2167–2170.
- Bilmes, J. (1988). The concept of preference in conversation analysis. *Language in Society*, 17(2), 161–181.
- Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* (Version 6.1.48) [Computer software]. <http://www.praat.org/>
- Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, 203.
<https://doi.org/10.1016/j.cognition.2020.104347>
- Bögels, S., Kendrick, K. H., & Levinson, S. C. (2020). Conversational expectations get revised as response latencies unfold. *Language, Cognition and Neuroscience*, 35(6), 766–779. <https://doi.org/10.1080/23273798.2019.1590609>
- Bögels, S., & Levinson, S. C. (2017). The brain behind the response: Insights into turn-taking in conversation from neuroimaging. *Research on Language and Social Interaction*, 50(1), 71–89.
- Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General*, 143(1), 295.
- Boland, A., Dickson, R., & Cherry, G. (2017). *Doing a systematic review: A student's guide* (2nd ed.). SAGE Publications Ltd.

- Bóna, J., & Bakti, M. (2020). The effect of cognitive load on temporal and disfluency patterns of speech: Evidence from consecutive interpreting and sight translation. *Target. International Journal of Translation Studies*, 32(3), 482–506.
<https://doi.org/10.1075/target.19041.bon>
- Bořil, T., & Skarnitzl, R. (2016). Tools rPraat and mPraat. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue* (pp. 367–374). Springer International Publishing. https://doi.org/10.1007/978-3-319-45510-5_42
- Bosker, H. R. (2017a). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, 79, 333–343.
<https://doi.org/10.3758/s13414-016-1206-4>
- Bosker, H. R. (2017b). How our own speech rate influences our perception of others. *Journal of Experimental Psychology: Learning Memory and Cognition*, 43(8), 1225–1238. <https://doi.org/10.1037/xlm0000381>
- Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: Behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, 33(8), 955–967.
- Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology*, 8, 1063.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech envelopes. *Neuroimage*, 233, 117958.
- Brosy, J., Bangerter, A., & Mayor, E. (2016). Disfluent Responses to Job Interview Questions and What They Entail. *Discourse Processes*, 53(5–6), 371–391.
<https://doi.org/10.1080/0163853X.2016.1150769>
- Brusco, P., Vidal, J., Beňuš, & Gravano, A. (2020). A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication*, 125, 24–40.
<https://doi.org/10.1016/j.specom.2020.09.004>
- Buanzur, T., Zellers, M., Namyalo, S., & Witzlack-Makarevich, A. (2018). A first investigation of the timing of turn-taking in Ruuli. *Proceedings of the Annual Conference of the International Speech Communication Association 2018, 2018-September*, 621–625. <https://doi.org/10.21437/Interspeech.2018-1254>

- Bull, M., & Aylett, M. P. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. *5th International Conference of Spoken Language Processing (ICSLP'98)*, 1175–1178.
<https://www.research.ed.ac.uk/en/publications/an-analysis-of-the-timing-of-turn-taking-in-a-corpus-of-goal-orientation>
- Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology*, 89(3), 319–325. <https://doi.org/10.1007/s00421-003-0807-5>
- Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in Cognitive Sciences*, 18(6), 300–309. <https://doi.org/10.1016/j.tics.2014.02.005>
- Caldwell, H. B. (2022). Sign and Spoken Language Processing Differences in the Brain: A Brief Review of Recent Research. *Annals of Neurosciences*, 29(1), 62–70.
<https://doi.org/10.1177/09727531211070538>
- Cambridge Scientific Abstracts, Inc. (2023). *Linguistics and Language Behavioural Abstracts*. <https://www.proquest.com/llba>
- Campbell, N. (2007). Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. *Proc XVIth International Congress of the Phonetic Sciences, Saarbrücken, Germany*, 343–348.
- Campbell, N. (2008). Individual traits of speaking style and speech rhythm in a spoken discourse. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5042, 107–120. https://doi.org/10.1007/978-3-540-70872-8_8
- Cangemi, F., Grice, M., Janz, A., Lucarini, V., Spaniol, M., & Vogeley, K. (2023). Content-free speech activity records: Interviews with people with schizophrenia. *Language Resources and Evaluation*, 58(3), 925–949.
<https://doi.org/10.1007/s10579-023-09666-z>
- Carlson, R., Gustafson, K., & Strangert, E. (2006). *Modelling hesitation for synthesis of spontaneous speech*. 3rd International Conference on Speech Prosody, SP 2006, Sweden.
- Chowdhury, S. A., Riccardi, G., & Alam, F. (2014). *Unsupervised Recognition and Clustering of Speech Overlaps in Spoken Conversations*. 62–66.

- Coffey, E. B. J., Arseneau-Bruneau, I., Zhang, X., Baillet, S., & Zatorre, R. J. (2021). Oscillatory Entrainment of the Frequency-following Response in Auditory Cortical and Subcortical Structures. *Journal of Neuroscience*, 41(18), 4073–4087. <https://doi.org/10.1523/JNEUROSCI.2313-20.2021>
- Corps, Knudsen, B., & Meyer, A. S. (2022). Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition*, 223. <https://doi.org/10.1016/j.cognition.2022.105037>
- Corps, R. E., Gambi, C., & Pickering, M. J. (2020). How do listeners time response articulation when answering questions? The role of speech rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(4), 781–802.
- Corrette, R. (2012). *Praat Vocal Toolkit* [Computer software]. <https://www.praatvocaltoolkit.com>
- Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction* (Vol. 25). John Benjamins Publishing.
- Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11(1), 51–62.
- De Boer, J. N., van Hoogdalem, M., Mandl, R. C. W., Brummelman, J., Voppel, A. E., Begemann, M. J. H., van Dellen, E., Wijnen, F. N. K., & Sommer, I. E. C. (2020). Language in schizophrenia: Relation with diagnosis, symptomatology and white matter tracts. *Npj Schizophrenia*, 6(1), 1–10. <https://doi.org/10.1038/s41537-020-0099-3>
- 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.
- De Vos, C., Torreira, F., & Levinson, S. C. (2015). Turn-timing in signed conversations: Coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2015.00268>
- Deppermann, A. (2013). Turn-design at turn-beginnings: Multimodal resources to deal with tasks of turn-construction in German. *Journal of Pragmatics*, 46(1), 91–121. <https://doi.org/10.1016/j.pragma.2012.07.010>
- Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, 27(5), 435–443. <https://doi.org/10.3758/BF03204461>

- Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11), 1664–1670.
- Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, 8.
<https://doi.org/10.3389/fnhum.2014.00311>
- Dingemanse, M., & Floyd, S. (2014). Conversation across cultures. In *The Cambridge handbook of linguistic anthropology* (pp. 447–480). Cambridge University Press.
- Dingemanse, M., & Liesenfeld, A. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, 1, 5614–5633.
- Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, 85, 761–768.
- Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*, 116(20), 10113–10121.
<https://doi.org/10.1073/pnas.1816414116>
- Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, 30, 412–431.
- Donoghue, T., Schaworonkow, N., & Voytek, B. (2022). Methodological considerations for studying neural oscillations. *European Journal of Neuroscience*, 55(11–12), 3502–3527. <https://doi.org/10.1111/ejn.15361>
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283.
- Duncan, S., & Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10(3), 234–247.
- Edlund, J., Heldner, M., & Hirschberg, J. (2009). *Pause and gap length in face-to-face interaction*. 2779–2782.
- Elsevier. (2023). *Scopus*. <https://www.scopus.com/>
- Fadden, L. M. (2008). *Prosodic profiles: Suspects' speech during police interviews*. Simon Fraser University.

- Fant, G., & Kruckenberg, A. (1996). On the quantal nature of speech timing. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 4, 2044–2047. <https://ieeexplore.ieee.org/abstract/document/607202/>
- Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1187–1199. <https://doi.org/10.1037/0278-7393.28.6.1187>
- Finlayson, I., Lickley, R. J., & Corley, M. (2012). Convergence of speech rate: Interactive alignment beyond representation. *25th Annual CUNY Conference on Sentence Processing, New York*.
- Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA*, 105(2), 203–206. <https://doi.org/10.5195/jmla.2017.88>
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (third). Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Franco-Gaviria, F., Amador-Jiménez, M., Millner, N., Durden, C., & Urrego, D. H. (2022). Quantifying resilience of socio-ecological systems through dynamic Bayesian networks. *Frontiers in Forests and Global Change*, 5, 889274.
- Franich, K. (2023). Prosodic predictors of temporal structure in conversational turn taking. *Proceedings of the International Congress of Phonetic Sciences. ICPhS, Prague*.
- Freud, D., Ezrati-Vinacour, R., & Amir, O. (2018). Speech rate adjustment of adults during conversation. *Journal of Fluency Disorders*, 57, 1–10. <https://doi.org/10.1016/j.jfludis.2018.06.002>
- Fry, D. B. (1975). Simple Reaction-Times to Speech and Non-Speech Stimuli. *Cortex*, 11(4), 355–360. [https://doi.org/10.1016/S0010-9452\(75\)80027-X](https://doi.org/10.1016/S0010-9452(75)80027-X)
- Fujiwara, N., Itoh, T., & Araki, K. (2007). *Analysis of changes in dialogue rhythm due to dialogue acts in task-oriented dialogues*. 4629. https://doi.org/10.1007/978-3-540-74628-7_73
- Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00751>

- Giannella Samelli, A., & Schochat, E. (2008). The gaps-in-noise test: Gap detection thresholds in normal-hearing young adults. *International Journal of Audiology*, 47(5), 238–245. <https://doi.org/10.1080/14992020801908244>
- Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension. *Cerebral Cortex*, 32(7), 1405–1418. <https://doi.org/10.1093/cercor/bhab287>
- Ginsparg, P. (2023). *arXiv*. <https://arxiv.org/>
- Girard-Groeber, S. (2015). The management of turn transition in signed interaction through the lens of overlaps. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00741>
- Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous Cortical Rhythms Determine Cerebral Specialization for Speech Perception and Production. *Neuron*, 56(6), 1127–1134. <https://doi.org/10.1016/j.neuron.2007.09.038>
- Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4), 511.
- Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, IEEE International Conference On*, 1, 517–520.
- Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., & Simon, J. Z. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, 77(5), 980–991.
- Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3), 601–634.
- Grömping, U. (2015). Variable importance in regression models. *WIREs Computational Statistics*, 7(2), 137–152. <https://doi.org/10.1002/wics.1346>
- Grosjean, F. (1979). A study of timing in a manual and a spoken language: American sign language and English. *Journal of Psycholinguistic Research*, 8(4), 379–405. <https://doi.org/10.1007/BF01067141>

- Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol*, 11(12), e1001752.
- Haegens, S., & Zion Golumbic, E. (2018). Rhythmic facilitation of sensory processing: A critical review. *Neuroscience & Biobehavioral Reviews*, 86, 150–165.
<https://doi.org/10.1016/j.neubiorev.2017.12.002>
- Harrison, A. A., Hwalek, M., Raney, D. F., & Fritz, J. G. (1978). Cues to deception in an interview situation. *Social Psychology*, 42(2), 156–161.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298(5598), 1569–1579.
<https://doi.org/10.1126/science.298.5598.1569>
- Hayashi, R. (1991). Floor structure of English and Japanese conversation. *Journal of Pragmatics*, 16(1), 1–30. [https://doi.org/10.1016/0378-2166\(91\)90003-G](https://doi.org/10.1016/0378-2166(91)90003-G)
- Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, 130(1), 508–513.
<https://doi.org/10.1121/1.3598457>
- Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555–568. <https://doi.org/10.1016/j.wocn.2010.08.002>
- Heldner, M., Edlund, J., Hjalmarsson, A., & Laskowski, K. (2011). Very short utterances and timing in turn-taking. *Twelfth Annual Conference of the International Speech Communication Association*.
- Hick, W. E. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, 4(1), 11–26. <https://doi.org/10.1080/17470215208416600>
- Holler, J., Alday, P. M., Decuyper, C., Geiger, M., Kendrick, K. H., & Meyer, A. S. (2021). Competition Reduces Response Times in Multiparty Conversation. *Frontiers in Psychology*, 12, 693124. <https://doi.org/10.3389/fpsyg.2021.693124>
- Holler, J., Casillas, M., H Kendrick, K., & C Levinson, S. (2016). *Turn-taking in human communicative interaction*. Frontiers Media SA.
- Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: Optimizing reciprocity. *Frontiers in Psychology*, 6.
<https://doi.org/10.3389/fpsyg.2015.00098>
- Howes, C., & Lavelle, M. (2023). Quirky conversations: How people with a diagnosis of schizophrenia do dialogue differently. *Philosophical Transactions of the Royal*

- Society B: Biological Sciences*, 378(1875), 20210480.
<https://doi.org/10.1098/rstb.2021.0480>
- Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1), 101–144.
<https://doi.org/10.1016/j.cognition.2002.06.001>
- Isard, J. C., Doherty, G. M., & Kowtkoy, S. D. (1993). *Conversational Games Within Dialogue*. Citeseer.
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5aa76cb178eb33fcb634774f2c47b7fd871813a5>
- Itoh, T., Kitaoka, N., & Nishimura, R. (2009). Subjective experiments on influence of response timing in spoken dialogues. *Proceedings of the Annual Conference of the International Speech Communication Association*, 1835–1838.
- Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a ‘standard maximum’ silence of approximately one second in conversation. *Conversation: An Interdisciplinary Approach/Multilingual Matters*.
- Jungers, M. K., & Hupp, J. M. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, 24(4), 611–624.
- Kayser, S. J., Ince, R. A. A., Gross, J., & Kayser, C. (2015). Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. *The Journal of Neuroscience*, 35(44), 14691–14701.
<https://doi.org/10.1523/JNEUROSCI.2243-15.2015>
- Kello, C. T., Bella, S. D., Médé, B., & Balasubramaniam, R. (2017). Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of The Royal Society Interface*, 14(135), 20170231.
- Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4), 255–289.
- Kitawaki, N., & Itoh, K. (1991). Pure delay effects on speech quality in telecommunications. *IEEE Journal on Selected Areas in Communications*, 9(4), 586–593.
- 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, 3071.

- Kotz, S. A., & Schwartz, M. (2010). Cortical speech processing unplugged: A timely subcortico-cortical framework. *Trends in Cognitive Sciences*, 14(9), 392–399.
- Kowal, S., & O’Connell, D. C. (1997). Theoretical ideals and their violation: Princess Diana and Martin Bashir in the BBC interview. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 7(3), 309–323.
<https://doi.org/10.1075/prag.7.3.02kow>
- Kubetschek, C., & Kayser, C. (2021). Delta/Theta band EEG activity shapes the rhythmic perceptual sampling of auditory scenes. *Scientific Reports*, 11(1), 1–15.
- Lelonkiewicz, J. R., & Gambi, C. (2020). Making oneself predictable in linguistic interactions. *Acta Psychologica*, 209.
<https://doi.org/10.1016/j.actpsy.2020.103125>
- Lepola, P., Needham, A., Mendum, J., Sallabank, P., Neubauer, D., & de Wildt, S. (2016). Informed consent for paediatric clinical trials in Europe. *Archives of Disease in Childhood*, 101(11), 1017–1025. <https://doi.org/10.1136/archdischild-2015-310001>
- 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>
- Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.
- Lukic, S., Thompson, C. K., Barbieri, E., Chiappetta, B., Bonakdarpour, B., Kiran, S., Rapp, B., Parrish, T. B., & Caplan, D. (2021). Common and distinct neural substrates of sentence production and comprehension. *NeuroImage*, 224, 117374. <https://doi.org/10.1016/j.neuroimage.2020.117374>
- Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6), 1001–1010.
- Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology*, 8(211).
- Makarova, V., & Zhou, X. (2006). *Prosodic characteristics in the speech of Chinese EFL learners*. 3rd International Conference on Speech Prosody, SP 2006, Department of Languages and Linguistics, University of Saskatchewan, Canada.
- McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, 1(1), 1–15.
<https://doi.org/doi:10.1017/S2041536210000012>

- Mertens, J., & de Ruiter, J. P. (2021). Cognitive and social delays in the initiation of conversational repair. *Dialogue and Discourse*, 12(1), 21–44.
<https://doi.org/10.5210/DAD.2021.102>
- Meyer, A. S., Alday, P. M., Decuyper, C., & Knudsen, B. (2018). Working together: Contributions of corpus analyses and experimental psycholinguistics to understanding conversation. *Frontiers in Psychology*, 9, 525.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (Version 1.7-14) [Computer software]. <https://CRAN.R-project.org/package=e1071>
- Meyer, L., Sun, Y., & Martin, A. E. (2019). Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, 1–11.
- Microsoft Excel. (2018). [Computer software]. Microsoft cooperation.
<https://office.microsoft.com/excel>
- Mooney, S., & Sullivan, G. C. (2015). Investigating an acoustic measure of perceived isochrony in conversation: Preliminary notes on the role of rhythm in turn transitions. *Univ. Penn. Work. Pap. Ling*, 21(2), 127–135.
- Mori, H. (2009). An analysis of switching pause duration as a paralinguistic feature in expressive dialogues. *Acoustical Science and Technology*, 30(5), 376–378.
<https://doi.org/10.1250/ast.30.376>
- Morrill, T., Baese-Berk, M., Heffner, C., & Dilley, L. (2015). Interactions between distal speech rate, linguistic knowledge, and speech environment. *Psychonomic Bulletin & Review*, 22(5), 1451–1457.
- Nagao, K., & De Jong, K. (2007). Perceptual rate normalization in naturally produced rate-varied speech. *The Journal of the Acoustical Society of America*, 121(5), 2882–2898.
- Nagaoka, C., Komori, M., Nakamura, T., & Draguna, M. R. (2005). Effects of Receptive Listening on the Congruence of Speakers' Response Latencies in Dialogues. *Psychological Reports*, 97(1), 265–274. <https://doi.org/10.2466/pr0.97.1.265-274>
- Nam, C. S., Choo, S., Huang, J., & Park, J. (2020). Brain-to-brain neural synchrony during social interactions: A systematic review on hyperscanning studies. *Applied Sciences (Switzerland)*, 10(19). <https://doi.org/10.3390/APP10196669>

- Nasreen, S., Hough, J., & Purver, M. (2021). *Detecting Alzheimer's Disease using Interactional and Acoustic features from Spontaneous Speech*. 1962–1966.
- Nevat, M., Khateb, A., & Prior, A. (2014). When first language is not first: An functional magnetic resonance imaging investigation of the neural basis of diglossia in Arabic. *European Journal of Neuroscience*, 40(9), 3387–3395.
<https://doi.org/10.1111/ejn.12673>
- Nevile, M. (2007). Talking without overlap in the airline cockpit: Precision timing at work. *Text and Talk*, 27(2), 225–249. <https://doi.org/10.1515/TEXT.2007.009>
- Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics*, 37(1), 46–65.
- Nguyen, V., Versyp, O., Cox, C. M. M., & Fusaroli, R. (2021). *A systematic review and Bayesian meta-analysis of the development of turn taking in adult-child vocal interactions*. PsyArXiv. <https://doi.org/10.31234/osf.io/3bak6>
- Oakey-Frost, D. N., Harris, J. A., Roberge, E. M., Andres, W. C., Rugo, K. F., Bryan, A. O., & Bryan, C. J. (2022). Verbal Response Latency as a Behavioral Indicator of Diminished Wish to Live in a Clinical Sample of Active Duty Army Personnel with Recent Suicidal Ideation. *Archives of Suicide Research*, 26(3), 1046–1059.
<https://doi.org/10.1080/13811118.2020.1848670>
- Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences*, 23(11), 913–926.
- O'Dell, M., Nieminen, T., & Lennes, M. (2012). Modeling turn-taking rhythms with oscillators. *Linguistica Uralica*, 48(3), 218–227.
<https://doi.org/10.3176/lu.2012.3.08>
- Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, 5(11), eaay6279.
<https://doi.org/10.1126/sciadv.aay6279>
- Ogden, R., & Hawkins, S. (2015). Entrainment as a basis for co-ordinated actions in speech. *International Congress of the Phonetic Sciences*.
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*.
<https://doi.org/10.1186/s13643-016-0384-4>
- Ovid Technologies, Ltd. (2023). *MEDLINE*. <https://www.nlm.nih.gov/bsd/medline.html>

- Ozaki, A., Hara, S., Kusakawa, T., Miyajima, C., Nishino, T., Kitaoka, N., Itou, K., Takeda, K., & et al.; European Media Laboratory GmbH (EML); Instituut voor Nederlandse Lexicologie (INL); Linguattec; Microsoft; Nuance. (2008). *In-car speech data collection along with various multimodal signals*. 1846–1851.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
<https://doi.org/10.1136/bmj.n71>
- Paschen, L., Fuchs, S., & Seifart, F. (2022). Final Lengthening and vowel length in 25 languages. *Journal of Phonetics*, 94, 101179.
<https://doi.org/10.1016/j.wocn.2022.101179>
- Paucke, M., Oppermann, F., Koch, I., & Jescheniak, J. D. (2015). On the costs of parallel processing in dual-task performance: The case of lexical processing in word production. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6), 1539–1552. <https://doi.org/10.1037/a0039583>
- Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, 3.
- Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 539–558.
- Peltonen, P. (2022). Connections between measured and assessed fluency in L2 peer interaction: A problem-solving perspective. *International Review of Applied Linguistics in Language Teaching*, 60(4), 983–1011. <https://doi.org/10.1515/iral-2020-0030>
- Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shaped. In *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge University Press.
- Pouw, W., & Holler, J. (2022). Timing in conversation is dynamically adjusted turn by turn in dyadic telephone conversations. *Cognition*, 222, 105015.
<https://doi.org/10.1016/j.cognition.2022.105015>

- Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90(6), 2956–2970.
- Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, 123(2), 1104–1113. <https://doi.org/10.1121/1.2821762>
- R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria. <http://www.R-project.org/>
- Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking Rate Affects the Perception of Duration as a Suprasegmental Lexical-stress Cue. *Language and Speech*, 54(2), 147–165. <https://doi.org/10.1177/0023830910397489>
- Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural Entrainment to Speech Modulates Speech Intelligibility. *Current Biology*, 28(2), 161–169. <https://doi.org/10.1016/j.cub.2017.11.033>
- Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00089>
- Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *Journal of the Acoustical Society of America*, 133(6), 471–477. <https://doi.org/10.1121/1.4802900>
- Roberts, F., Margutti, P., & Takano, S. (2011). Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5), 331–354. <https://doi.org/10.1080/0163853X.2011.558002>
- 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>
- Rohatgi, A. (2017). *WebPlotDigitizer*. Austin, Texas, USA. https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/6556939
- Rühlemann, C., & Gries, S. Th. (2020). Speakers advance-project turn completion by slowing down: A multifactorial corpus analysis. *Journal of Phonetics*, 80, 100976. <https://doi.org/10.1016/j.wocn.2020.100976>

- Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7–55). Elsevier.
- Sakuma, J., Fujie, S., & Kobayashi, T. (2023). *Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction*. 369–374.
<https://doi.org/10.1109/SLT54892.2023.10023458>
- Sandamirskaya, Y., & Schöner, G. (2008). *Dynamic field theory and embodied communication* (Vol. 4930). https://doi.org/10.1007/978-3-540-79037-2_14
- Schegloff, E. A. (1987). Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organization. *Talk and Social Organisation*, 1, 70–85.
- Schoenenberg, K., Raake, A., & Koeppel, J. (2014). Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end. *International Journal of Human-Computer Studies*, 72(5), 477–487.
<https://doi.org/10.1016/j.ijhcs.2014.02.004>
- Schultz, B. G., O'Brien, I., Phillips, N., McFarland, D. H., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics*, 37(5), 1201.
- Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action—Candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, 10(4), 295–302.
- Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1), 289.
- Scrucca, L., Fraley, C., Murphy, B. T., & Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC. <https://mclust-org.github.io/book/>
- Shimojima, A., Katagiri, Y., Koiso, H., & Swerts, M. (2002). Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses. *Speech Communication*, 36(1–2), 113–132.
- Sneck, S. (1987). *Assessment of Chronography in Finnish-English Telephone Conversation: An Attempt at a Computer Analysis*. Jyväskylä Cross-Language Studies, No. 14. ERIC. <https://eric.ed.gov/?id=ED292303>

- Song, J. (2018). *The effects of adverse conditions on speech recognition by non-native listeners: Electrophysiological and behavioural evidence* [University College London]. <https://discovery.ucl.ac.uk/id/eprint/10041104/>
- Song, J., & Iverson, P. (2016). Cross-linguistic perception of continuous speech: Neural entrainment to the speech amplitude envelope. *The Journal of the Acoustical Society of America*, 140(4), 3335–3335. <https://doi.org/10.1121/1.4970635>
- Sørensen, A. J. M., Fereczkowski, M., & MacDonald, E. N. (2021). Effects of Noise and Second Language on Conversational Dynamics in Task Dialogue. *Trends in Hearing*, 25, 233121652110244. <https://doi.org/10.1177/23312165211024482>
- Stiefel, K. M., & Ermentrout, G. B. (2016). Neurons as oscillators. *Journal of Neurophysiology*, 116(6), 2950–2960. <https://doi.org/10.1152/jn.00525.2015>
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592. <https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>
- Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1074–1095. <https://doi.org/10.1037/0096-1523.7.5.1074>
- Tanner, J., Sonderegger, M., Kendall, T., Stuart-Smith, J., Mielke, J., Thomas, E., & Dodsborn, R. (2024, March 26). *Dialect and speaker variation, and the role of anticipatory shortening in English articulation rate*. Colloquium of the British Association of Academic Phoneticians, Cardiff, Wales.
- Templeton, E. M., Chang, L. J., Reynolds, E. A., LeBeaumont, M. D. C., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 119(4). <https://doi.org/10.1073/pnas.2116915119>
- Ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1–2), 80–86.
- Teng, X., Tian, X., Doelling, K., & Poeppel, D. (2018). Theta band oscillations reflect more than entrainment: Behavioral and neural evidence demonstrates an active chunking process. *European Journal of Neuroscience*, 48(8), 2770–2782. <https://doi.org/10.1111/ejn.13742>

- Timbrook, J., Olson, K., & Smyth, J. D. (2022). Your Best Estimate is Fine. Or is It? *Journal of Official Statistics*, 38(4), 1097–1123. <https://doi.org/10.2478/jos-2022-0047>
- Tobin, K. (1987). The Role of Wait Time in Higher Cognitive Level Learning. *Review of Educational Research*, 57(1), 69–95.
<https://doi.org/10.3102/00346543057001069>
- Tomasello, M. (2010). *Origins of human communication*. MIT press.
- Torreira, F., & Bögels, S. (2022). Vocal reaction times to speech offsets: Implications for processing models of conversational turn-taking. *Journal of Phonetics*, 94, 101175.
- Trimboli, C., & Walker, M. B. (1984). Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology*, 20(4), 297–311.
- Van Bree, S., Sohoglu, E., Davis, M. H., & Zoefel, B. (2021). Sustained neural rhythms reveal endogenous oscillations supporting speech perception. *PLOS Biology*, 19(2). <https://doi.org/10.1371/journal.pbio.3001142>
- Van Den Akker, O., Peters, G.-J., Bakker, C., Carlsson, R., Coles, N. A., Corker, K. S., Feldman, G., Mellor, D. T., Moreau, D., & Nordström, T. (2020, September). *Inclusive Systematic Review Registration Form*. MetaArXiv.
<https://doi.org/10.31222/osf.io/3nbea>
- Van Os, M., De Jong, Nivja H, & Bokser, H. R. (2020). Fluency in Dialogue: Turn-Taking Behavior Shapes Perceived Fluency in Native and Nonnative Speech. *Language Learning: A Journal of Research in Language Studies*, 70(4), 1–34.
- Van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV Corpus: A Free Dialog Video Corpus. *LREC*, 501–508.
- Vesper, C. V., Sebanz, N., van der Wel, R. P. R. D., & Knoblich, G. (2011). Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Experimental Brain Research*, 211(3–4), 517–530.
- Vranjes, J., & Oben, B. (2022). Anticipation and timing of turn-taking in dialogue interpreting: A quantitative study using mobile eye-tracking data. *Target. International Journal of Translation Studies*, 34(4), 627–651.
<https://doi.org/10.1075/target.20121.vra>
- Warner, R. M. (1979). Periodic Rhythms in Conversational Speech. *Language and Speech*, 22(4), 381–396. <https://doi.org/10.1177/002383097902200407>

- Warren, P. (1985). *The temporal organisation and perception of speech*. [PhD Thesis]. University of Cambridge.
- Wehrle, S. (2021). *A Multi-Dimensional Analysis of Conversation and Intonation in Autism Spectrum Disorder* [PhD Thesis]. University of Cologne.
- Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. *Proceedings of the International Conference of Phonetic Sciences*.
- White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, 63, 38–54.
- White, L., Mattys, S., Knight, S., Saunders, T., & Macbeath, L. (2022). Temporal expectations and the interpretation of timing cues to word boundaries. *Proc. Speech Prosody 2022*, 322–326.
- White, L., Mattys, S., & Wiget, L. (2012). Segmentation cues in conversational speech: Robust semantics and fragile phonotactics. *Frontiers in Psychology*, 3, 375.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- 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., Wiemann, J. M., & Zimmerman, D. H. (1984). Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3(3), 159–183. <https://doi.org/10.1177/0261927X8400300301>
- Wilson, T. P., & Zimmerman, D. H. (1986). The Structure of Silence between Turns in Two-party Conversation. *Discourse Processes*, 9(4), 375–390. <https://doi.org/10.1080/01638538609544649>
- Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.
- Xu, Q., & Ye, D. (2015). Temporal integration reflected by frequency following response in auditory brainstem. *Bio-Medical Materials and Engineering*, 26, S767–S778.

- Xue, J., Xu, H., Wang, S., Hao, T., Yang, Y., Zhang, X., Song, Y., Li, Y., & Zhao, J. (2021). Design and synthesis of 2D rGO/NiO heterostructure composites for high-performance electrochromic energy storage. *Applied Surface Science*, 565, 150512.
- Yngve, V. H. (1970). On getting a word in edgewise. *Papers from the Sixth Regional Meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, 567–578. <https://cir.nii.ac.jp/crid/1571417124968122496>
- Yoshida, M., Miyake, Y., & Furuyama, N. (2008). *Temporal development of pragmatics and dynamics in conversation for building consensus*. 2419–2425. <https://doi.org/10.1109/ICSMC.2008.4811657>

Appendix A1. PRISMA Checklists

Section/topic	#	Checklist item	Location(s) reported
TITLE			
Title	1	Identify the report as a systematic review.	Chapter title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	NA
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	3.1.1
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	3.1.4
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	3.2.2 3.2.7 3.2.5
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	3.2.1
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	3.2.1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	3.2.3
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	3.2.6
Data items	10a	List and define all outcomes for which	3.2.6

Section/topic	#	Checklist item	Location(s) reported
		data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	3.2.6
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	3.2.3 3.2.4
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	3.2.7
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	3.2.7
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	3.2.6
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	3.2.7
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	3.2.7

Section/topic	#	Checklist item	Location(s) reported
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	3.2.7 3.3.6
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	3.3.6
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	No risk of reporting bias was performed
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	No certainty assessment was performed
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	3.2.3 3.2.5 3.3.1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	3.2.3
Study characteristics	17	Cite each included study and present its characteristics.	A3
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	3.2.4
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	3.3.6 – 3.3.12
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	3.3.6 3.3.8 3.3.9
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g.	3.3.6 – 3.3.12

Section/topic	#	Checklist item	Location(s) reported
		confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	3.3.6 – 3.3.12
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	No sensitivity analysis performed.
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	No risk of bias assessment performed.
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	No certainty assessment performed.
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	3.3 – 3.4
	23b	Discuss any limitations of the evidence included in the review.	3.3.4
	23c	Discuss any limitations of the review processes used.	3.4
	23d	Discuss implications of the results for practice, policy, and future research.	3.4
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	3.2
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	3.2
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	3.2.5 3.2.6 Table 3.4
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Thesis acknowledgements (piii)

Section/topic	#	Checklist item	Location(s) reported
Competing interests	26	Declare any competing interests of review authors.	No competing interests
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	3.2.6

Table A 1 PRISMA checklist (Page et al., 2021).

Section/topic	#	Checklist item	Location(s) reported
INFORMATION SOURCES AND METHODS			
Database name	1	Name each individual database searched, stating the platform for each.	3.2.1
Multi-database searching	2	If databases were searched simultaneously on a single platform, state the name of the platform, listing all of the databases searched.	3.2.1
Study registries	3	List any study registries searched.	No registries searched
Online resources and browsing	4	Describe any online or print source purposefully searched or browsed (e.g., tables of contents, print conference proceedings, web sites), and how this was done.	3.2.1
Citation searching	5	Indicate whether cited references or citing references were examined, and describe any methods used for locating cited/citing references (e.g., browsing reference lists, using a citation index, setting up email alerts for references citing included studies).	3.2.1
Contacts	6	Indicate whether additional studies or data were sought by contacting authors, experts, manufacturers, or others.	3.2.1
Other methods	7	Describe any additional information sources or search methods used.	3.2.1
SEARCH STRATEGIES			
Full search strategies	8	Include the search strategies for each database and information source, copied and pasted exactly as run.	Table 3.2
Limits and restrictions	9	Specify that no limits were used, or describe any limits or restrictions applied to a search (e.g., date or time period, language, study design) and provide justification for their use.	3.2.1
Search filters	10	Indicate whether published search filters were used (as originally designed or modified), and if so, cite the filter(s) used.	3.2.1
Prior work	11	Indicate when search strategies from other literature reviews were adapted or reused for a substantive part or all of the search, citing the previous review(s).	NA
Updates	12	Report the methods used to update the search(es) (e.g., rerunning searches, email alerts).	NA

Dates of searches	13	For each search strategy, provide the date when the last search occurred.	3.2 Table 3.2
PEER REVIEW			
Peer review	14	Describe any search peer review process.	3.2.3
MANAGING RECORDS			
Total Records	15	Document the total number of records identified from each database and other information sources.	3.2.3
Deduplication	16	Describe the processes and any software used to deduplicate records from multiple database searches and other information sources.	3.2.3

Table A 2 PRISMA-S checklist (Rethlefsen et al., 2021).

Appendix A2. Examples of Records Excluded Based on Abstract

The following two examples are abstracts from records that were excluded based on information in the abstract alone:

Basset et al. (2001) measured the timing of nasal airflow to draw inferences regarding the behaviour of the velum during and around the production of nasals. They do not use this measurement to quantify conversational timing.

“The goal of this paper is to compare the degree of anticipatory and carryover phenomena in the behaviour of the velum during the production of French spontaneous speech, and to compare the data with the same speech material, but read. Airflows through the nose and the mouth were taken as an indirect indication of the velum behaviour. French is a language of special interest because there is a phonological contrast between oral and nasal vowels. The results show a tendency for nasal airflow to start before the nasal and a strong propensity to spread after the nasal. No differences could be established between speaking styles (spontaneous vs. read speech) regarding nasal airflow anticipation and carryover.” (Basset et al., 2001)

Nevat et al. (2014) collected fMRI responses of participants while they were reading written words in literary Arabic, Hebrew and a non-literary variety of Arabic. Their study is not concerned with spoken conversation.

“In Arabic, the language used for everyday conversation ('spoken Arabic' - SA) differs markedly from literary Arabic (LA), which is used for written communication and formal functions. This fact raises questions regarding the cognitive status of the two varieties and their processing in the brain. Previous studies using auditory stimuli suggested that LA is processed by Arabic native speakers as a second language. The current study examined this issue in the visual modality. Functional magnetic resonance imaging (fMRI) responses were collected while Arabic-Hebrew bilinguals performed a semantic categorization task on visually presented words in LA, SA and Hebrew. Performance on LA was better than SA and Hebrew, which did not differ from each other. Activation in SA was stronger than in LA in left inferior frontal, precentral, parietal and occipito-temporal regions, and stronger than in Hebrew in left precentral and parietal regions. Activation in SA was also

less lateralized than activation for LA and Hebrew, which did not differ from each other in terms of lateralization, though activation for Hebrew was more extensive in both hemispheres than activation for LA. Altogether, these results indicate an advantage for LA in the current study, presumably due to participants' proficiency in reading in this language. Stronger activation for SA appears to be due to the relative unfamiliarity of written word forms in SA, which could also explain differences in performance between the two languages. However, the stronger activation observed in the left parietal cortex may also reflect stronger associations among words in SA.” (Nevat et al., 2014)

Appendix A3. Table of Included Reports

Report ID	Aimed to:	Summary of finding regarding turn timing
(Aubanel et al., 2011)	Describe differences in speech between conversations held while another conversation is happening in the background, or in silence.	When another conversation is present, turn timing is slower.
(Beattie, 1978)	Test whether speaker gaze has a floor-appointing function by describing the relationship between the presence of gaze and the duration of turn timing.	Turns that ended without the speaker gazing towards their interlocutor were responded to with faster turn timing, except in the case of incomplete turns (those without turn-ending cues).
(Bögels et al., 2020)	Test whether speakers can, in principle, start planning their response early in spontaneous conversation, if turn timing is shorter when speakers can start planning earlier, and whether the neural signature of response planning found in experimental studies extend to spontaneous conversation.	Speakers could start planning their turn after around 1/3rd of the turn had been uttered. A longer planning time was associated with shorter turn timing, and longer responses were associated with longer turn timing. A positive ERP was observed around 500ms after the speaker could start planning, but the expected posterior decrease in power in the alpha frequency band was not observed.
(Brosy et al., 2016)	Test the effect of turn timing on answer type (narrative or non-narrative) and success in job interviews.	Mean total pause time before the answer significantly predicts hiring recommendations. Total pause time (time until the relevant response, including hesitations and within-speaker pauses) predicts response type. Mean turn timing is reported.
(Buanzur et al., 2018)	Replicate some of the findings of Stivers et al. (2009) in Ruuli.	The effect of polar/non-polar and type-confirming/type non-confirming is the same in Ruuli as in the other languages in Stivers et al. (2009). Polar questions are responded to faster than WH. Type-confirming polar questions are responded to faster than type-non-confirming.

Table A 3 A list of all included reports (part 1 of 6).

Report ID	Aimed to:	Summary of finding regarding turn timing
(Bull & Aylett, 1998)	Describe the effect of several task-related variables on turn timing, as well as the presence of eye-contact.	Task complexity, task familiarity, conversational game boundary, role (of the speaker in the game), eye contact and individual differences all significantly affect turn timing. Individual differences did not interact with the other effects.
(Chowdhury et al., 2014)	Attempt to build an algorithm that can do unsupervised recognition of competitive and non-competitive overlap.	Using a clustering algorithm, cases of overlap are classified in two groups. The clusters appear to correspond to competitive and non-competitive overlaps.
(De Vos et al., 2015)	Test to what extent the turn taking system is dependent on modality by comparing three different measures of turn timing in conversations in NGT (sign language of the Netherlands) to reports of turn timing.	In signed conversations, one of three operationalisation of turn timing results in measurements similar to spoken conversations of Stivers et al. (2009). The triadic conversation was not significantly different from the dyadic conversations.
(Fadden, 2008)	Compare several prosodic aspects of suspect's responses to interviewer questions during a police interview, including comparing turn timing between negative, affirmative and relevant responses by first time offenders and repeat offenders.	There was a significant interaction between offender group and response type. In post hoc testing, affirmative and relevant responses were significantly different, and negative and relevant responses. First time offenders have slower turn timing for a relevant response. There was a non-significant positive correlation between response length of a confession and turn timing.
(Heldner & Edlund, 2010)	Characterise the distribution of turn timing and discuss how many turn transitions could in theory be simple reactions to the end of the previous turn.	41% to 45% of turn switches were longer than 200ms.

Table A 4 A list of all included reports (part 2 of 6).

Report ID	Aimed to:	Summary of finding regarding turn timing
(Heldner et al., 2011)	Compare the distribution of turn timing before and after short turns to before and after longer turns, to determine what the effect is of excluding short utterances from analyses.	The turn timing distribution of longer turns had a lower central tendency (mean and highest density) and a wider variance than those related to shorter turns.
(Howes & Lavelle, 2023)	Compare turn timing statistics and qualitative differences in conversations between neurotypical speakers and conversations in which a person with a schizophrenia diagnosis was present (and the other interlocutors were unaware of their diagnosis).	There was a significant difference in overall turn timing between the groups. Post-hoc analysis suggested this effect arises because neurotypical-to-neurotypical turns switches are slower in conversations where a speaker with schizophrenia was present than in the conversations where only neurotypical speakers were present.
(Kowal & O'Connell, 1997)	Report an analysis of several aspects of a BBC interview with princess Dianna, to test 1) Clark's ideal delivery theory, 2) Sacks and Schlegoff's simplest systematics and 3) Grice's maxim of quantity, as measurements of performance quality, working from the assumption that this interview is an example of excellent performance.	Turn timing is reported.
(Levinson & Torreira, 2015)	Discuss how many overlaps are consistent with Sacks et al.'s (1974) no-gap-no-overlap model in a review paper on the implications of turn timing for processing models of language.	75% of overlaps are shorter than 374, consistent with the no-gap-no-overlap model.

Table A 5 A list of all included reports (part 3 of 6).

Report ID	Aimed to:	Summary of finding regarding turn timing
(Mertens & de Ruiter, 2021)	Evaluate how cognitive and social variables affect turn timing.	The best predictive model included both cognitive and social predictors of the timing of other-initiated repair requests. The authors particularly highlight that discourse context (operationalised as the previous turn timing), overlapping speech during the previous turn, previous turn duration, and repair request type affect turn timing.
(Mori, 2009)	Describe the effects of the perceived emotional state of the previous utterance on turn timing.	The perceived emotional state of an utterance, as determined in a rating study of isolated utterances (i.e., not including the turn timing) is significantly correlated to turn timing for some dimensions of emotionality: pleasantness, dominance, credibility, interest, positivity. When backchannels are excluded, turn timing is significantly correlated to pleasantness, aroused/sleepiness, credibility, positivity. The turn timing of backchannels is faster than of non-backchannels.
(Nagaoka et al., 2005)	Test the difference in turn timing between dialogues in which participants were expected to debate vs. in which they were expected to try to find a compromise.	Turn timing was faster in the compromise condition. Turn timing between speakers only correlated in the second half of the compromise condition, not in the first half or in the debate condition.
(Ozaki et al., 2008)	Report the construction of a multi-modal driving corpus.	When participants were speaking to a human navigator while being the driver of a car, the driver's mean turn timing was negative. While the participant was driving their turn timing was non-significantly faster than when the car was standing still.

Table A 6 A list of all included reports (part 4 of 6).

Report ID	Aimed to:	Summary of finding regarding turn timing
(Pouw & Holler, 2022)	Describe how speakers over-adjust turn timing based on the previous turn's turn timing by testing whether there is an autocorrelation with lag 1 between turn timing of both speakers in a conversation.	There is an autocorrelation with lag 1 between turn timing of both speakers in a conversation. There is no noteworthy autocorrelation at further lags and the same pattern does not hold for single speaker turn timing time series, or shuffled or simulated data. There was no correlation between factors related to turn timing (speech rate, final vowel and final consonant, or turn duration) and the antipersistence value (the degree to which the current value is in a different direction from the previous value) of the turn timing. There was only an extremely weak correlation with turn duration. Longer turns are followed by a more anti-persistent turn timing.
(S. G. Roberts et al., 2015)	Compare the importance of processing-related variables and sequence-organisation related variables in turn timing using Random Forest and describing the relationship between turn timing and some individual variables.	Both processing-related variables, and sequence-organisation variables are important. An additional indexical variable (speaker gender) was not important.
(Shimojima et al., 2002)	Describe the relationship between prosodic features of echoic responses in Japanese and their information- and dialogue function.	Information that is strongly integrated in the body of knowledge of the responder (according to annotator judgements) is associated with faster echoic responses.
(Sørensen et al., 2021)	Describe the effect of speaking in noise and in an L2 on conversational dynamics, particularly turn timing and inter-pausal unit (IPU) length.	In noise, the median and inter-quartile range (IQR) of the turn timing distribution are slightly higher. In L2, the median is slightly lower but the IQR higher. For longer first IPU, turn timing became faster and more narrowly distributed. For shorter second IPU, turn timing became faster and more narrowly distributed.

Table A 7 A list of all included reports (part 5 of 6).

Report ID	Aimed to:	Summary of finding regarding turn timing
(Stivers et al., 2009)	Test whether patterns in turn taking are shared across ten languages.	Turn timing averages in polar questions are slightly different between the ten languages, and so are annotators' judgements of when a response sounded delayed. However, the same factors affected turn timing in all languages in the same way.
(Ten Bosch et al., 2005)	Show that shallow annotations of large corpora can provide data on turn-taking, and how machine learning can provide semi-automatic enrichments.	Inter-turn pause durations by interlocutors are correlated.
(Trimboli & Walker, 1984)	Test if an observation by a previous study (Beattie, 1978), that turn timing after turns with no turn-ending cue was faster than turn timing after turns with a turn-ending cue, is due to the competitive nature of the conversations, by testing the effects of competitiveness, gender and completeness (turn ending cue presence) on turn timing.	Turn timing after a turn without a turn-ending cue in a cooperative conversation is slower than in a competitive conversation. In a cooperative conversation, turn timing after a turn without a cue is slower than after one with a cue. For a competitive conversation, turn timing after a turn without a cue was faster than after those with a cue.

Table A 8 A list of all included reports (part 6 of 6).

Appendix A4. Additional Grouping Plots

Chapter 3 includes forest plots that show the means, medians and modes of the included samples, which were colour-coded by conversation- and methodological variables. For completeness, the following plots depict the grouping of data by variables not visualised in Chapter 3. Figure A 1 to Figure A 4 visualise the entire selection of samples, Figure A 5 and Figure A 6, visualise subset 1 and Figure A 7 to Figure A 9 visualise subset 2.

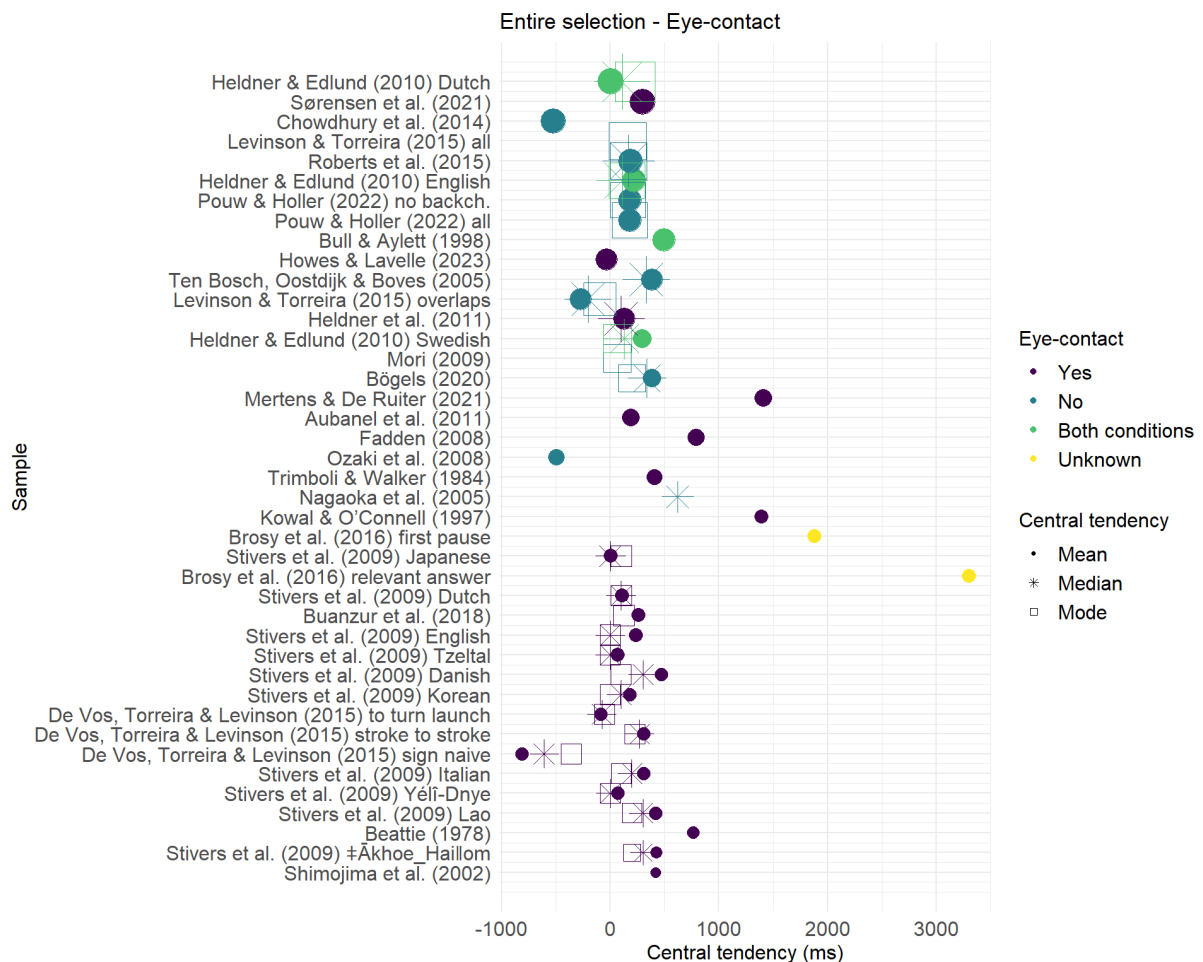


Figure A 1 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Eye-contact. Point size reflects the natural logarithm of the sample size.

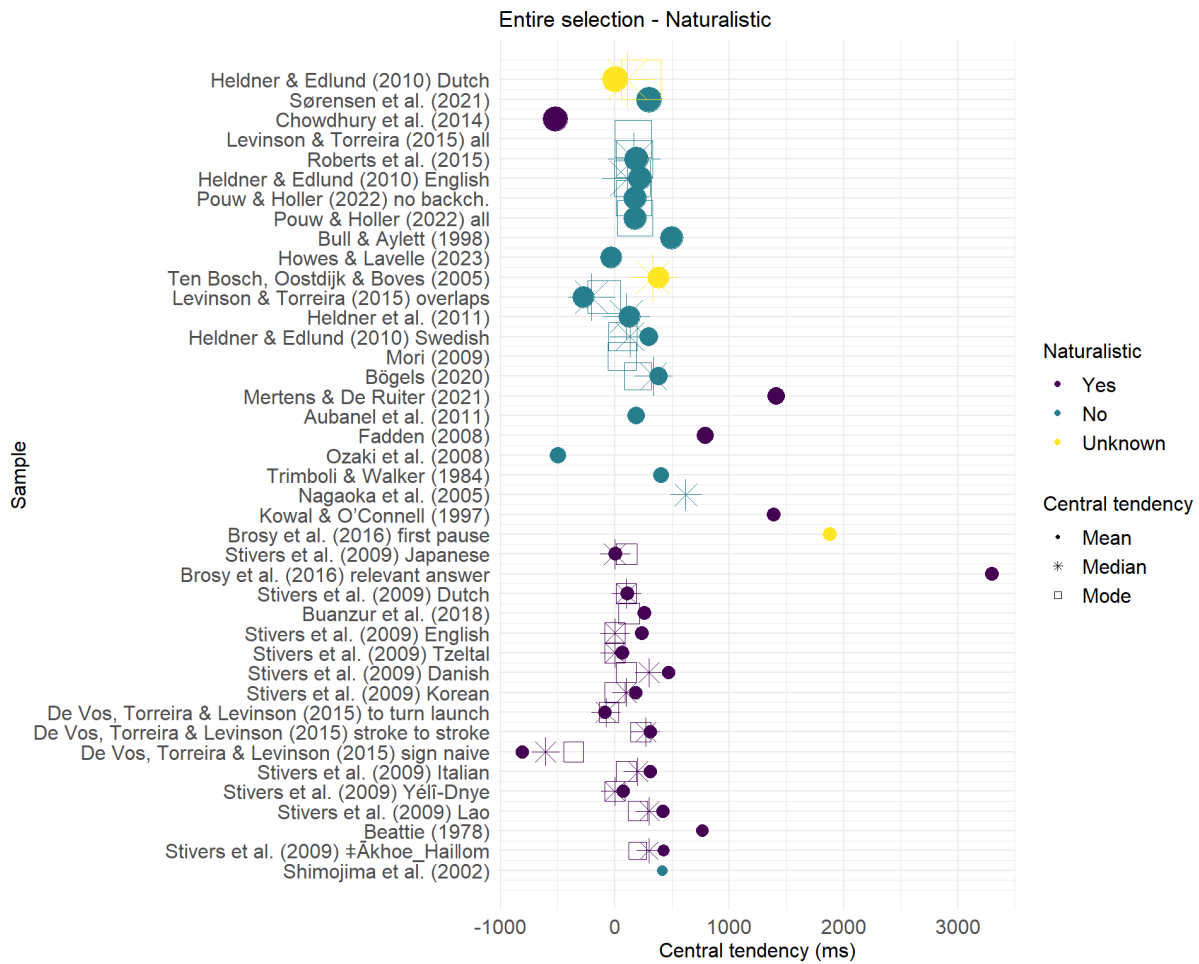


Figure A 2 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Naturalistic. Point size reflects the natural logarithm of the sample size.

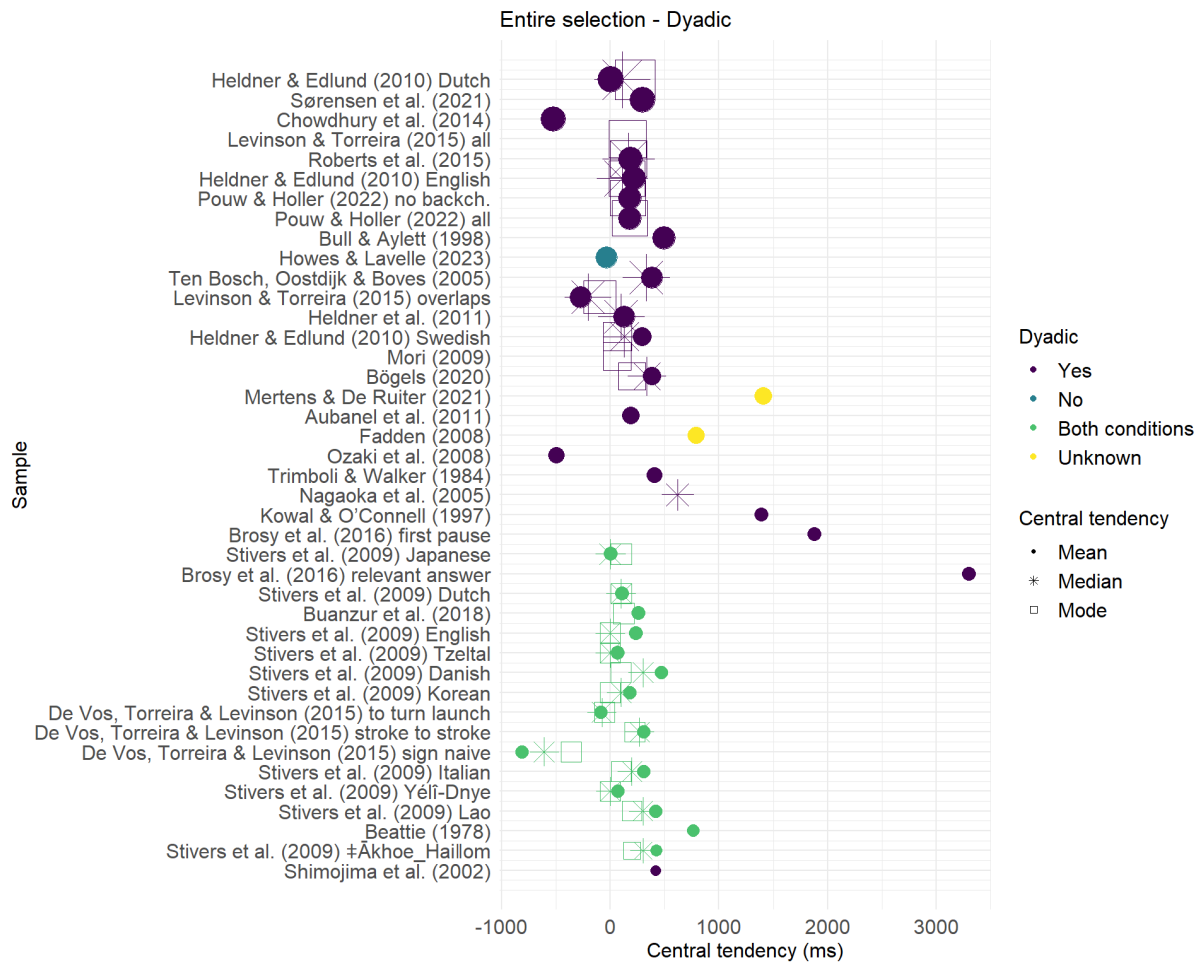


Figure A 3 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Dyadic. Point size reflects the natural logarithm of the sample size.

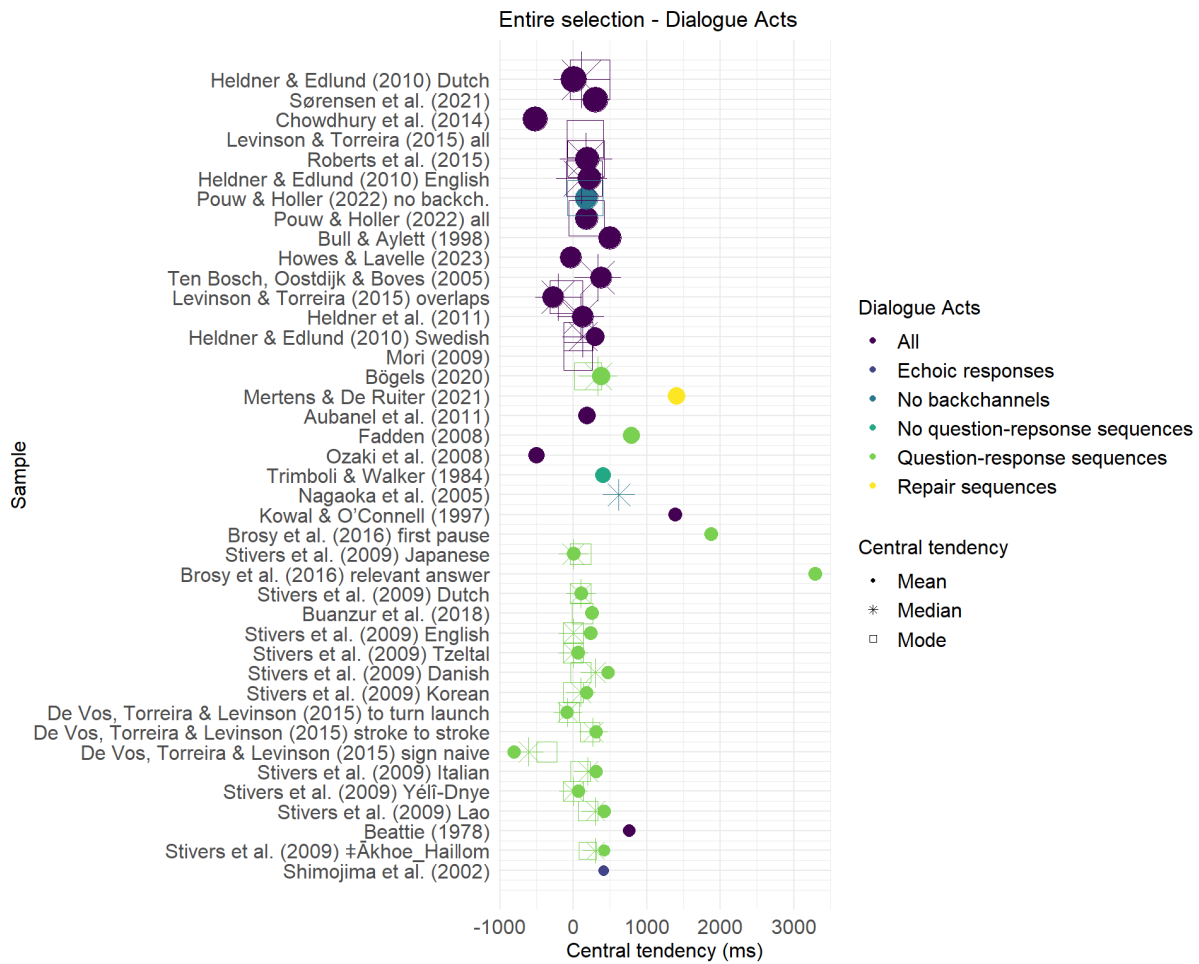


Figure A 4 The mean, median and mode of samples in the entire selection, arranged by sample size. The sample estimates are colour-coded by the variable Dialogue Acts. Point size reflects the natural logarithm of the sample size.

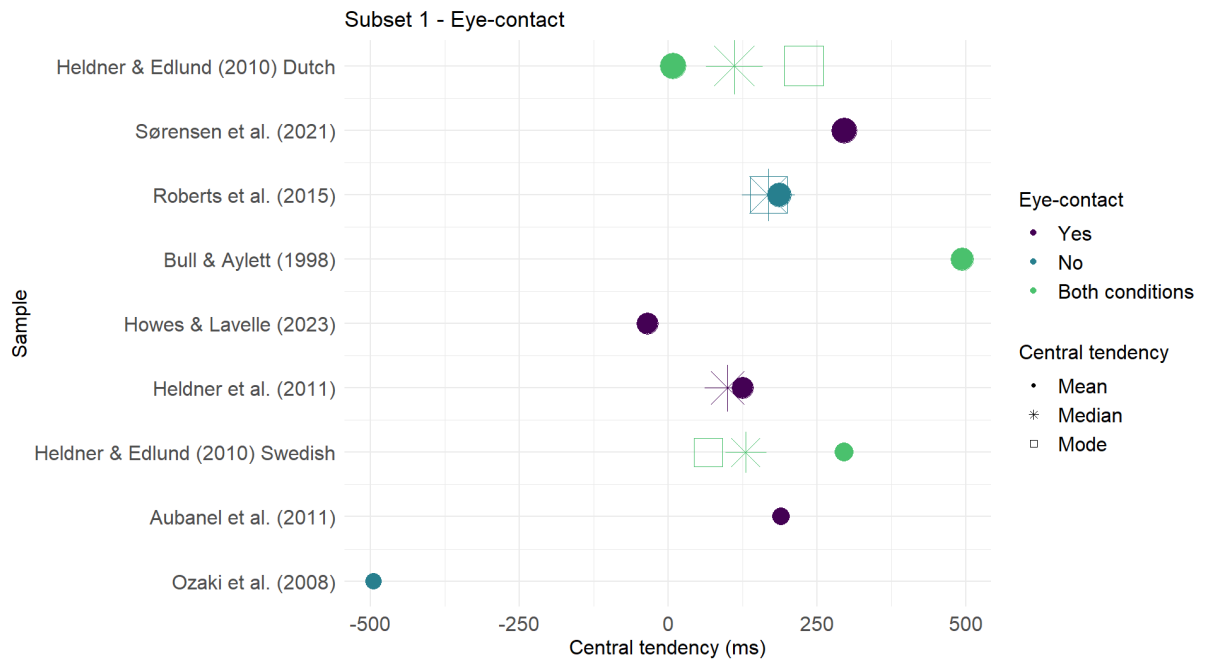


Figure A 5 The mean, median and mode of samples in subset 1, arranged by sample size. The sample estimates are colour-coded by the variable Eye-contact. Point size reflects the natural logarithm of the sample size.

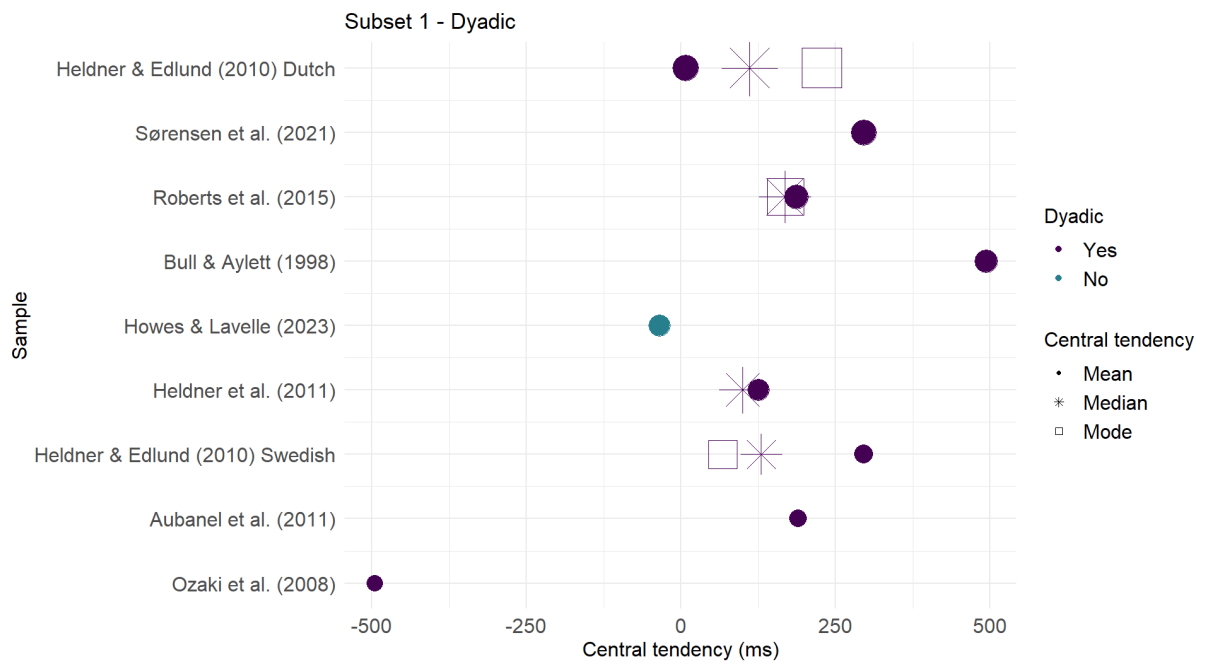


Figure A 6 The mean, median and mode of samples in subset 1, arranged by sample size. The sample estimates are colour-coded by the variable Dyadic. Point size reflects the natural logarithm of the sample size.

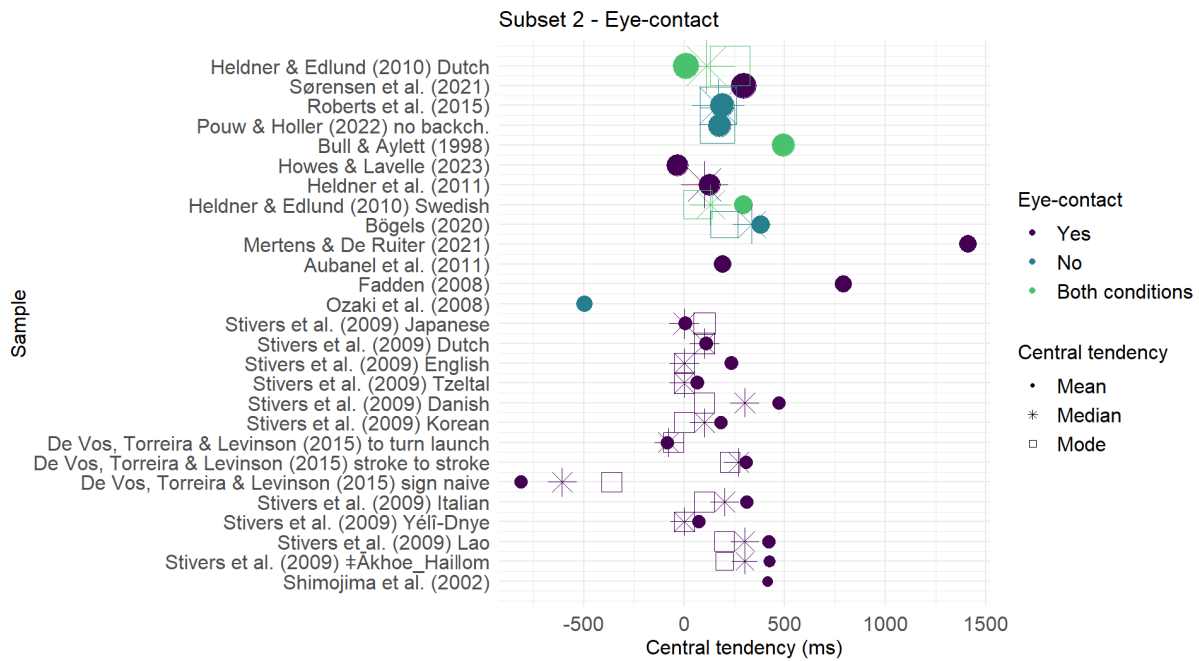


Figure A 7 The mean, median and mode of samples in subset 2, arranged by sample size. The sample estimates are colour-coded by the variable Eye-contact. Point size reflects the natural logarithm of the sample size.

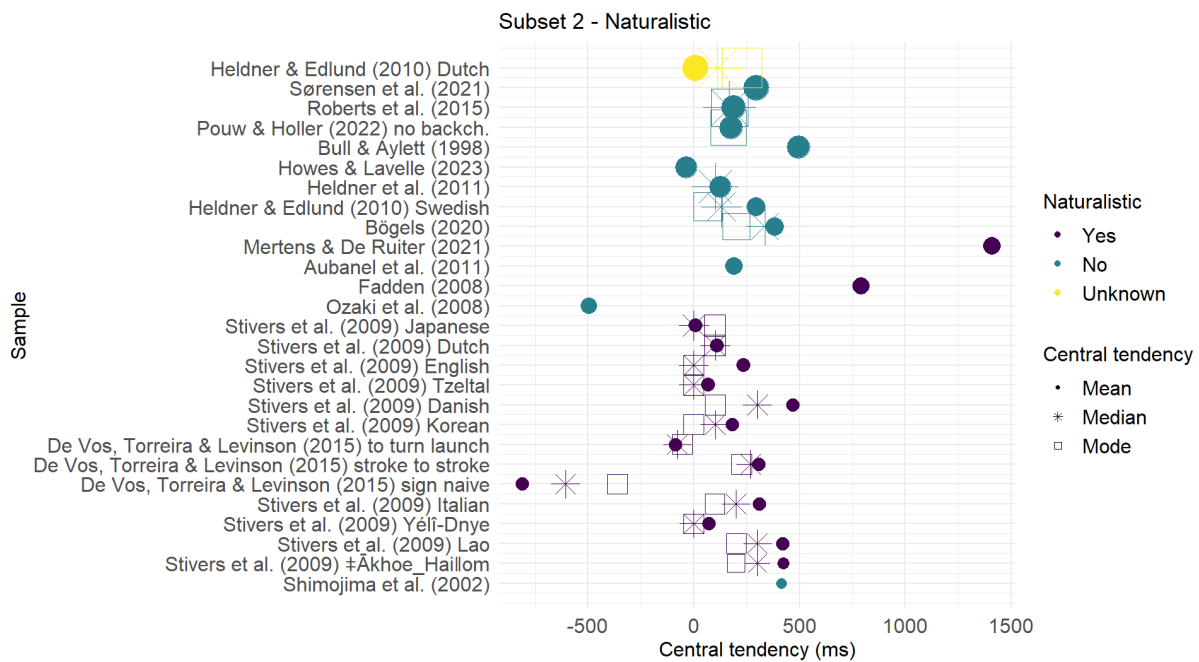


Figure A 8 The mean, median and mode of samples in subset 2, arranged by sample size. The sample estimates are colour-coded by the variable Naturalistic. Point size reflects the natural logarithm of the sample size.

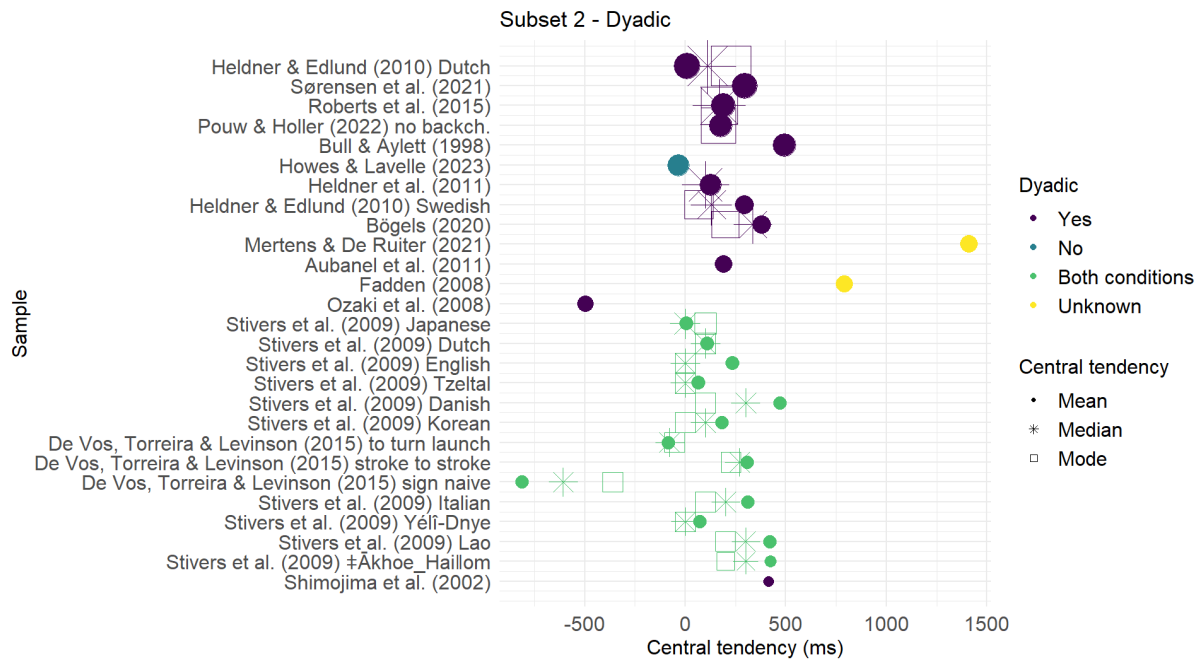


Figure A 9 The mean, median and mode of samples in subset 2, arranged by sample size. The sample estimates are colour-coded by the variable Dyadic. Point size reflects the natural logarithm of the sample size.

Appendix A5. Visualisations of Distributions

Symmetrical distributions

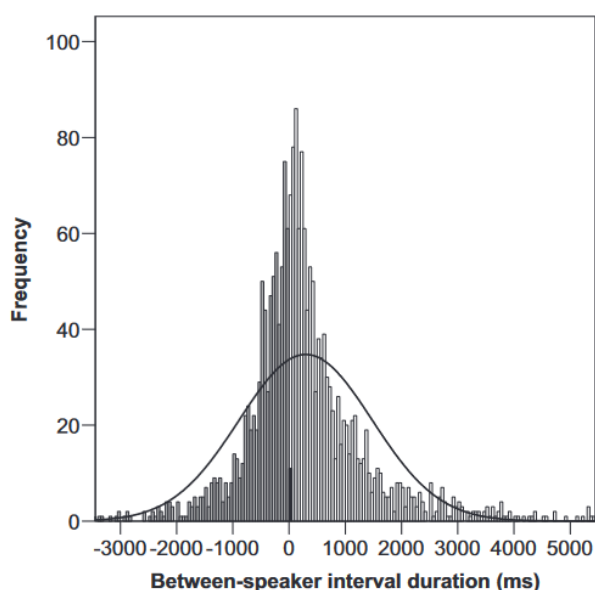


Figure A 10 Histogram with normal distribution overlayed (Heldner & Edlund, 2010) Scottish English

Reprinted from: Journal of Phonetics, 38/4, M. Heldner and J. Edlund, Pauses, gaps and overlaps in conversations, p555–568, Copyright (2010) Elsevier, with permission from Elsevier.

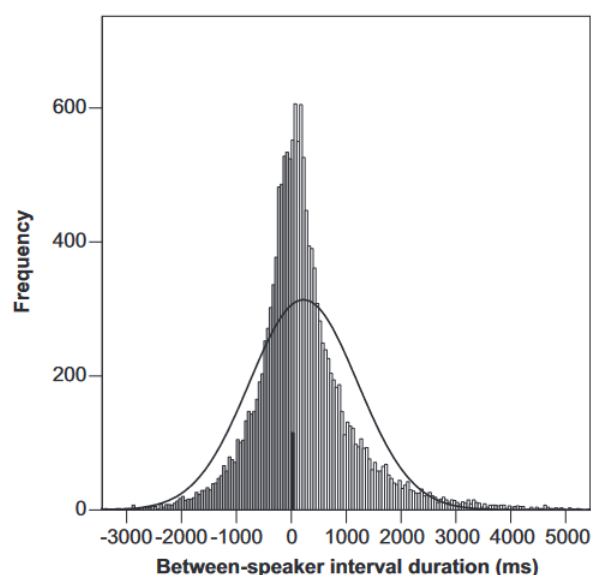


Figure A 11 Histogram with normal distribution overlayed (Heldner & Edlund, 2010) Swedish

Reprinted from Journal of Phonetics, 38/4, M. Heldner and J. Edlund, Pauses, gaps and overlaps in conversations, p555–568, Copyright (2010) Elsevier, with permission from Elsevier.

Symmetrical distributions

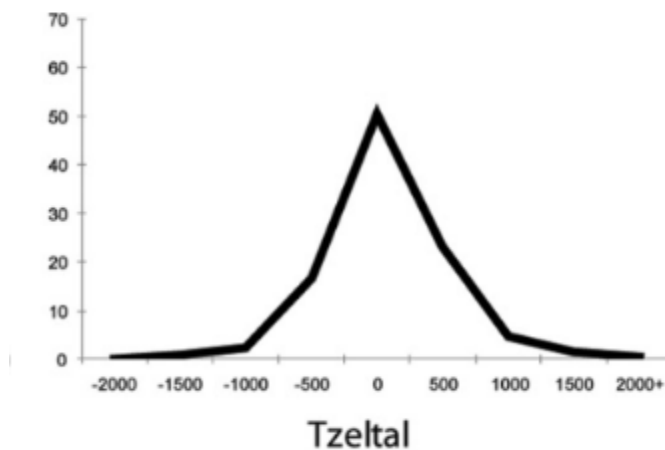


Figure A 12 Histogram (Stivers et al., 2009) Tzeltal

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

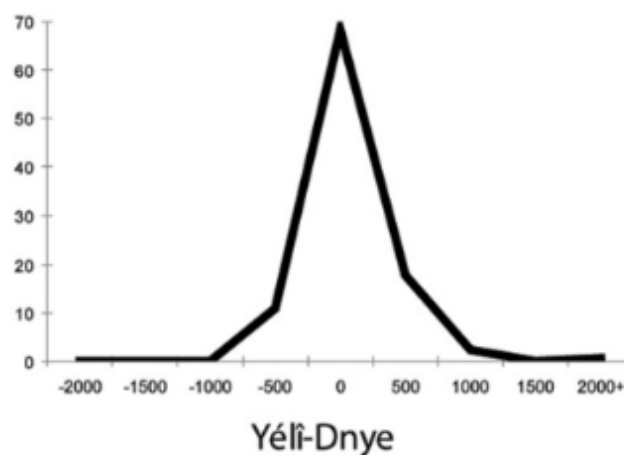


Figure A 13 Histogram (Stivers et al., 2009) Yéllî Dnye

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106> Copyright belongs to the authors. Reprinted with permission.

Left-skewed distributions

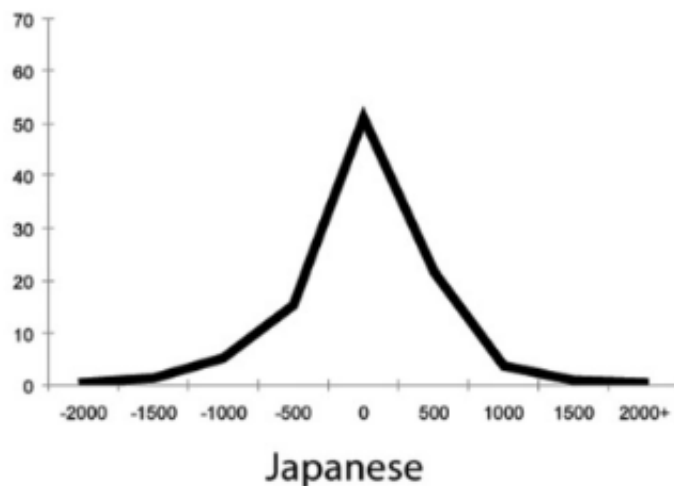


Figure A 14 Histogram (Stivers et al., 2009) Japanese

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

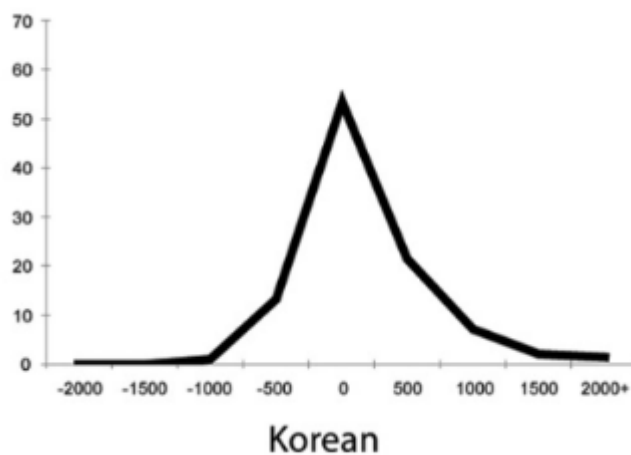


Figure A 15 Histogram (Stivers et al., 2009) Korean

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

Left-skewed distributions

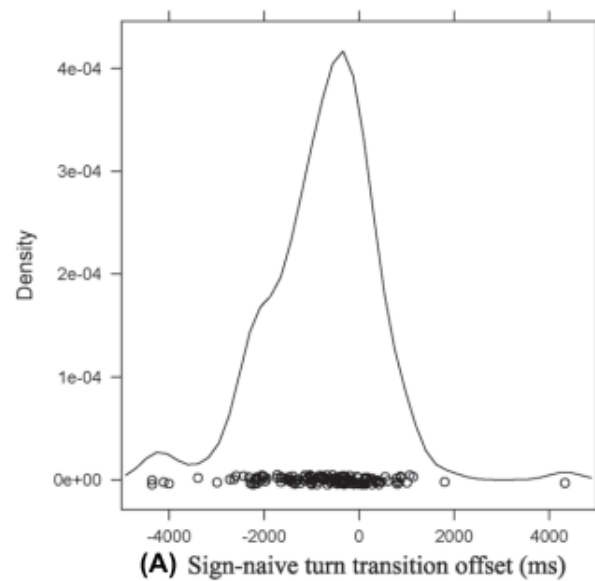


Figure A 16 Density plot (De Vos, Torreira & Levinson, 2015) sign naive

From: De Vos, C., Torreira, F., & Levinson, S. C. (2015). Turn-timing in signed conversations: Coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2015.00268>. Published under a CC BY 4.0 licence.

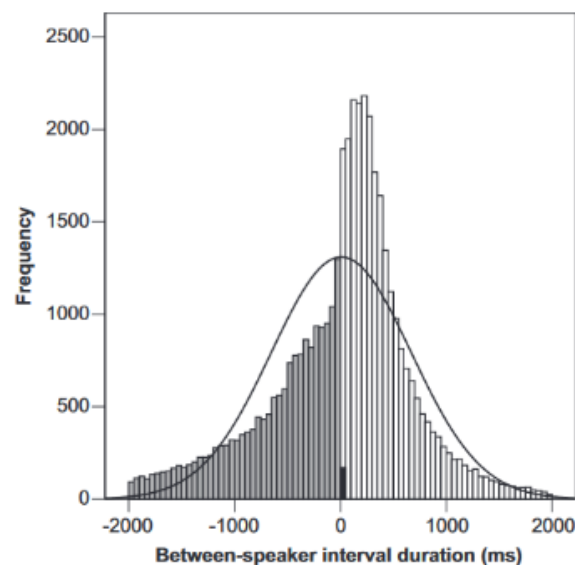


Figure A 17 Histogram with normal distribution overlaid (Heldner & Edlund, 2010) Dutch

Reprinted from Journal of Phonetics, 38/4, M. Heldner and J. Edlund, Pauses, gaps and overlaps in conversations, p555–568, Copyright (2010) Elsevier, with permission from Elsevier.

Left-skewed distributions

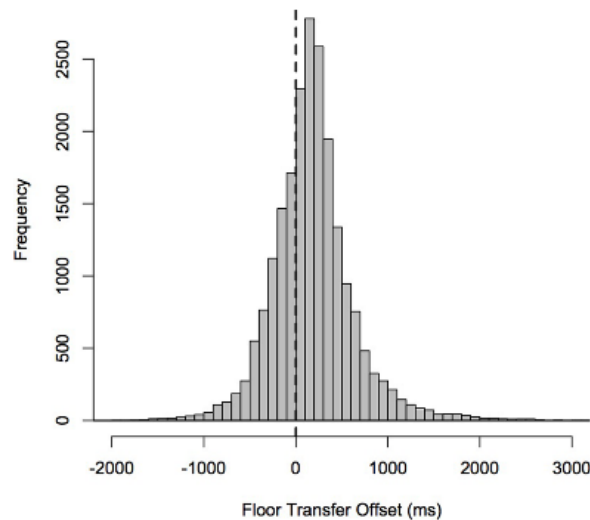


Figure A 18 Histogram (Levinson & Torreira, 2015) all

From: Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*.

<https://doi.org/10.3389/fpsyg.2015.00731>. Published under a CC BY 4.0 licence.

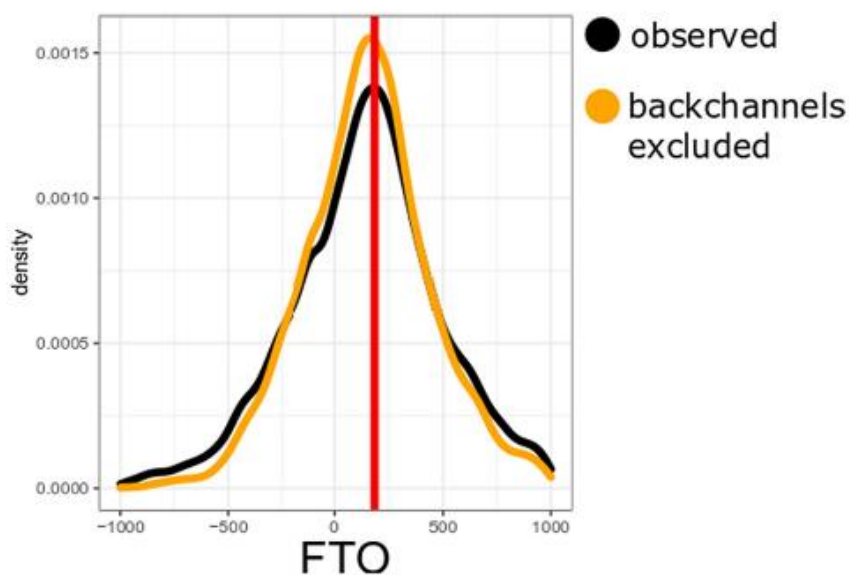


Figure A 19 Density plot (Pouw & Holler, 2022) ms

From: Pouw, W., & Holler, J. (2022). Timing in conversation is dynamically adjusted turn by turn in dyadic telephone conversations. *Cognition*, 222, 105015.

<https://doi.org/10.1016/j.cognition.2022.105015>. Published under a CC BY 4.0 licence.

Left-skewed distributions

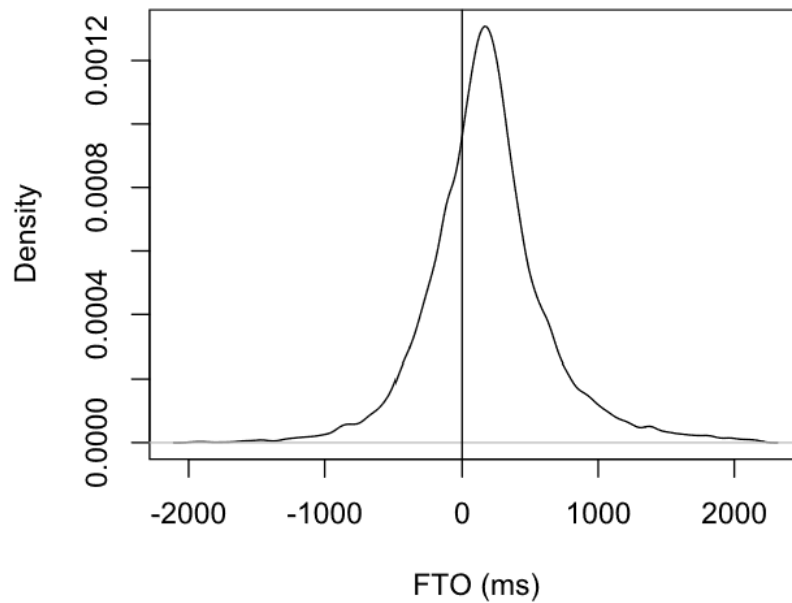


Figure A 20 Density plot (Roberts et al., 2015)

From: Roberts, 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(509). <https://doi.org/10.3389/fpsyg.2015.00509>. Published under a CC BY 4.0 licence.

Right-skewed distributions

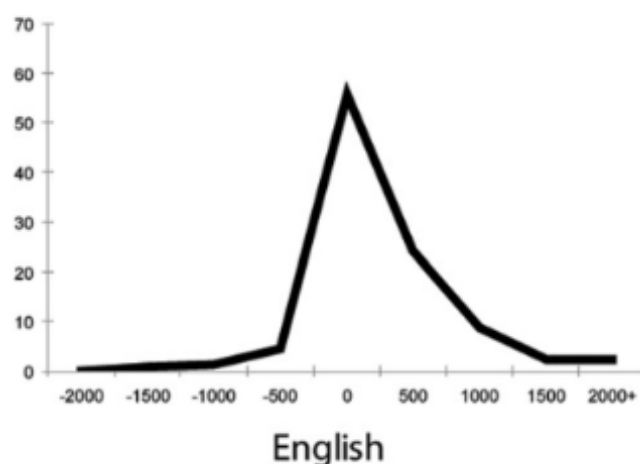


Figure A 21 Histogram (Stivers et al., 2009) English

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

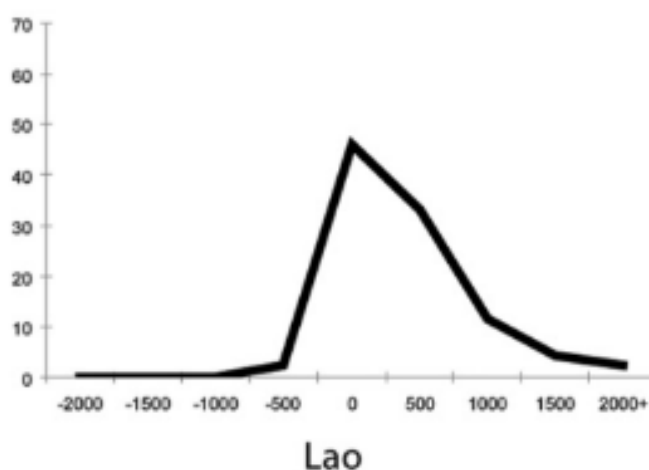


Figure A 22 Histogram (Stivers et al., 2009) Lao

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

Right-skewed distributions

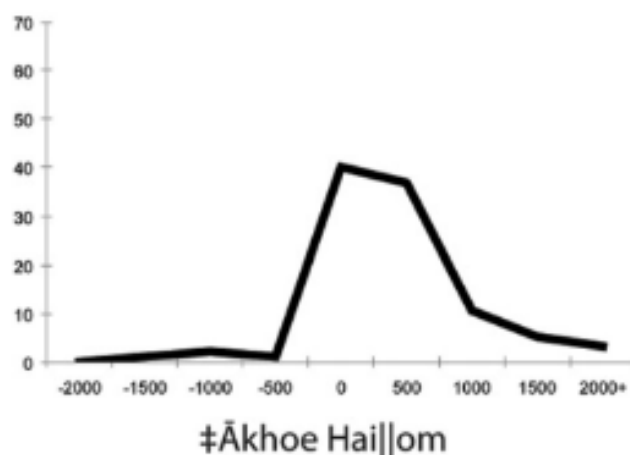


Figure A 23 Histogram (Stivers et al., 2009) #Akhoe Hai||om milliseconds, binsize = 100ms

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

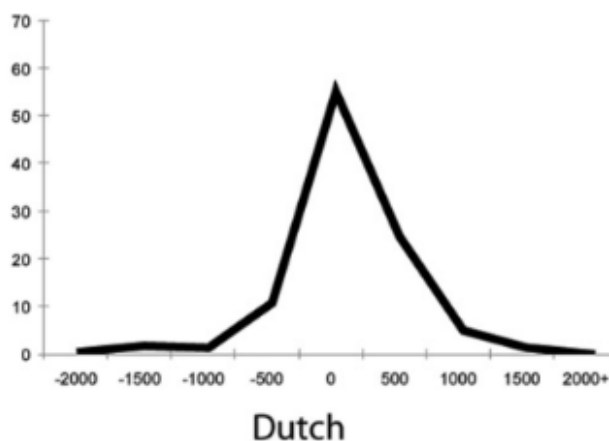


Figure A 24 Histogram (Stivers et al., 2009) Dutch

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

Right-skewed distributions

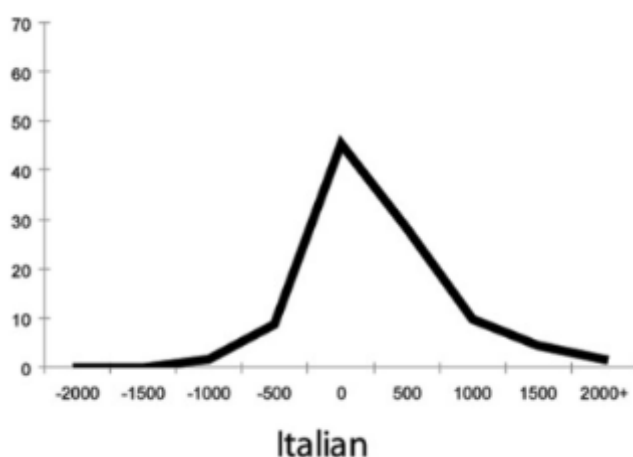


Figure A 25 Histogram (Stivers et al., 2009) Italian

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

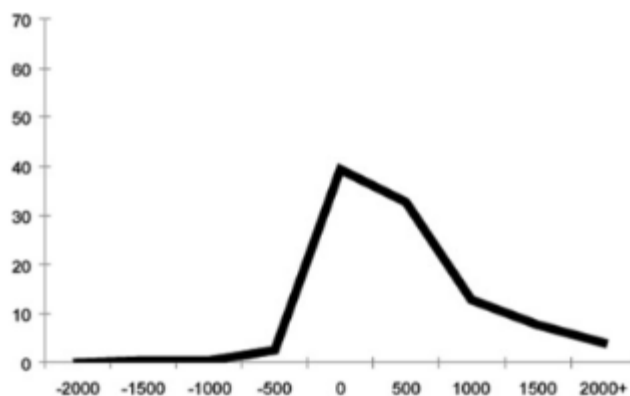


Figure A 26 Histogram (Stivers et al., 2009) Danish

From: Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

<https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>. Copyright belongs to the authors. Reprinted with permission.

Right-skewed distributions

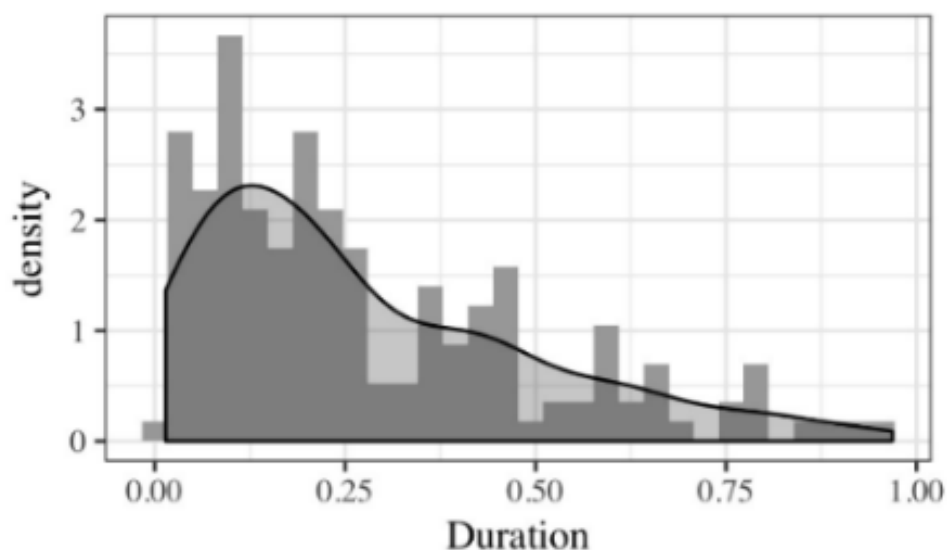


Figure A 27 Histogram and density plot, seconds (Buanzur et al., 2018)

From: Buanzur, T., Zellers, M., Namyalo, S., & Witzlack-Makarevich, A. (2018). A first investigation of the timing of turn-taking in Ruuli. *Proceedings of the Annual Conference of the International Speech Communication Association 2018, 2018-September*, 621–625. <https://doi.org/10.21437/Interspeech.2018-1254>. Copyright belongs to ISCA. Reprinted with permission from the first author.

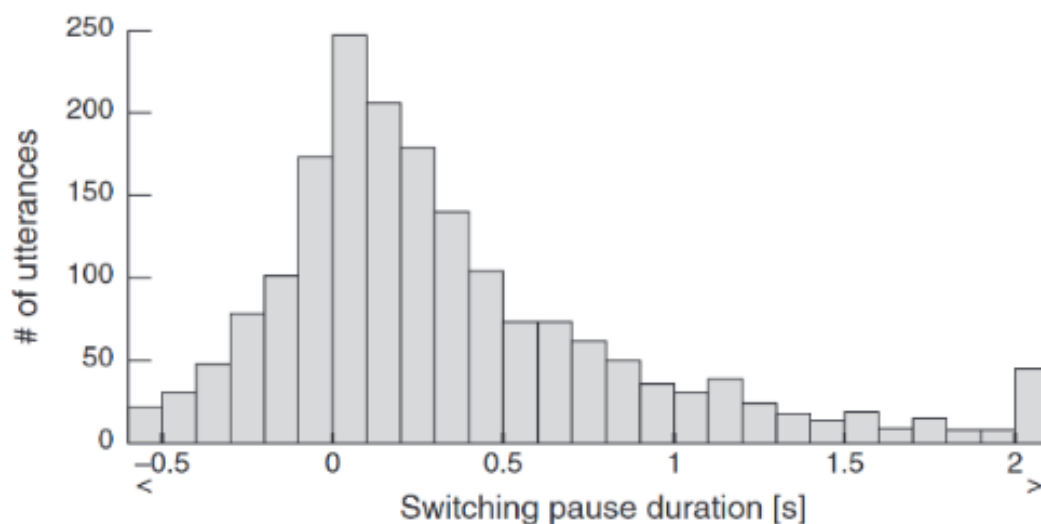


Figure A 28 Histogram (Mori, 2009)

From: Mori, H. (2009). An analysis of switching pause duration as a paralinguistic feature in expressive dialogues. *Acoustical Science and Technology*, 30(5), 376–378. <https://doi.org/10.1250/ast.30.376>. Copyright belongs to the author. Reprinted with permission.

Right-skewed distributions

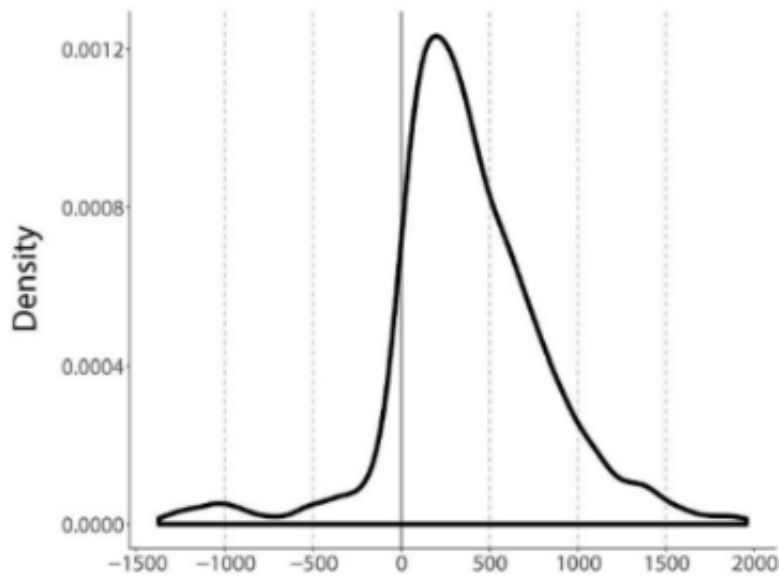


Figure A 29 Density plot, ms (Bögels, 2020)

From: Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, 203.

<https://doi.org/10.1016/j.cognition.2020.104347>. Published under a CC BY 4.0 licence.

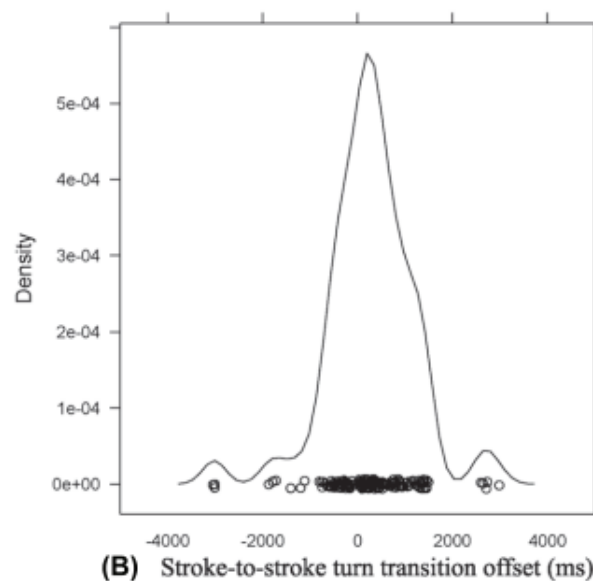


Figure A 30 Density plot (De Vos, Torreira & Levinson, 2015) stroke-to-stroke

From: De Vos, C., Torreira, F., & Levinson, S. C. (2015). Turn-timing in signed conversations: Coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*.

<https://doi.org/10.3389/fpsyg.2015.00268>. Published under a CC BY 4.0 licence.

Appendix A6. Rating Study Participant Information Sheet

Participant Information

Welcome to this study! Please read the following information carefully:

1. You are invited to take part in a research study.
2. This study is conducted by Damar Hoogland as part of her PhD research at Newcastle University and supervised by Dr. Laurence White from the School of Education, Communication & Language Sciences at Newcastle University and Dr. Kai Alter from the Biosciences Institute at Newcastle University.
3. The purpose of this study is to find out how people perceive conversations.
4. If you agree to take part in this study, you will be asked to listen to fragments of conversations and rate them. Then you will be asked to type your answer to a few questions.
5. You are free to decide not to participate. If you decide to participate you are free to withdraw at any time during the study without negative consequences for you. If you withdraw, all your data will be excluded.
6. All your responses will be kept anonymous. There will be no way for anyone to connect your name to your responses after you complete this study, not even the researchers. If you previously corresponded with Damar Hoogland about your participation time, she will delete this correspondence.
7. It will be impossible to identify you from any research reports that may be published.
8. We may decide to publish some of the data gathered in this study on an online and openly accessible repository for scientific data. It will be impossible for anyone to identify you from this data.
9. If you would like to receive a report of the findings of this research, a link to the data repository (if we decide to publish the data), or have any questions about this study, you can email Damar Hoogland at d.hoogland2@newcastle.ac.uk. She will then send you a summary of the findings once the research is completed.

10. Any concerns about this study should be addressed to the School of Education, Communication & Language Sciences Ethics Committee, Newcastle University via email to: ecls.researchteam@newcastle.ac.uk.

Feel free to pause this experiment to take a note of this email address. You can pause it by clicking ESC.

Consent form

- I agree to participate in this study, the purpose of which is to study people's perception of conversations.
- I declare that I have understood the nature and purpose of the research.
- I have read the participant information sheet and understand the information provided.
- I have been informed that I may decline to answer any questions or withdraw from the study without penalty of any kind.
- I have been informed that all of my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that the researchers may wish to share non-identifying information, e.g. my responses to questions, in a public repository. I understand that it will not be possible to identify me as an individual from this information alone.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is d.hoogland2@newcastle.ac.uk.
- I have recorded the investigator's email.

By clicking 'agree', I agree to the statements above. ☐ Agree ☐ Disagree

Appendix A7. Participant Instructions

English instructions

Instructions

You will hear fragments of interactions between two people. You will hear each interaction only once.

For each interactions, we would like to know whether you thought the reaction to the question sounded early, on time, or late.

After the sound has stopped playing, you can indicate your response on a scale. Please use the whole scale.

Please respond every time, even if you are not sure or if you don't have a strong opinion.

There are no right and wrong responses, so just use your gut feeling.

Sometimes you will hear "Please press W" instead of a fragment of an interaction. In those cases, please press the key with the letter 'w' instead of the 'next' button to continue.

You can take a break whenever you like.

The first four rounds are practice rounds.

English instructions repeat

Remember, for each interaction, we would like to know whether you thought the reaction to the question sounded early, on time, or late.

Please use the whole scale, and respond every time.

Whenever you hear "Please press W", please press the key with the letter 'W' to continue instead of the 'next' button.

There are no right and wrong responses, so just use your gut feeling.

You can take a break whenever you like.

Dutch instructions

Instructies

Je hoort straks een serie fragmenten van interacties tussen twee mensen. Je zult elk fragment één keer te horen krijgen.

Nadat je naar een interactie geluisterd hebt, willen wij graag weten of je vond dat de

reactie op de VRAAG vroeg, op tijd, of laat klonk.

Als het geluid opgehouden is kun je je antwoord aangeven op een schaal.

Gebruik alsjeblief de hele schaal, van VROEG naar LAAT.

Vul je antwoord in voor elk fragment, ook als je niet zeker bent of geen sterke mening hebt. Er zijn geen goede of foute antwoorden in dit onderzoek, dus gebruik je intuïtie.

Soms hoor je “druk SPATIE om door te gaan”, in plaats van een fragment van een interactie. In dat geval, gebruik dan de spatietoets om door te gaan naar het volgende fragment.

Je kunt pauze houden wanneer je wilt.

Je hoort nu eerst vier voorbeelden.

Dutch instructions repeat

Nogmaals: voor elk fragment willen we graag weten of je vond dat de de reactie op de VRAAG vroeg, op tijd of laat klonk.

Gebruik de hele schaal, en beoordeel elk fragment. Als je “druk SPATIE om door te gaan” hoort, druk dan de spatietoets in om door te gaan in plaats van op “doorgaan” te klikken. .

Er zijn geen goede of foute antwoorden, dus gebruik je intuïtie. .

Je kunt pauze houden wanneer je wilt.

Appendix A8. A Continuous Analysis of Means Standardised Ratings

In addition to the categorical analysis in Section 5.3.4, Mean Standardised Rating was also analysed as a continuous variable, to test whether it has a continuous modulating effect on the relationship between Question Speech Rate and TTT in this dataset.

Mixed effect linear regression models were built for the Full dataset and Gaps (observations where $TTT \geq 0ms$), predicting TTT from Mean Standardised Ratings, Question Speech Rate, and Final Rhyme Duration (for Gaps), and all possible two-way and three-way interactions. In a backward stepwise selection procedure, the significance of the contribution of each predictor was tested using Likelihood Ratio Tests, starting with the three-way interaction, then the two-way interactions, and then the main predictors. Non-significant predictors were dropped from the model at each step, and significant predictors ($\alpha = .05$) or predictors that showed a trend towards significantly contributing to the model fit ($\alpha = .1$) were retained. A random intercept for Speaker – i.e., the speaker who uttered the response in the question-response sequence – was also included in all models.

If the relationship between turn timing and speech rate is modulated by the perceived timeliness of turns, the interaction between Question Speech Rate and Mean Standardised Rating is expected to significantly improve the model fit. Specifically, the perception-action entrainment hypothesis would predict a positive interaction coefficient: there should be a negative relationship between Question Speech Rate and TTT in timely turns, which is less steep when turns are perceived as late, and steeper when turns are perceived as early.

In the Full dataset, Question Speech Rate $\times$ Mean Standardised Rating did not improve the model fit, $\chi^2(1) = .11$, $p = .74$, and neither did Question Speech Rate, $\chi^2(1) = .67$, $p = .413$. Only Mean Standardised Rating significantly improved the model fit. The final model is summarised in Table A 9 (next page).

In Gaps, the three-way interaction Question Speech Rate $\times$ Final Rhyme Duration $\times$ Mean Standardised Rating and two-way interactions Question Speech Rate $\times$ Mean Standardised Rating, and Final Rhyme Duration $\times$ Mean Standardised Rating, significantly improved the model fit, and Question Speech Rate $\times$ Final Rhyme Duration marginally improved the model fit, so they were retained in the model while testing the main effects.

None of the main effects significantly improved the model fit: Question Speech Rate, $\chi^2 (1) = 2.06, p = .152$; Final Rhyme Duration, $\chi^2 (1) = 2.38, p = .123$; or Mean Standardised Rating, $\chi^2 (1) = .94, p = .333$. The final model is summarised in Table A 10.

In the Gaps dataset, the coefficients of the Question Speech Rate $\times$ Mean Standardised Rating is *positive*, as the perception-action entrainment hypothesis entails, and so is the Final Rhyme Duration $\times$ Mean Standardised Rating. However, these interaction terms are difficult to interpret without main effects. Visualisations with categorically split data in Figure A 31 suggest that the relationship between Question Speech Rate and TTT, instead of becoming less negative as Mean Standardised Rating increases, turns over from slightly negative to slightly positive to negative. The other two predictors similarly show a more subtle pattern. This may indicate that the model has overfitted.

Predictor	Coefficient	Std. Error	χ^2	p-value
Intercept	254.45	23.60		
Mean Standardised Rating	519.96	33.65	170.15	<.001

Table A 9 Model coefficients, standard errors and LRT statistics for the final model for the Full dataset with Mean Standardised Rating as continuous predictor.

Predictor	Coefficient	Std. Error	χ^2	p-value
Intercept	452.31	41.67		
Question Speech Rate $\times$ Mean Standardised Rating	81.40	13.40	4.41	.036
Final Rhyme Duration $\times$ Mean Standardised Rating	1.94	.51	4.27	.039
Question Speech Rate $\times$ Final Rhyme Duration	-.05	.03	3.46	.063
Question Speech Rate $\times$ Final Rhyme Duration $\times$ Mean Standardised Rating	-.40	.13	4.08	.043

Table A 10 Model coefficients, standard errors and LRT statistics for the final model for the Gaps dataset with Mean Standardised Rating as continuous predictor.

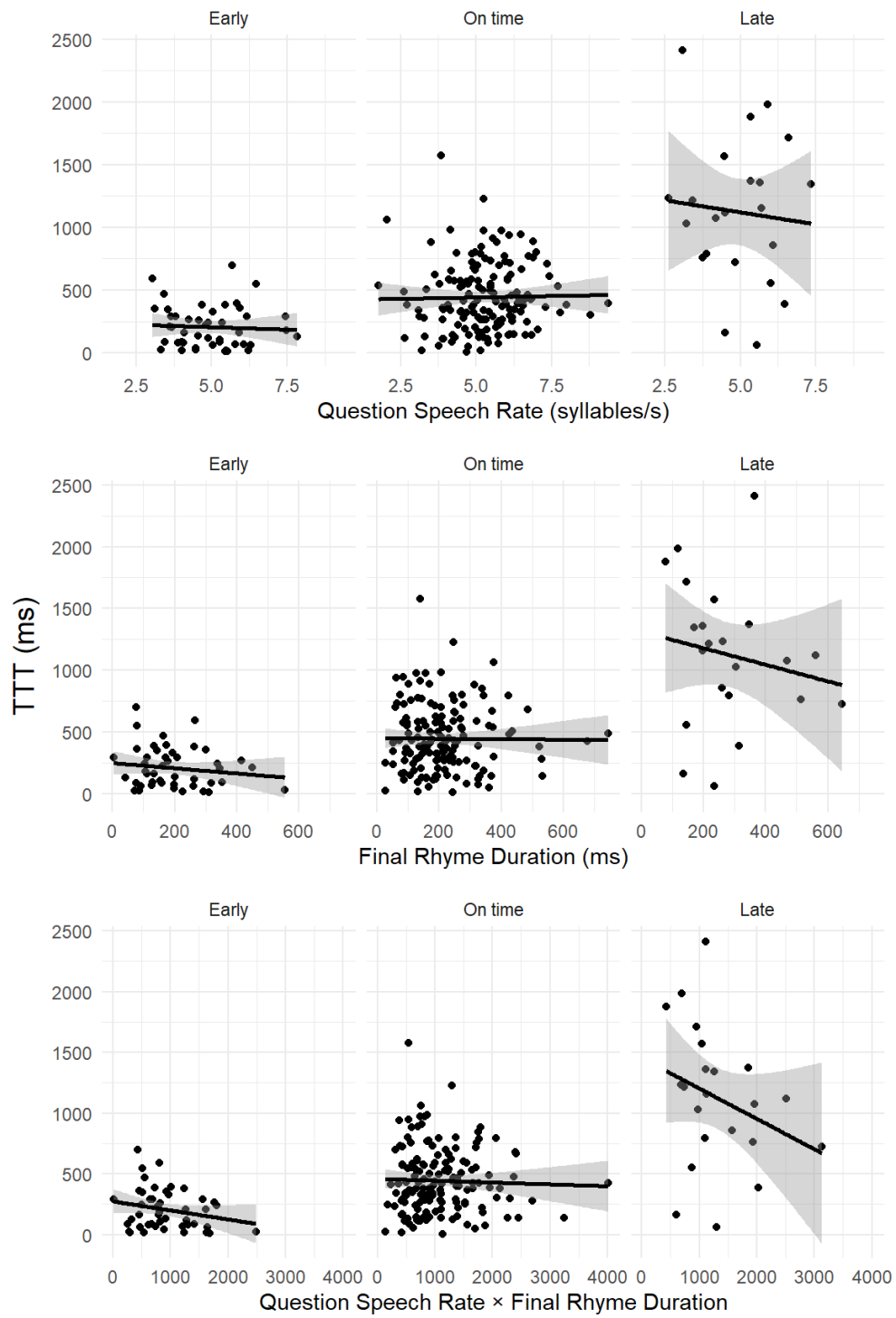


Figure A 31 Illustration of the interaction effects with categorically split data. Subsets consist of one-third of the Mean Standardised Rating scale. The Question Speech Rate $\times$ Final Rhyme Duration $\times$ Mean Standardised Rating is visualised as the product of Question Speech Rate and Final Rhyme Duration against TTT.