Trinity College Dublin, The University of Dublin

School of Computer Science and Statistics

Doctor of Philosophy

Self-Adaptation to Interlingual Machine Mediated Communication

---

Author:
Hayakawa Akira

Supervisors:
Dr. Saturnino Luz
Dr. Carl Vogel

A thesis submitted to Trinity College Dublin, The University of Dublin
October 8, 2021

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Hayakawa Akira Campbell

Dated: October 8, 2021

Acknowledgements

I would like to thank my supervisors Dr. Saturnino Luz and Dr. Carl Vogel for their belief, support and guidance throughout my research. Without their advice and insight this thesis would not have been possible. I am sure you expected more, but I hope that this meets some of your expectations. Dr. Luz provided the idea of collecting the ILMT-s2s corpus and studying the biosignals, which I was not able to fulfill. I hope Dr. Haider's use of the corpus achieved at least some of what you initially wanted. Dr. Vogel kindly took me in after Dr. Luz's departure to Edinburgh University, and since I am not the best at communicating, I am sure he had his troubles understanding what I was doing and what was going on in my mind. To those I have offended during the time I have been here at Trinity College Dublin, The University of Dublin, I am sorry. It was not intentional. Finally I express my gratitude towards my Viva Voice examiners, Professor Khurshid Ahmad and Professor Laszlo Hunyadi, for their questions and guidance they provided to complete this thesis.

Hayakawa Akira

Trinity College Dublin, The University of Dublin

October 8, 2021

— Note —

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin, The University of Dublin.

Abstract

One of Claude Shannon's greatest contributions to computer science is considered to be the Information Theory. Put simply, what is the minimum amount of information (data) required to obtain an acceptable outcome. Computer mediated communication has been around since the mid-twentieth century, but it is only recently that the balance of requirement and output have started to level out and it has been incorporated into the life of the masses. The next forecasted step will be to incorporate Speech-to-Text Machine Translation (S2S-MT) into the video-conferencing network for the masses.

The possibility to learn is a privilege of life and hence not always fully equal. This thesis is my contribution to this task of equality, where I have; (i) provided much-needed Speech-to-Speech Machine Translation data which has been systematically collected in two typologically distinct languages (English & Portuguese) with a close to even gender balanced, (ii) used the monolingual 'gold standard' Edinburgh University's HCRC corpus as a basis for selecting two key segments (or Map Tasks) in the corpus for creating a dialogue between Information Giver in one language and the response of the Information Follower in another language by the mediation of a computer system, (iii) jointly created a program pipeline which comprises speech recognition, speech transcription, text-to-speech, and machine translation systems together with a good graphical-user interface and (iv) provided experimental design which include necessary statistical tests.

From the collected data, I have; (i) identified two map tasks from the HCRC corpus that are easy to route, but long in duration, use opposing directions in the route and have a close to even gender balance, (ii) identified that the speed in which the subjects talk is significantly different between computer mediated dialogues and the 'gold standard' non-mediated data, (iii) identified that with computer mediated dialogues the speech rates may differ between gender and (iv) identified that the length and the type of information the subjects communicate differs in length and type. There is more to be done, but hopefully information continues to increase in the open domain.

My fear that underlines the motivation of this study is the adaptability of the untrained people that will use this communication method, and the limitation of fundings that will create an imbalance in the quality of the used infrastructure (network speeds, affordable equipment), and the precision of the S2S-MT used. Practice makes perfect, is a phrased used often, but when practice is not possible, the tools should provide the helping hand. To do so, we must see how the human is interacting within a computer-mediated dialogue setting and compare it to the non-mediated dialogue setting.

Contents

Declaration i

Acknowledgements ii

Abstract iii

Contents iv

List of Figures xi

List of Tables xix

List of Abbreviations xxvi

1 Introduction 1

1.1 Motivations . . . 1

1.2 The Question . . . 2

1.2.1 Task Oriented Interactions . . . 3

1.2.2 Research Question . . . 4

1.3 Contributions and Key Findings . . . 5

1.4 Associated Publications . . . 7

1.4.1 Publications that Used the Data Collected in this Study . . . 7

1.4.2 Publications and Presentations Related to this Study . . . 7

1.5 Overview of the Research . . . 12

1.5.1 Introduction . . . 12

1.5.2 Structure of the Thesis . . . 14

2 Literature Review 18

2.1 Speech-to-Speech Translation Systems . . . 18

2.1.1 S2S-MT Projects to Date . . . 19

2.1.2 Map Task Corpora . . . 20

2.2 Aspects Influencing and Being Influenced by S2S-MT Systems . . . 20

2.2.1 Speech Rate . . . 21
2.2.2 Dialogue Acts and Communication Medium . . . 22
2.2.3 Perceived Emotion (Cognitive State) . . . 24
2.2.4 Different Types of Talk . . . 25
2.3 Conclusion . . . 26

3 The Datasets **27**
3.1 Introduction . . . 27
3.2 The Non-Mediated Dialogue Dataset . . . 27
3.2.1 Introduction of the HCRC Edinburgh Map Task Corpus . . . 27
3.2.2 The Selection of the Two Maps . . . 29
3.2.3 The Subjects of Map 01 & Map 07 . . . 30
3.2.4 The Dialogues of Map 01 & Map 07 . . . 31
3.2.5 The Transcription & Annotation Data . . . 31
3.2.6 Summary of the Data . . . 33
3.3 The Computer-Mediated Dialogue Dataset . . . 34
3.3.1 The Subjects of the Data . . . 34
3.3.2 The Dialogues . . . 35
3.3.3 The Transcription & Annotation of the Data . . . 35
3.3.4 The Subject Survey . . . 40
3.3.5 Summary of the Data . . . 41
3.4 Summary . . . 42

4 Data Created for this Study **43**
4.1 Introduction . . . 43
4.1.1 Main Design Concept of the Corpus Collection . . . 44
4.1.2 Initial Objectives of the Corpus . . . 44
4.1.3 Structure of this Chapter . . . 45
4.2 The Communication Tool . . . 46
4.2.1 Basic Structure of the System . . . 46
4.2.2 The Recognition Method . . . 47
4.2.3 The Transcription and Translation Method . . . 50
4.2.4 The Output Method . . . 51
4.2.5 The User Interface . . . 54
4.2.6 System Log (XML) . . . 57
4.3 The Data Collection . . . 59
4.3.1 Separation of Subjects . . . 59
4.3.2 The Subjects . . . 59

4.3.3 Recording Environment . . . 60
4.3.4 Recording Devices . . . 61
4.3.5 The Map Task . . . 62
4.3.6 ILMT-s2s System User Survey . . . 72
4.3.7 Biosignal Recording . . . 75
4.3.8 Eye Gaze Recording . . . 75
4.4 Processing the Data . . . 76
4.4.1 Synchronisation . . . 77
4.4.2 Transcription . . . 79
4.4.3 Annotation . . . 82
4.4.4 Calculating the Speech Rate . . . 96
4.4.5 Calculating the ASR Performance (WER & F-Score) . . . 97
4.4.6 20 Frame/Window Normalization . . . 100
4.5 Summary . . . 102

5 Speech Rate Difference 103

5.1 Introduction . . . 103
5.1.1 Methodological Reason to Using Speech Rate Difference . . . 103
5.1.2 Practical Reason to Using Speech Rate Difference . . . 104
5.1.3 The Dataset from the Perspective of *Words per Minute* . . . 105
5.1.4 The Point of Concern with *Words per Minute* . . . 105
5.1.5 The Point of Concern with *Syllables per Second* . . . 107
5.1.6 Other Concerns . . . 107
5.2 Dataset . . . 108
5.3 Preparations for Calculating the Speech Rate Difference . . . 108
5.3.1 The Dataset from the Perspective of the TTS Output . . . 110
5.4 Calculating the Speech Rate Difference . . . 112
5.4.1 Creating the Speech Rate Difference . . . 112
5.4.2 Simplifying the Speech Rate Difference . . . 113
5.5 Definition . . . 113
5.6 Conclusion . . . 113
5.7 Summary . . . 115

6 The Three Talk Types 117

6.1 Introduction . . . 117
6.2 Dataset . . . 118
6.3 Methods . . . 118
6.3.1 Talk Type Analysis of Each Subject . . . 118

6.3.2 Talk Type Analysis by Utterance Word Count . . . 118
6.3.3 Talk Type Analysis by Grouping . . . 119
6.4 Results . . . 119
6.4.1 Talk Type Analysis of Each Subject . . . 121
6.4.2 Talk Type Analysis by Utterance Word Count . . . 123
6.4.3 Talk Type Analysis by Grouping . . . 128
6.5 Discussion . . . 130

7 Speech Rate Differences Between Datasets 133

7.1 Introduction . . . 133
7.2 Dataset . . . 134
7.3 Methods . . . 134
7.3.1 Visualisation of the Dialogues' Speech Rate . . . 134
7.3.2 Comparison Between *Off-Talk Other* and the Non-Mediated Dialogues . . . 134
7.3.3 Identifying Speech Rate patterns in the Computer-Mediated Dialogues . . . 135
7.3.4 Identifying Speech Rate patterns in the Non-Mediated Dialogues . . . 135
7.3.5 Comparison Between *On-Talk* and the Non-Mediated Dialogues . . . 135
7.4 Results . . . 136
7.4.1 Visualisation of the Dialogues' Speech Rate . . . 136
7.4.2 Comparison Between *Off-Talk Other* and the Non-Mediated Dialogues . . . 138
7.4.3 Identifying Speech Rate patterns in the Computer-Mediated Dialogues . . . 139
7.4.4 Identifying Speech Rate patterns in the Non-Mediated Dialogues . . . 139
7.4.5 Comparison Between *On-Talk* and the Non-Mediated Dialogues . . . 140
7.4.6 Verifications with Adjusted Values . . . 140
7.5 Summary . . . 142
7.6 Discussion . . . 144

8 Dialogue Acts Differences Between Datasets 148

8.1 Introduction . . . 148
8.2 Dataset . . . 148
8.3 Methods . . . 148
8.3.1 Comparison of Dialogue Act Utterance Count and Word Count . . . 149
8.3.2 Visualisation of the Dialogue Patterns . . . 149
8.4 Results . . . 150
8.4.1 Comparison of Dialogue Act Utterance Count and Word Count . . . 150
8.4.2 Visualisation of the Dialogue Patterns . . . 151
8.5 Ad-hoc Analysis . . . 153
8.6 Summary . . . 154

8.7 Discussion . . . 159

9 ASR Performance & Speech Rate 161

9.1 Introduction . . . 161
9.2 Dataset . . . 162
9.3 Methods . . . 162
9.3.1 Calculating the ASR Performance (WER & F-Score) . . . 162
9.3.2 Visualisation of the Speech Rate and ASR Performance . . . 162
9.3.3 Speech Rate & ASR Performance Analysis . . . 162
9.4 Results . . . 163
9.4.1 Calculating the ASR Performance (WER & F-Score) . . . 163
9.4.2 Visualisation of the Speech Rate and ASR Performance . . . 164
9.4.3 Speech Rate & ASR Performance Analysis . . . 165
9.5 Ad-hoc Analysis . . . 170
9.5.1 Removal of Dialogues with System Crashes . . . 170
9.5.2 Analysis Results After System Crash Dialogues are Removed . . . 172
9.6 Summary . . . 173
9.7 Discussion . . . 173

10 ASR Performance & Perceived Emotions Analysis 176

10.1 Introduction . . . 178
10.2 Dataset . . . 179
10.3 Methods . . . 179
10.3.1 Linking Perceived Emotion Labels with Communication Items . . . 179
10.3.2 Direction of the Analysis . . . 180
10.3.3 Analysis by Grouping . . . 180
10.4 Results . . . 181
10.4.1 Balancing Perceived Emotion Labels' Count . . . 181
10.4.2 Linking Perceived Emotion Labels with Communication Items . . . 182
10.4.3 Direction of the Analysis . . . 182
10.4.4 Analysis by Grouping . . . 184
10.5 Summary . . . 189
10.6 Discussion . . . 189

11 User Survey Study 191

11.1 Introduction . . . 191
11.2 Dataset Used for the Analysis . . . 193
11.3 Methods of the Analysis . . . 193

11.3.1 Visualisation of the 7 Point Likert Scale Results . . . 193
11.3.2 Analysis of the 7 Point Likert Scale Results . . . 193
11.3.3 Analysis of Results Divided into Groups . . . 193
11.4 Results of the Analysis . . . 194
11.4.1 Visualisation of the 7 Point Likert Scale Results . . . 194
11.4.2 Analysis of Results as a Whole — One Group . . . 195
11.4.3 Analysis of Results Divided into Groups . . . 198
11.5 Summary . . . 211

12 Discussions **213**

12.1 Question One . . . 213
12.2 Question Two . . . 214
12.3 Question Three . . . 214
12.4 Future works . . . 217
12.5 Conclusion . . . 218

13 Conclusions **219**

13.1 Introduction . . . 219
13.2 Thesis Contribution and Findings . . . 220
13.3 Looking back at the ILMT-s2s corpus . . . 222
13.4 Conclusions . . . 223

Appendix A ILMT-s2s corpus **224**

A.1 ILMT-s2s corpus Data Summary . . . 224
A.2 ILMT-s2s corpus Data Processing Workflows . . . 226
A.3 Speech Rate & ASR Precision of All Dialogues . . . 229
A.4 Speech Rate of The Three Talk Types . . . 245
A.5 ILMT-s2s User Survey Details and Replies . . . 254
A.5.1 User Survey Statements and Questions . . . 254
A.5.2 ILMT-s2s User Survey Layout . . . 255
A.5.3 ILMT-s2s User Survey Replies . . . 260
A.5.4 The Fifteen 7 Point Likert Scale Statements' Open Text Field Comments . . . 261
A.5.5 The Five Open Text Field Questions' Comments . . . 270
A.6 ILMT-s2s User Survey Results . . . 277
A.6.1 User Survey Statements' Median and Mean . . . 278
A.6.2 User Survey Stacked Bar Charts and Density Plots . . . 280
A.6.3 User Survey Significance Test Results . . . 288

Appendix B The HCRC Edinburgh Map Task corpus **298**

B.1 The Four Route Types of the HCRC Edinburgh Map Task corpus Maps 298

B.2 The Map Differences Between Information Giver and the Information Follower 299

B.3 Speech Rate of 16 HCRC Edinburgh Map Task corpus Dialogues 303

B.3.1 HCRC Edinburgh Map Task corpus Dialogues After 20 Window Normalization 303

B.3.2 Line Charts of HCRC Edinburgh Map Task corpus Dialogues' Speech Rate . . 309

Appendix C Dialogue Act Summary **325**

Glossary **337**

Bibliography **345**

List of Figures

1.1 *Communication flow of S2S-MT system — English to Japanese example* 2

1.2 *Overview of the Analysis Conducted in this Study* 12

1.3 *Overview of the Data Used for the Analysis in this Study* 13

3.1 *The 4 route patterns of the HCRC Edinburgh Map Task corpus maps* 30

3.2 *Image of subject pairs and recording situation with deviation score and duration of dialogues* 30

3.3 *Bar chart of dialogue act occurance within the HCRC Edinburgh Map Task corpus* . . 32

3.4 *Speech rate difference density plot of Information Giver and Information Follower dialogue data* 33

3.5 *Image of subject pairs and recording situation* 34

3.6 *Bar chart of dialogue acts: Total of On-Talk and Off-Talk variant's combined as well as separated — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item* 36

3.7 *Distribution of perceived emotions within the ILMT-s2s corpus — Categorised by group. The "n" indicates the number of labels within each grouping of "Setting", "Role", and "Language".* 37

3.8 *Speech rate difference density plot of English and Portuguese dialogue data* 38

3.9 *WER: ASR performance density plot of Eng. and Portuguese dialogue data* 39

3.10 *Reverse F-Score: ASR performance density plot of Eng. and Portuguese dialogue data* 39

3.11 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale replies* . 41

4.1 *ILMT-s2s System used to collect the data* 46

4.2 *Turn flow diagram when using the ILMT-s2s System. Items (a & b) related to this section are highlighted.* 47

4.3 *Turn flow diagram when using the ILMT-s2s System. Items (b) related to this section are highlighted.* 50

4.4 *Turn flow diagram when using the ILMT-s2s System. Items (a, b, c & d) related to this section are highlighted.* 51

4.5 *Turn flow diagram when using the ILMT-s2s System. Items (a, b, c & d) related to this section are highlighted.* 54
4.6 *User Interface of the ILMT-s2s System* 55
4.7 *System process indicator* 56
4.8 *Interlocutor action indicators* 56
4.9 *Turn flow diagram when using the ILMT-s2s System. Items (b & c) related to this section are highlighted.* 57
4.10 *Image of subject pairs and recording situation* 60
4.11 *Subjects during recordings* 61
4.12 *An example of the Information Giver and Information Follower maps* 63
4.13 *An example of the differences found in the maps — Definition from "Language and Speech" Journal Paper Article: The HCRC Map Task Corpus. Anderson et al. (1991)* 64
4.14 *An example of the differences found in the maps — Definition from the HCRC Map Task Corpus Website (Last verified: 2017/10/10)* 67
4.15 *Redefining the definition of identical location* 69
4.16 *Maps used in the ILMT-s2s Data Collection with differences highlighted* 71
4.17 *The 4 route patterns of the HCRC Edinburgh Map Task corpus maps* 71
4.18 *7 point Likert scale user survey actual layout example* 73
4.19 *Open text field question user survey actual layout example* 74
4.20 *10 – 20 location system layout map* 76
4.21 *Image of changing the data to a corpus* 76
4.22 *Screenshot of data synchronisation process* 77
4.23 *Example of Wavesurfer transcribing configuration for right channel audio* 80
4.24 *Example of ELAN perceived emotion annotation* 83
4.25 *Image of dialogue segmentation — Actual timing from Dialogue 15* 84
4.26 *Distribution of perceived emotions within the ILMT-s2s corpus — Categorised by group. The "n" indicates the number of labels within each grouping of "Setting", "Role", and "Language".* 84
4.27 *Example of ELAN dialogue act annotation* 86
4.28 *Bar chart of dialogue acts: Total of On-Talk and Off-Talk variant's combined as well as separated — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item* 86
4.29 *Bar chart of dialogue acts: Proportion of On-Talk and Off-Talk variant's combined as well as separated, with the total of each variant being 100% — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item* 88
4.30 *Example of ELAN facial expressions and head movement annotation* 89
4.31 *Example of segmentation irregularity* 93

4.32 *Representation of ELAN data files* 95
4.33 *Speech rate difference density plot of English and Portuguese dialogue data* 96
4.34 *Levenshtein distance between human transcription and ASR result* 97
4.35 *Precision (P): Ratio of matching words in the alignment to the ASR result* 97
4.36 *Recall (R): Ratio of matching words in the alignment to the transcribed text* 98
4.37 *F-Score (F_0): Harmonic mean of P and R* 98
4.38 *WER: ASR performance density plot of Eng. and Portuguese dialogue data* 99
4.39 *Reverse F-Score: ASR performance density plot of Eng. and Portuguese dialogue data* 99
4.40 *ILMT-s2s corpus utterance count and duration comparison* 100
4.41 *Speech rate related and ASR accuracy for all subjects combined — Dialogue duration standardised into 20 frame windows* 101

5.1 *Mdn wpm for the six provided samples — Subject utterance* 106
5.2 *Transcription image from Praat* 109
5.3 *Scatter plot of ILMT-s2s corpus utterance durations by word count* 110
5.4 *Mdn wpm for the six provided samples — TTS output* 111
5.5 *Scatter plot of word count by subject utterance duration, word count by TTS output duration, and subject utterance duration by TTS output duration* 112
5.6 *Example speech rate comparison — Subject utterance & TTS output* 115
5.7 *Example of speech rate duration difference by role and gender* 116
5.8 *Example of speech rate comparison of speakers — Y-axis is the speech rate difference, X-axis is the frame when all dialogues are divided into 20 frames, using the method discribed in § 4.4.6* 116

6.1 *Data sheet with On-Talk highlighted in yellow and utterances slower than the TTS reference output in red. The actual values in the data sheet rows do not matter for this illustration.* 120
6.2 *Boxplot of all word, single word, and multi word utterance speech rates of all thirty subjects* 124
6.3 *Boxplot of all word, single word, and multi word utterance speech rates of all thirty subjects* 125
6.4 *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.* . . 129
6.5 *10 sec. interval speech rate difference comparison for first 100 sec. of all dialogues* . . . 131
6.6 *First (n) utterances with 100% accurate ASR results and following ($5 - n$) utterances with ASR errors* 131
6.7 *Talk type speech rate difference comparison separated by IG, IF and dialogue quarter* . 132

7.1 *Mean speech rate of the HCRC Edinburgh Map Task corpus (n = 24 subjects and 16 dialogues) and the ILMT-s2s corpus (n = 15 subjects and 15 dialogues)* 137
7.2 *Speech rates boxplots of both datasets* 138
7.3 *Speech rate comparison with the values from the ILMT-s2s corpus shifted up so that the median male speech rate matches that of the HCRC Edinburgh Map Task corpus (ILMT-s2s corpus values raised by 41.19%)* 141
7.4 *Density plot of gender speech rates in both datasets — Without value adjestments* . . . 143
7.5 *Density plot of gender speech rates in both datasets — After ILMT-s2s corpus values raised by 41.19% to align the median of all values with the HCRC Edinburgh Map Task corpus* 143
7.6 *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.* . . 144
7.7 *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.* . . 145
7.8 *Scatter plot of ILMT-s2s corpus utterance durations by word count* 146
7.9 *Scatter plot of ILMT-s2s corpus utterance durations by word count* 147

8.1 *Histogram & density plot of word count per utterance by dataset & role.* 149
8.2 *Boxplot of dialogue act word count by dataset and role* 151
8.3 *Density plot — Difference of the frequency of x word count utterances per dialogue act per dataset/role.* 151
8.4 *Ratio comparison of dialogue act count and word count of roles between datasets* . . . 152
8.5 *Bar chart role ratio comparison of dialogue acts within the HCRC Edinburgh Map Task corpus* 153
8.6 *Bar chart role ratio comparison of dialogue acts within the ILMT-s2s corpus* 154
8.7 *Bar chart role ratio comparison of dialogue acts word count within datasets* 154
8.8 *Frequency of occurrence differences between the datasets: Grey highlighted section indicate 90% coverage.* 155
8.9 *Bar chart role ratio comparison of dialogue acts within HCRC Edinburgh Map Task corpus (top and middle) and boxplot comparison of said dataset dialogue act word count (bottom).* 156
8.10 *Bar chart role ratio comparison of dialogue acts within ILMT-s2s corpus (top and middle) and boxplot comparison of said dataset dialogue act word count (bottom).* 157
8.11 *Bar chart of dialogue acts: Total of On-Talk and Off-Talk variant's combined as well as separated — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item* 160

9.1 *ASR performance density plot of English and Portuguese dialogue data* 163
9.2 *Speech rate related and ASR accuracy for all subjects combined — Dialogue duration standardised into 20 frame windows* 164

9.3 *ILMT-s2s corpus On-Talk utterance count per dialogue* 166
9.4 *Plot of subject's speech rate and Reverse F-Score by gender (20 frame window)* 167
9.5 *Plot of subject's speech rate and Reverse F-Score by role (20 frame window)* 168
9.6 *Plot of subject's speech rate and Reverse F-Score by setting (20 frame window)* 169
9.7 *Boxplot of speech rate adaptation to the ILMT-s2s System by group (20 frame window)* 169
9.8 *Boxplot of Reverse F-Score by group (20 frame window)* 170
9.9 *Speech rate density plot of dialogues w/ and w/o crashes (20 frame window)* 171
9.10 *Boxplot of speech rate adaptation to the ILMT-s2s System by group (20 frame window: Boxplots with "w/o Crash" included are displayed behind the boxplots with the "w/ Crash" removed)* 172
9.11 *Boxplot of ILMT-s2s System ASR Reverse F-Score by group (20 frame window: Boxplots with "w/o Crash" included are displayed behind the boxplots with the "w/ Crash" removed)* 172
9.12 *Plots of subject's speech rate and Reverse F-Score — Dialogues with crashes* 174

10.1 *Stacked bar chart of subject survey Likert scale replies — Role, setting and language* . 178
10.2 *Distribution of perceived emotions within the ILMT-s2s corpus — Summarised by groupings role, setting and language* 179
10.3 *Linkage example of perceived emotion and communication item* 180
10.4 *Flow of hypothesis (H1 and H2) — WER value is the same what ever the preceding or following perceived emotion may be (amused, frustrated, or surprised).* 181
10.5 *Linkage example of perceived emotion and communication item* 182
10.6 *Flow of Hypothesis (H1 and H2) — WER value is the same what ever the preceding or following cognitive state may be (Amused, Frustrated, or Surprised).* 183
10.7 *Plot of Tukey HSD test results of* $H_0{}^2$ *ANOVA output. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)* 184
10.8 $H_0{}^2$ *linkage example of perceived emotion and communication item* 184
10.9 *Bar chart of English subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused) divided by role and setting* 185
10.10 *Bar chart of Portuguese subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused) divided by role and setting* 185
10.11 *Plot of Tukey HSD test results of "w/ Video–En–On-Talk" WER ANOVA output. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)* 187
10.12 *Plot of Tukey HSD test results of "w/ Video–Pt–On-Talk" WER ANOVA output. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)* 187
10.13 *Means plot of "w/ Video–On-Talk" perceived emotion utterance WER. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)* 190

10.14 *Means plot of "w/o Video–On-Talk" perceived emotion utterance WER. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)* 190

11.1 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale replies* . 194
11.2 *Stacked bar chart of subject survey Likert scale replies — All Subjects* 195
11.3 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale replies* . 197
11.4 *Stacked barchart of survey Likert scale replies — Setting, role, language, gender, & map* 198
11.5 *Stacked bar chart of subject survey Likert scale replies — Setting* 205
11.6 *Stacked bar chart of subject survey Likert scale replies — Role* 206
11.7 *Stacked bar chart of subject survey Likert scale replies — Language* 207
11.8 *Stacked bar chart of subject survey Likert scale replies — Setting and role* 208
11.9 *Stacked bar chart of subject survey Likert scale replies — Role, setting and language* . 210
11.10 *7 point Likert score result grouping median comparison — w/ Video & w/o Video* . . 212
11.11 *7 point Likert score result grouping median comparison — IF & IG* 212

12.1 *Effect size (Cohen's d) of difference of word count per dialogue act per dataset/role.* . . 215
12.2 *Bar chart of Portuguese subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused), with a stacked bar chart of each grouping's user survey Likert score results* 216
12.3 *Bar chart of English subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused), with a stacked bar chart of each grouping's user survey Likert score results* 216

A.1 *Image of audio synchronisation* 227
A.2 *Portion of data that was edited out and portion that remained* 228
A.3 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Different Grouping* 230
A.4 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Gender 01* 231
A.5 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Gender 02* 232
A.6 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Role 01* 233
A.7 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Role 02* 234
A.8 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Spoken Language 01* 235
A.9 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Spoken Language 02* 236
A.10 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues without & with video 01* 237
A.11 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues without & with video 02* 238
A.12 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues w/ Video that crash 01* 239
A.13 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues w/ Video that crash 02* 240
A.14 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 1 to 4* 241
A.15 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 5 to 8* 242

A.16 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 9 to 12* 243
A.17 *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 13 to 15* 244
A.18 *Image of first page of the User Survey — 1 of 5 pages* 255
A.19 *Image of second page of the User Survey — 2 of 5 pages* 256
A.20 *Image of third page of the User Survey — 3 of 5 pages* 257
A.21 *Image of fourth page of the User Survey — 4 of 5 pages* 258
A.22 *Image of fifth page of the User Survey — 5 of 5 pages* 259
A.23 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by video setting — without live video streaming and with live video streaming* . 281
A.24 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by subject role — between Information Giver (IG) and Information Follower (IF)* 282
A.25 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by subject role — between Portuguese speaking and English speaking subjects* . 283
A.26 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by Gender — male (♂) or female (♀)* 284
A.27 *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by used map — map 01 or map 07* 285
A.28 *Stacked bar chart of subject survey Likert scale results differentiated by the setting and role of the subject — "w/o Video – IG", "w/o Video – IF", "w/ Video – IG" and "w/ Video – IF"* 286
A.29 *Density plot of subject survey Likert scale results differentiated by the setting and role of the subject — "w/o Video – IG", "w/o Video – IF", "w/ Video – IG" and "w/ Video – IF"* 287

B.1 *The 4 route patterns of the HCRC Edinburgh Map Task corpus maps* 298
B.2 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Grouping 01* 309
B.3 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Grouping 02* 310
B.4 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Role 01* . . . 311
B.5 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Role 02* . . . 312
B.6 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Gender 01* . 313
B.7 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Gender 02* . 314
B.8 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Eye Contact 01* 315
B.9 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Eye Contact 02* 316
B.10 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Map 01* . . . 317
B.11 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Map 02* . . . 318

B.12 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Familiarity 01* 319
B.13 *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Familiarity 02* 320
B.14 *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 1 to 4* . . . 321
B.15 *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 5 to 8* . . . 322
B.16 *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 9 to 12* . . 323
B.17 *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 13 to 16* . 324

C.1 *Bar Chart Role Ratio Comparison of Dialogue Acts within HCRC Edinburgh Map Task Corpus (Top and Middle) and Boxplot Comparison of said corpus Dialogue Act Word Count (Bottom).* 331
C.2 *Bar Chart Role Ratio Comparison of Dialogue Acts within ILMT-s2s Corpus (Top and Middle) and Boxplot Comparison of said corpus Dialogue Act Word Count (Bottom).* . . 332
C.3 *Frequency of Occurrence Differences Between the Role within the HCRC Map Task corpus: Grey highlighted sections indicate 90% coverage.* 333
C.4 *Frequency of Occurrence Differences Between the Role within the ILMT-s2s corpus: Grey highlighted sections indicate 90% coverage.* 334
C.5 *Bar Chart Comparison of IG Dialogue Act Percentage of Each Corpus* 335
C.6 *Word Count Boxplot of IG Dialogue Acts of Each Corpus* 335
C.7 *Bar Chart Comparison of IF Dialogue Act Percentage of Each Corpus* 336
C.8 *Word Count Boxplot of IF Dialogue Acts of Each Corpus* 336

List of Tables

3.1 *Summary of HCRC Edinburgh Map Task corpus dialogue duration by maps* 28
3.2 *Summary of HCRC Edinburgh Map Task corpus dialogue used* 29
3.3 *Ranking of HCRC Edinburgh Map Task corpus dialogues by maps* 29
3.4 *HCRC Edinburgh Map Task corpus data summary — Dialogues* 31
3.5 *Summary of dialogue acts annotated in the HCRC Edinburgh Map Task corpus* 32
3.6 *Summary of subject utterance duration divided by TTS output duration — "*$1-(S/T)$*"* 33
3.7 *Summary of the ILMT-s2s corpus duration and utterance count* 35
3.8 *ILMT-s2s corpus data transcription summary* 35
3.9 *Summary of dialogue acts annotated in the ILMT-s2s corpus* 36
3.10 *Summary of perceived emotions annotated in the ILMT-s2s corpus* 37
3.11 *Summary of ILMT-s2s corpus Talk Types* 38
3.12 *Summary of subject utterance duration divided by TTS output duration — "*$1-(S/T)$*"* 38
3.13 *Summary of computer-mediated dialogue WER & Reverse F-Score values* 39
3.14 *List of survey statements about the ILMT-s2s System* 40
3.15 *Median value per statement of subject survey Likert results* 40

4.1 *Recording devices per subject* 61
4.2 *List of HCRC Edinburgh Map Task corpus map Information Giver landmarks that could be defined as Absent/Present* 66
4.3 *Ranking of HCRC Edinburgh Map Task corpus dialogues by maps* 70
4.4 *List of survey statements about the ILMT-s2s System* 72
4.5 *List of survey questions about the ILMT-s2s System* 74
4.6 *Text reply count of ILMT-s2s corpus user survey* 75
4.7 *ILMT-s2s corpus data statistics* 78
4.8 *Synchronised files of the ILMT-s2s corpus* 79
4.9 *ILMT-s2s corpus data transcription summary* 81
4.10 *Description of perceived emotions annotation labels* 82
4.11 *Summary of perceived emotions annotated in the ILMT-s2s corpus* 83
4.12 *Description of dialogue acts annotation labels* 85

4.13 *Summary of dialogue acts annotated in the ILMT-s2s corpus* 87
4.14 *Description of facial expressions and head movements annotation labels* 89
4.15 *Summary of facial expressions and head movements annotated in the ILMT-s2s corpus* 90
4.16 *Description of the talk type labels* 90
4.17 *ILMT-s2s System XML log offset time* 92
4.18 *Summary of ILMT-s2s corpus Talk Types* 94
4.19 *Summary of subject utterance duration divided by TTS output duration — "$1-(S/T)$"* 96
4.20 *Summary of computer-mediated dialogue WER values* 99
4.21 *Summary of computer-mediated dialogue Reverse F-Score values* 99
4.22 Sample window alteration example 100
4.23 *Summary of 20 frame window dialogue data* 101

5.1 *Summary of word count, duration and wpm* 104
5.2 *Summary of wpm per utterance* 105
5.3 *Summary of single and two word subject utterance duration and wpm sample* 106
5.4 *Summary of duration and wpm of the subject utterances output by the TTS system* . . 109
5.5 *Summary of single and two word TTS output duration and wpm sample* 111
5.6 *Summary of subject utterance duration divided by TTS output duration — "S/T"* . . . 112
5.7 *Summary of subject utterance duration divided by TTS output duration — "$1-(S/T)$"* 113
5.8 *Comparison with TTS duration sample* 114

6.1 *Summary and standard deviation of the 2 talk types (All subjects combined)* 120
6.2 *The 13 subjects with significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (all utterances)* 121
6.3 *The 17 subjects without significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (all utterances)* 121
6.4 *The 21 subjects with significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)* 122
6.5 *The 9 subjects without significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)* 122
6.6 *The 21 subjects with significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)* 123
6.7 *The 9 subjects without significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)* 123
6.8 *Summary and standard deviation of 2 talk types of single word, and multi word utterance speech rates of all thirty subjects* 124

6.9 *Summary and standard deviation of 3 talk types of single word, and multi word utterance speech rates of all thirty subjects* 124
6.10 *The 14 subjects with significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (w/o 1 word utterances)* 125
6.11 *The 16 subjects without significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (w/o 1 word utterances)* 126
6.12 *The 17 subjects with significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)* 126
6.13 *The 13 subjects without significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)* 127
6.14 *The 17 subjects with significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)* 127
6.15 *The 13 subjects without significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)* 128
6.16 *ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance by groupings)* 128
6.17 *Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance by groupings)* 129

7.1 *Sample window alteration example* 134
7.2 *Summary of the various speech rates* 139
7.3 *Summary of the speech rates with a significant difference.* 139
7.4 *Summary of the speech rates in the non-mediated dialogue data.* 140
7.5 *Significance test and effect size between datasets (ILMT-s2s corpus values raised by 41.19%)* 141
7.6 *Summary and standard deviation of 4 and 5 word utterances (wpm)* 146

8.1 *Summary of the whole of the ILMT-s2s corpus and data from Map 01 and Map 07 of the HCRC Edinburgh Map Task corpus (Off-Talk dialogue acts of the ILMT-s2s corpus have not been included in this summary)* 150
8.2 *Summary after standardising the dialogue acts of both datasets and after the Portuguese speaking subjects were removed from the ILMT-s2s corpus* 150
8.3 *Summary of dialogue act word count by dataset and role.* 150

9.1 *Sample window alteration example* . . . 162
9.2 *Summary of computer-mediated dialogue F-Score values* . . . 163
9.3 *Summary of computer-mediated dialogue WER values* . . . 164
9.4 *Summary of 20 frame window dialogue data* . . . 164
9.5 *ILMT-s2s corpus On-Talk utterance count summary* . . . 165
9.6 *Speech Rate Adaptation to the ILMT-s2s System by Group (20 frame window)* . . . 166
9.7 *Speech rate significance test and effect size between between groupings (20 frame window)* 166
9.8 *ASR performance significance test and effect size between between groupings (20 frame window)* . . . 167
9.9 *Proportion of system crashes within the ILMT-s2s corpus* . . . 170
9.10 *Comparison of speech rate adaptation to the ILMT-s2s System by groups before and after dialogues with system crashes are removed (20 frame window)* . . . 171
9.11 *Significance test and effect size between groupings (20 frame window: Without system crash dialogues)* . . . 172

10.1 *Summary and standard deviation of dialogue duration — by setting* . . . 181
10.2 *Summary of perceived emotions annotation linkage in the ILMT-s2s corpus.* . . . 182
10.3 *Subject's perceived emotions by setting — WER* . . . 186
10.4 *Subject's perceived emotions by setting of w/ Video — WER* . . . 186
10.5 *English subjects' perceived emotions WER* . . . 188
10.6 *Portuguese subjects' perceived emotions WER* . . . 188
10.7 *Subject's perceived emotions WER — w/ Video setting* . . . 189

11.1 *Participant's requirements expressed* prior *to teleconferencing experience and revised requirement for video* after *8 months experience.* . . . 192
11.2 *Median value per statement of subject survey Likert results* . . . 195
11.3 *Median, mean and standard deviation value per statement of subject survey Likert results* 195
11.4 *Median value per statement of subject survey Likert results* . . . 196
11.5 *Median, mean and standard deviation value per statement of subject survey Likert results* 196
11.6 *Median, mean and standard deviation value per statement of subject survey Likert results* 199
11.7 *Significance test and effect size between Likert scale results of "w/o Video : w/ Video"* 199
11.8 *Median, mean and standard deviation value per statement of subject survey Likert results* 200
11.9 *Significance test and effect size between Likert scale results of "Information Giver (IG) : Information Follower (IF)"* . . . 200
11.10 *Median, mean and standard deviation value per statement of subject survey Likert results* 200
11.11 *Significance test and effect size between Likert scale results of "Portuguese speaking subject (Pt) : English speaking subject (En)"* . . . 200
11.12 *Median and mean value per statement of subject survey Likert results* . . . 201

11.13 *Median and mean value per statement of subject survey Likert results* 201
11.14 *Median, mean and standard deviation value per statement of subject survey Likert results* 201
11.15 *Significance test and effect size between Likert scale results of "w/ Video – IF : w/o Video – IG" — Part 1 of 6* 203
11.16 *Significance test and effect size between Likert scale results of "w/ Video : w/o Video of IF" — Part 2 of 6* 203
11.17 *Significance test and effect size between Likert scale results of "IF : IG of w/ Video" — Part 3 of 6* 203
11.18 *Significance test and effect size between Likert scale results of "w/o Video : w/ Video of IG" — Part 4 of 6* 204
11.19 *Significance test and effect size between Likert scale results of "IG : IF of w/o Video" — Part 5 of 6* 204
11.20 *Significance test and effect size between Likert scale results of "w/o Video – IF : w/ Video – IG" — Part 6 of 6* 204
11.21 *Median, mean and standard deviation value per statement of subject survey Likert results* 205
11.22 *Median, mean and standard deviation value per statement of subject survey Likert results* 206
11.23 *Median, mean and standard deviation value per statement of subject survey Likert results* 207
11.24 *Median, mean and standard deviation value per statement of subject survey Likert results* 208

A.1 *ILMT-s2s corpus data summary — Media Files* 224
A.2 *ILMT-s2s corpus data summary — Dialogues* 225
A.3 *HCRC Edinburgh Map Task corpus data summary — Dialogues* 225
A.4 *Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (all utterances)* 246
A.5 *ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)* 247
A.6 *Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)* 248
A.7 *Summary and standard deviation of 2 talk type all word, single word, and multi word utterance speech rates of all thirty subjects.* 249
A.8 *Summary and standard deviation of 3 talk type all word, single word, and multi word utterance speech rates of all thirty subjects.* 249
A.9 *Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (only 1 word utterances)* 250
A.10 *Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (w/o 1 word utterances)* 251

A.11 *ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)* 252
A.12 *Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)* 253
A.13 *List of survey statements and questions about the ILMT-s2s System* 254
A.14 *ILMT-s2s user Likert survey results* 260
A.15 *Median value of the thirty Subjects' Likert scale results* 278
A.16 *Mean value of the thirty Subjects' Likert scale results* 279
A.17 *Significance test and effect size between Likert scale results of "w/o Video : w/ Video"* 290
A.18 *Significance test and effect size between Likert scale results of "Information Giver (IG) : Information Follower (IF)"* 290
A.19 *Significance test and effect size between Likert scale results of "Portuguese speaking subject (Pt) : English speaking subject (En)"* 291
A.20 *Significance test and effect size between Likert scale results of "Male : Female"* 291
A.21 *Significance test and effect size between Likert scale results of "map 01 : map 07"* . . 292
A.22 *Significance test and effect size between Likert scale results of "w/ Video – Information Follower (IF) : w/o Video – Information Giver (IG)" — Part 1 of 6* 292
A.23 *Significance test and effect size between Likert scale results of "w/ Video – Information Follower (IF) : w/o Video – Information Follower (IF)" — Part 2 of 6* 293
A.24 *Significance test and effect size between Likert scale results of "w/ Video – Information Follower (IF) : w/ Video – Information Giver (IG)" — Part 3 of 6* 293
A.25 *Significance test and effect size between Likert scale results of "w/o Video – Information Giver (IG) : w/ Video – Information Giver (IG)" — Part 4 of 6* 294
A.26 *Significance test and effect size between Likert scale results of "w/o Video – Information Giver (IG) : w/o Video – Information Follower (IF)" — Part 5 of 6* 294
A.27 *Significance test and effect size between Likert scale results of "w/o Video – Information Follower (IF) : w/ Video – Information Giver (IG)" — Part 6 of 6* 295
A.28 *Significance test and effect size between Likert results from subjects w/o live video streaming and w/ live video streaming* 295
A.29 *Significance test and effect size between Likert results from subjects with the role of Information Giver (IG) and Information Follower (IF)* 296
A.30 *Significance test and effect size between Likert results from English speaking and Portuguese speaking subjects* 296
A.31 *Significance test and effect size between Likert results from male and female subjects* . 297
A.32 *Significance test and effect size between Likert results from subjects using map 01 and subjects using map 07* 297

B.1 *HCRC Edinburgh Map Task corpus Map difference between role — Route type 01* . . . 299
B.2 *HCRC Edinburgh Map Task corpus Map difference between role — Route type 02* . . . 300
B.3 *HCRC Edinburgh Map Task corpus Map difference between role — Route type 03* . . . 301
B.4 *HCRC Edinburgh Map Task corpus Map difference between role — Route type 04* . . . 302
B.5 *20 Window Values of Groupings within the HCRC Edinburgh Map Task corpus* 304
B.6 *20 Window Values of Dialogues 01 to 04 of the HCRC Edinburgh Map Task corpus* . . 305
B.7 *20 Window Values of Dialogues 05 to 08 of the HCRC Edinburgh Map Task corpus* . . 306
B.8 *20 Window Values of Dialogues 09 to 12 of the HCRC Edinburgh Map Task corpus* . . 307
B.9 *20 Window Values of Dialogues 13 to 16 of the HCRC Edinburgh Map Task corpus* . . 308

C.1 *Summary and standard deviation of Dialogue Acts Word Count by Role within the HCRC Edinburgh Map Task corpus.* 327
C.2 *Summary and standard deviation of Dialogue Acts Word Count by Role within the ILMT-s2s corpus.* 328
C.3 *Dialogue act occurrence count per corpora separated by subject role* 329
C.4 *Dialogue act word count per corpora separated by subject role* 330

List of Abbreviations

.	P-value range: $0.1 \geq P \geq 0.05$
*	P-value range: $0.05 \geq P \geq 0.01$
**	P-value range: $0.01 \geq P \geq 0.001$
***	P-value range: $0.001 \geq P$
δ	Delta (Cliff's delta results)
♀	Female
♂	Male
1080i	1080 Interlace video
1080p	1080 Progressive video
ANOVA	Analysis of variance
API	Application Programming Interface
ASR	Automatic Speech Recognition
BR	Brazil
BVP	Blood Volume Pulse
CP	Continuation/Contact and Perception
CPU	Continuation/Contact, Perception and Understanding
EC	Eye Contact
EEG	Electroencephelography
ELAN	The EUDICO Linguistic Annotator (Computer software)
En	English
Eng	English
eps	Encapsulated PostScript (Vector image format)
ES	Effect Size
FLAC	Free Lossless Audio Codec (Audio coding format)
FPS	Frames Per Second
H2C	Human-to-Computer
H2H	Human-to-Human
HCRC	Human Communication Research Centre
HR	Heart Rate
IF	Information Follower
IG	Information Giver

m	Minute
M	Means
Mdn	Median
MT	Machine Translation
OOT	Other forms of Off-Talk
PCM	Pulse-code modulation
Por	Portuguese
PSSUQ	The Post-Study System Usability Questionnaire
Pt	Portuguese
PT	Portugal
ROT	Read Off-Talk
s	Second
S.01–S.15	Statement 1 – Statement 15 of user survey
S2S	Speech-to-Speech
S2S-MT	Speech-to-Speech Machine Translation
S2T	Speech-to-Text
S2T-MT	Speech-to-Text Machine Translation
SC	Skin Conductance
SD	Standard Deviation
sps	Samples Per Second
svg	Scalable Vector Graphics
T2T-MT	Text-to-Text Machine Translation
TCD	Trinity College Dublin
TTS	Text-to-Speech
UI	User Interface
utt	Utterance
w/	with
w/o	without
WER	Word Error Rate
WPM	Words Per Minute
XML	Extensible Markup Language
C-STAR	Consortium for Speech Translation Advanced Research
TC-STAR	Technology and Corpora for Speech to Speech Translation
ATR	Advanced Telecommunications Research Institute
BBN	BBN Technologies (Bolt, Beranek and Newman)
EMIME	Effective Multilingual Interaction in Mobile Environments
VoiceTra	Voice Translator
ILMT-s2s	Inter Lingual Machine Translation Speech-to-Speech (System/corpus) or Inter Lingual Map Task Speech-to-Speech (System/corpus)

Chapter 1

Introduction

This thesis aims to provide design cues for future Speech-to-Speech Machine Translation (S2S-MT) systems by providing insights on how a user adapts to such a system. It investigates the factors that might influence the user experience and can also be captured by the S2S-MT system in action. The factors explored are user's preference of modality, speech rate, effect of ASR performance, and language usage, that are observed in the light of sociological factors such as gender, task role, native language and possibility to have eye-contact. This exploration is split into different studies presented in this thesis, that aim, as a whole to, understand how the users adapt their communication style and interaction, when talking to another person who speaks a different language through a computer mediated Speech-to-Speech Machine Translation (S2S-MT) system. This chapter introduces the subject of this thesis by presenting the motivations, details the research question, outlines the main contributions, provides the associated publications in which parts of this work have been published and gives the structure of the thesis.

1.1 Motivations

As a way of communication, S2S-MT systems are becoming a reality. Microsoft has released the Skype Translator (W. D. Lewis, 2015), and the Japanese Ministry of Internal Affairs and Communication has announced that they target S2S-MT to have the capabilities of being used in detailed business negotiations by 2030, with testing of such systems already started using the VoiceTra system (Matsuda et al., 2013) in hospitals, hotels, tourist sites and shopping centres.[1] Although automatic S2S-MT systems are already in commercial deployment they are still not adapted to the way in which people actually behave when using them, specifically how dialogues between users change between Human-to-Human (H2H) communication and computer mediated interlingual communication. People are already experienced with H2H communication and some are even familiar with communicating in a

[1] Source: Ministry of Internal Affairs and Communications website (https://www.soumu.go.jp/main_content/000678485.pdf) (accessed on 2021/09/30).

foreign language, which are subjects that have been abundantly explored by researchers. However, few are familiar with communication via an interpreter, let alone an interpreter that cannot be interrupted — which is what S2S-MT systems are offering (Figure 1.1). This important technological progress that S2S-MT systems bring to everyday life, being able to converse directly with another person that is not speaking your language and without the need of a human interpreter, offers a vast range of possibilities for people to understand each other by virtually by-passing the language barrier. But this also raises a number of questions. Some questions are concerned with the machine translation itself, that remains open to interpretation and improvement, but others concern how this technology will be accepted and integrated into people's lives. How a user adapt to this type of technology is still in great need for answers.

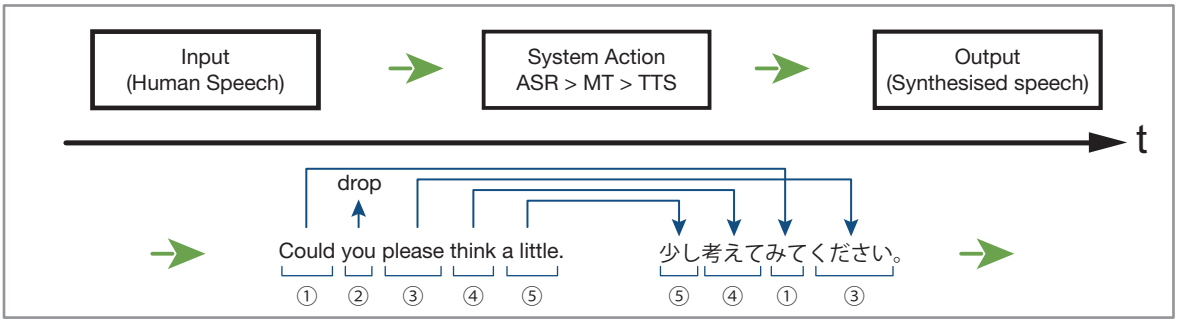


Figure 1.1: *Communication flow of S2S-MT system — English to Japanese example*

To fill this gap, I conducted a series of exploratory studies of language-technology mediated task oriented dialogues, for which I created a dataset using an interlingual communication system ("The Communication Tool" in §4.2) to collect a detailed log of all interactions on fifteen task based dialogues. The aim of this thesis is to look into how people adapt their communication style and their interaction when talking to their interlocutor who speaks a different language though the ILMT-s2s System for a task oriented dialogue.

1.2 The Question

There has been a lot of research into Human-to-Human (H2H) communication and Human-to-Computer (H2C) communication (Branigan, Pickering, Pearson, & McLean, 2010) and Machine Translation (MT) mediated communications as well (Yamashita & Ishida, 2006b). However, as mentioned by Hara & Iqbal (2015), little has been published in S2S-MT mediated communication. The Karlsruhe Institute of Technology in Germany is one step closer and uses an Speech-to-Text Machine Translation (S2T-MT) system to translate lectures from German into English (Müller, Fünfer, Stüker, & Waibel, 2016) and following the evaluation of the system from student user feedback, they are now trying to reduce the latency caused due to the characteristics of interpreting (Niehues et al., 2016) (since the financial gains from such a system are high, little is presented within the article).

Stepping aside from the trends in academia, it is important to look at a simple idea. As Picard

(1995) and Norman (2002) mention, it is important that the design of a system does not frustrate the user. So as not to frustrate the user it is important to understand how the user will use and adapt to such a system, and this user adaptation is what I will base my research question upon; the observation and exploration of how users adapt their communication style and interaction when using an S2S-MT system to mediate their communication. I focus on one type of dialogue particularly fit to observe interactional adaptation: task based dialogues, and map tasks in particular. A map task is a exercise were an information giver guides and information follower through a map with landmarks (the information giver having the map with the route drawn on it while the follower do not, more details are given in the following chapters 3 and 4). The next section (§ 1.2.1) details how map task dialogues are appropriate as an object of study for this thesis.

1.2.1 Task Oriented Interactions

From the spectrum of unconstrained dialogue recordings on one hand and speech readouts on the other, this thesis investigates task oriented dialogues that have minimal constraints, yet have a comparable face-to-face H2H dataset available. Of the six primary dialogue types (Walton & Krabbe, 1995) of "Information-seeking dialogues", "Inquires", "Persuasion dialogues", "Negotiations", "Deliberations" and "Eristic", by using the HCRC Edinburgh Map Task (Anderson et al., 1991) maps to elicit conversation, this study mainly investigates dialogues that are structured around *information-seeking*. Although a number of studies of linguistic phenomena and adaptation strategies in map task dialogues have been carried out (Reitter & Moore, 2007; Newlands, Anderson, & Mullin, 2003) to the best of my knowledge, little research has been done to investigate communicative behaviour in map task dialogues in a computer mediated interlingual (S2S-MT) setting.

Spoken dialogues can be seen as a joint action in which interlocutors coordinate their verbal and non-verbal behaviour and adapt their linguistic choices to each other. This often results in the phenomenon referred to as convergence or alignment which consists in a tendency by the interlocutors to adopt and re-use each other's language patterns in the course of authentic interaction (Giles, Mulac, Bradac, & Johnson, 1987). According to the interactive alignment model (Pickering & Garrod, 2004) this linguistic coordination in dialogue occurs at the level of the lexicon, grammar, and pronunciation; and represents one way in which interlocutors achieve understanding in dialogue.

Previous studies have argued that convergence is a property of human dialogue that seems to persist even when one of the interlocutors is replaced by a conversational H2C interface. Reeves & Nass (1996, p. 28) have even gone so far as to assume that "It is possible to take a psychology research paper about how people respond to other people, replace the word 'human' with the word 'computer,' and get the same result". Subjects have a tendency to hyperarticulate with slower and louder speech (Oviatt, Bernard, & Levow, 1998) as if talking to a baby or a person hard of hearing, align their choice of words with the output of the S2S-MT system (Schneider & Luz, 2011) and align

their speech rate to a digital agent (L. Bell, Gustafson, & Heldner, 2003) as if they were talking with a fellow human, and they will even show prosodic alignment to the digital agent (Suzuki & Katagiri, 2007). The exploratory studies presented in this thesis investigates if this findings applies to S2S-MT mediated task oriented H2H communication as well.

1.2.2 Research Question

As mentioned in the previous section, the aim of this thesis is to observe and explore how subjects adapt their communication style and interaction, when using an S2S-MT system installed on each subject's computer to mediate interlingual communication over the internet in a task oriented exercise. The three main questions to be answered about the computer-mediated dialogue data (the ILMT-s2s corpus) collection are, (i) how are the subjects adapting to the interlingual computer mediated communication style, (ii) how does the communication style compare with the direct Human-to-Human communication style of the non-mediated dialogue data (the HCRC Edinburgh Map Task corpus), and (iii) why would the subjects of the computer-mediated dialogue data collection be adapting, if, they are at all adapting to the interlingual computer mediated interaction. To answer these questions, I have studied the following aspects of computer mediated interlingual communication.

- Preference of Modality

  Do users have a preference in communication modality (with or without eye-contact)?

- Difference in Speech Rate

  Does the subject adapt their speech rate to the ILMT-s2s System?

- Effect of Automatic Speech Recognition (ASR) Errors

  How does the subject react to the errors of the ILMT-s2s System?

- Language Usage Difference

  How does the communication style of the computer-mediated dialogue data compare to that of the non-mediated dialogue data?

This work provides data fit for the exploration of these questions and present the reasoning associated with observations made along the data creation. These specific task-oriented computer-mediated interlingual communication setups are meant to be used as insight to open or strengthen new research avenues in a more controlled experimental setting.

1.3 Contributions and Key Findings

The three main questions to be answered about the computer-mediated dialogue data are, (i) how are the subjects adapting to the interlingual computer mediated communication style, (ii) how does the communication style compare with the direct Human-to-Human communication style of the non–mediated dialogue dataset, and (iii) why would the subjects of the computer-mediated dialogue data collection be adapting, if, they are at all adapting to the interlingual computer mediated interaction. Detailed explanations to the answer of these questions can be found in Chapter 5 to Chapter 11 of this thesis. This section gives a simple summary:

(i) How are the subjects adapting to the interlingual computer mediated communication style?

From the data of the computer-mediated dialogue dataset, it has been identified that the subjects used three different basic speech rates. One while talking to the system that was the slowest, a second when talking to themselves, and the third when talking directly a fellow human which was the fastest. When the subjects of the computer-mediated dialogue dataset talked to their interlocutor through the S2S-MT system (the ILMT-s2s System) ASR performance did affect the speech rate, but more prominent was the gender difference with female subjects speaking slower than male subjects and the difference in the role of the subject with the Information Giver talking slower than the Information Follower.

(ii) How does the communication style compare with the direct Human-to-Human communication style of the non-mediated dialogue data?

The comparison of the interlingual computer mediated communication of the computer-mediated dialogue dataset and the direct Human-to-Human communication of the non-mediated dialogue dataset indicated that when the subjects in both corpora were talking directly to a fellow human, their speech rate was very similar. Also the fact that the Information Giver spoke slower than the Information Follower was another similarity between the datasets, but no speech rate difference was identified between the gender of the subjects in the non-mediated dialogue dataset, indicating that the identification of female subjects speaking slower than the male subjects might be an effect of the computer mediated communication. Furthermore, the dialogue acts of the two datasets showed clear differences. The communication style used for the interlingual computer-mediated dialogues was closer to high latency video-conferencing communication with a reduction in communicated backchannels, shorter utterances for the Information Giver and longer utterances from the Information Follower, indicating that the cost (required effort)[2] of

[2]To quote Newlands et al. (2003, p. 344), "The use of precise instruction giving in the CMC context also seems to be a sensible way of adapting to the restrictive nature of this text-based communicative context. It may take less collaborative effort to ground precise instructions in this written context than the rather underspecified messages often found in spoken interactions; this would reduce the number of turns (or cycles of presentation and acceptance phases) required to establish mutual understanding. In contrast, in spoken interactions the use of highly specified instructions might be counter-productive, as they would require several repetitions before the information could be assimilated correctly. Very precise oral instructions could also be deemed to violate Grice's Cooperative Principle and the associated conversational maxims of quantity, quality, relation and manner (Grice, 1975)."

communication had risen and used with more consideration.

(iii) Why would the subjects of the computer-mediated dialogue data collection be adapting, if, they are at all adapting to the interlingual computer mediated interaction?

The subjects of the computer-mediated dialogue dataset did adapt to the ILMT-s2s System. From the comparison with the non-mediated dialogue dataset it is clear that there is a difference between the two datasets, but there is no definitive answer to the question of why. However there are indications that the ASR performance is not the only reason why the subjects spoke more slowly to the ILMT-s2s System. Of the subjects who had a perfect ASR result for their first utterances, the onset of reduced speech rate occurred before the first ASR error, so both the difficulty of the task on hand (the map task) and the ASR performance can be attributed as the cause of slower speech. From the subject survey results, there was a clear indication that the subjects using the w/ Video setting, subjects that were able to see their interlocutor, and had the role of Information Follower, following the instructions, didn't perceive the ILMT-s2s System as highly as subjects using the w/o Video setting, subjects who couldn't see their interlocutor, or had the role of Information Giver, providing the instructions. However, for the different role of the subject didn't provide a significant difference in the ASR performance, and even though there was a significant difference for the setting, the effect size was small. It is not possible to give a definitive answer to the "why", yet, indications to future system developers can be given, with the following approaches. The fact that the female subjects used a slower speech rate than the male subjects: an effort that did not reward the female subjects with a better ASR performance, hence a notice that speaking slowly will not help should be given at the start of system usage. The fact that the subjects clearly prefer the communication modality where they cannot see the interlocutor: manage the expectations of the users so that they do not use a register too closely related to direct Human-to-Human communication style when they can see the interlocutor.

1.4 Associated Publications

1.4.1 Publications that Used the Data Collected in this Study

The publications listed below are studies that used the computer-mediated dialogue data as part of the dataset of their studies. My contribution to these publications with regards to the analysis that is performed is the bare minimal,[3] and my inclusion as a co-author is due to the grace of the other authors more than any direct contribution that I made.

(i) **Reverdy, J.**, Hayakawa, A. & Vogel, C. (2018).
Alignment in a Multimodal Interlingual Computer-Mediated Map Task Corpus. *In Proceedings of the Multimodal Corpora Workshop* 2018 at LREC'18. Miyazaki, Japan: European Language Resources Association (ELRA).

(ii) **Haider, F.**, Hayakawa, A., Luz, S., Vogel, C. & Campbell, N. (2018).
On-Talk and Off-Talk Detection: A Discrete Wavelet Transform Analysis of Electroencephalogram. *In Proceedings of ICASSP 2018: the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 960 – 964). Calgary, Alberta, Canada: IEEE.

(iii) **Haider, F.**, Hayakawa, A., Vogel, C., Campbell, N. & Luz, S. (2019).
Analysing patterns of right brain-hemisphere activity prior to speech articulation for identification of system-directed speech. *Speech Communication* (Vol. 107, pp. 18 – 25). Elsevier.

1.4.2 Publications and Presentations Related to this Study

The work in this thesis builds upon the following publications that I had the privilege to publish. To the best of my knowledge, analysis and results obtained by the co-authors are not included in this thesis, however ideas are mature during the process of questioning each other over the contents of the paper, and it is thanks to them that I was able to write this thesis.[4]

- Some feedback received from these presentations are noted bellow the presentation title, but they may not have been reviewed within this thesis.

(i) **Hayakawa, A.**, Campbell, N., & Luz, S. (2014).
Interlingual Map Task Corpus Collection. *The Scottish Information & Computer Science Alliance: Speech Technology and Human Computer Interaction Workshop.* Edinburgh, United Kingdom.

- At this workshop, I had the privilege to talk to Dr. Holly Branigan and explain the idea behind collecting data for this study. The point she was most interested about was, unfortunately,

[3] The description of the data is correct and the analysis/methods, results and conclusion have been scrutinised to the best of my knowledge and represent the data in a true and honest manner.

[4] Permission has been obtained from the co-authors to incorporate the text of the works within this thesis.

an aspect that was dropped from the actual data collection due to time restrictions of the subjects — how the subject survey would be conducted. Dr. Branigan agreed with the idea that the subjects should be questioned in-person while going over their actions in the video recordings of the task — this is to minimise the difference between what the subjects' thought they did and what they actually did.

- Contents related to this publication can be found in chapter 4.

(ii) **Hayakawa, A.**, Campbell, N., & Luz, S. (2014).

Interlingual Map Task Corpus Collection. *In Proceedings of INTERSPEECH'14: the 15th Annual Conference of the International Speech Communication Association* (pp. 189 – 191). Singapore: ISCA.

- This "Show & Tell" demonstration was the first presentation of a running ILMT-s2s System to the scientific community. During the session, Dr. Gérard Bailly told me that eye tracking data might be interesting, and Dr. John Hansen indicated that the study of the human interacting with the system as a whole is of interest. Unfortunately, the eye tracking data is one of the items that I have yet to study within the data, but Dr. Hansen's point came true thanks to the numerous papers that were published using the collected computer-mediated dialogue data.
- Contents related to this publication can be found in chapter 4.

(iii) **Hayakawa, A.**, Cerrato, L., Campbell, N., & Luz, S. (2015).

A Study of Prosodic Alignment In Interlingual Map-Task Dialogues. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of ICPhS XVIII (18th International Congress of Phonetic Sciences).* Glasgow, United Kingdom: The University of Glasgow. ISBN 978-0-85261-941-1. Paper number 760.1-5.

- In this paper, I presented the speech rate patterns alongside the word-error-rate patterns of the computer system that mediated the dialogues within the collected data. The most challenging question from this presentation was "Are the subjects talking to a human or a computer?", to which I replied that the dialogue resembled communication towards a computer rather than a fellow human.
- Contents related to this publication can be found in chapter 9.

(iv) **Hayakawa, A.**, Haider, F., Cerrato, L., Campbell, N., & Luz, S. (2015).

Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems. *In Proceedings of INTERSPEECH'15: the 16th Annual Conference of the International Speech Communication Association* (pp. 2539 – 2543). Dresden, Germany: ISCA.

- In this paper, Dr. Fasih Haider and I presented the detection of the subjects' perceived emotion from the collected biosignal data and the association between the ASR performance and

the perceived emotion. Interest towards this paper were basically only directed towards the biosignal analysis method performed by my co-author Dr. Haider.

- Contents related to the perceived emotions reported in this publication can be found in chapter 10.

(v) Cerrato, L., **Hayakawa, A.**, Campbell, N., & Luz, S. (2016).

A Speech-to-Speech, Machine Translation Mediated Map Task: An Exploratory Study. *Future and Emergent Trends in Language Technology: First International Workshop, FETLT 2015, Seville, Spain, 2015, Revised Selected Papers,* (pp. 53 – 64). Cham: Springer International Publishing.

- My contribution to this paper was providing the data for the qualitative study performed by Dr. Loredana Cerrato, the quantitative analysis of the data and other minor tasks. This paper was successfully presented by the first author at FETLT 2015, to what I have been told was a positive reception.
- Contents related to the quantitive analysis in this publication can be found in chapter 9.

(vi) **Hayakawa, A.**, Luz, S., Cerrato, L., & Campbell, N. (2016).

The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 605 – 612). Portorož, Slovenia: European Language Resources Association (ELRA).

- This paper was published to introduce the completed computer-mediated dialogue data collection to the community. No questions of interest were asked during the presentation.
- Contents related to this publication can be found in chapter 4.

(vii) **Hayakawa, A.**, Haider, F., Luz, S., Cerrato, L., & Campbell, N. (2016).

Talking to a System and Oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. *In Proceedings of Speech Prosody 2016 (SP8)* (pp. 776 – 780). Boston, Massachusetts, USA: ISCA.

- In this paper, the three different speech rates that were identified in the computer-mediated dialogue data were analysed and also detected using the biosignal. Though I am the first author of this paper, I asked my co-author Dr. Fasih Haider to present this paper. As with the INTERSPEECH'15 paper, I have been told that the main questions were around the biosignal analysis which were performed by my co-author Dr. Haider. The study of the biosignals was continued by Dr. Haider and published as indicated in § 1.4.1.
- As a side note, at INTERSPEECH'16, I asked Dr. Anton Batliner about *On-Talk* and *Off-Talk* and was encouraged to continue the research into this topic.
- Contents related to the speech rate reported in this publication can be found in chapter 6.

(viii) **Hayakawa, A.**, Luz, S. & Campbell, N. (2016).

Talking to a System and Talking to a Human: A study from a Speech-to-Speech, Machine Translation mediated Map Task. *In Proceedings of INTERSPEECH'16: the 17th Annual Conference of the International Speech Communication Association* (pp. 1422 – 1426). San Francisco, California, USA: ISCA.

- In this paper, I presented how the dialogue acts differed within the non-mediated dialogue dataset and the computer-mediated dialogue dataset. The conclusion that the cost of communication had increased in the computer-mediated dialogue data was not challenged with the results showing a reduction in short acknowledgement utterances and an increase in utterance length from the Information Follower. No questions of interest were asked during the presentation.
- Contents related to this publication can be found in chapter 8.

(ix) **Hayakawa, A.**, Vogel, C., Luz, S. & Campbell, N. (2017).

Speech Rate Comparison when Talking to a System and Talking to a Human: A study from a Speech-to-Speech, Machine Translation mediated Map Task. *In Proceedings of INTERSPEECH'17: the 18th Annual Conference of the International Speech Communication Association* (pp. 3286 – 3290). Stockholm, Sweden: ISCA.

- In this paper, I presented the observation that parts of the speech rate between the non-mediated dialogue dataset and the computer-mediated dialogue dataset were similar, but also showed a greater difference between gender in the computer-mediated dialogue data. The main interest this paper received was the method used to calculate the speech rate and due to this interest, I decided to write the paper published at LREC'18.
- Since this paper presented results that indicated that there was a different approach between gender, I used caution when presenting this finding that I thought would be good for improving usability in software design. Therefore I was surprised when a researcher associated with one of the top academics in this field came up to me and mentioned that the results do not help gender equality and shouldn't have been published.
- Contents related to this publication can be found in chapter 7.

(x) **Hayakawa, A.**, Vogel, C., Luz, S. & Campbell, N. (2017).

Perception Changes With and Without the Video Channel: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. *In Proceedings of CogInfoCom 2017: the 8th IEEE International Conference on Cognitive InfoCommunications* (pp. 401 – 406). Debrecen, Hungary: IEEE.

- I see this paper as a six page version of this thesis. Fortunately no points of concern were raised except the fact that the title is too long.
- Contents related to this publication can be found in chapter 11.

(xi) **Hayakawa, A.**, Vogel, C., Luz, S. & Campbell, N. (2018).

Speech Rate Calculations with Short Utterances: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)* (pp. 3176 – 3183). Miyazaki, Japan: European Language Resources Association (ELRA).

- Following the interest towards the speech rate difference calculation method at INTERSPEECH'17, this paper was written to describe in detail, how and why I use the speech rate difference instead of the more orthodox words-per-minute or syllables-per-second. During the session, Dr. Hiroya Fujisaki quizzed me and indicated that human speakers, even well trained news readers cannot speak at a single rate without variation or repeat phrases at the same speech rate and didn't question the conclusion of using the computer text-to-speech module as a reference to calculate the speech rate. However, to conclude, Dr. Fujisaki told me to try and find a method to accurately calculate speech rate without TTS references, even if it takes me 30 years to do so, indicating that his *"Fujisaki-model"* was not created in a year. Furthermore, Dr. Fujisaki quizzed me on why single word utterances having a bigger variation in speech rate than multi-word utterances, with his theory being related to the various muscle initiation processes when one starts to speak from a state of non-activity. Though I believe this is one of the reasons, I also believe that lung capacity, may have a factor in this. This idea come from my low capacity that makes it difficult for me to speak loudly or utter long segments of words without pause.
- Contents related to this publication can be found in chapter 5.

(xii) **Hayakawa, A.** (2019).

タスク型会話で自動翻訳システムを使ったユーザーの反応分析と比較 (Analysis of User Reaction to Machine Translated Task Based Dialogue). 第17回音情報処理研究交流会 (The 17$^{\text{th}}$ Researcher's Gathering on Sonic Information Processing). Osaka, Japan: Panasonic Corporation.

- As an invited speaker, I presented a summary of the work presented within this thesis. The main relief from this presentation was my past co-workers who were present reassuring me that finding unpopular results such as the speech rate difference between genders was a finding that would improve products and not something to shy away from.

1.5 Overview of the Research

1.5.1 Introduction

As illustrated in Figure 1.2, three types of analysis were performed on the data of the computer-mediated dialogue dataset. (i) A comparison of the computer-mediated dialogue dataset with data from the non-mediated dialogue dataset to see how the data compares with direct Human-to-Human (H2H) communication, (ii) analysis of the dataset to see what the data says about the subjects and (iii) analysis of the user survey to try to understand what the subjects of the computer-mediated dialogue data collection felt about the ILMT-s2s System. Two datasets were used in this study, a subset of the HCRC Edinburgh Map Task corpus (non-mediated dialogue data) and parts of the ILMT-s2s corpus (computer-mediated dialogue data). The ILMT-s2s corpus was collected for the purpose of this study, but not all of the collected data was used. For this reason, as illustrated in Figure 1.3, the data used in this study are described in Chapter 3 while the details of the corpus creation are described in Chapter 4. Chapter 5 explains in detail how and why the speech rate indications values used in this study were obtained, but it has been removed from Chapter 4 due to it's length.

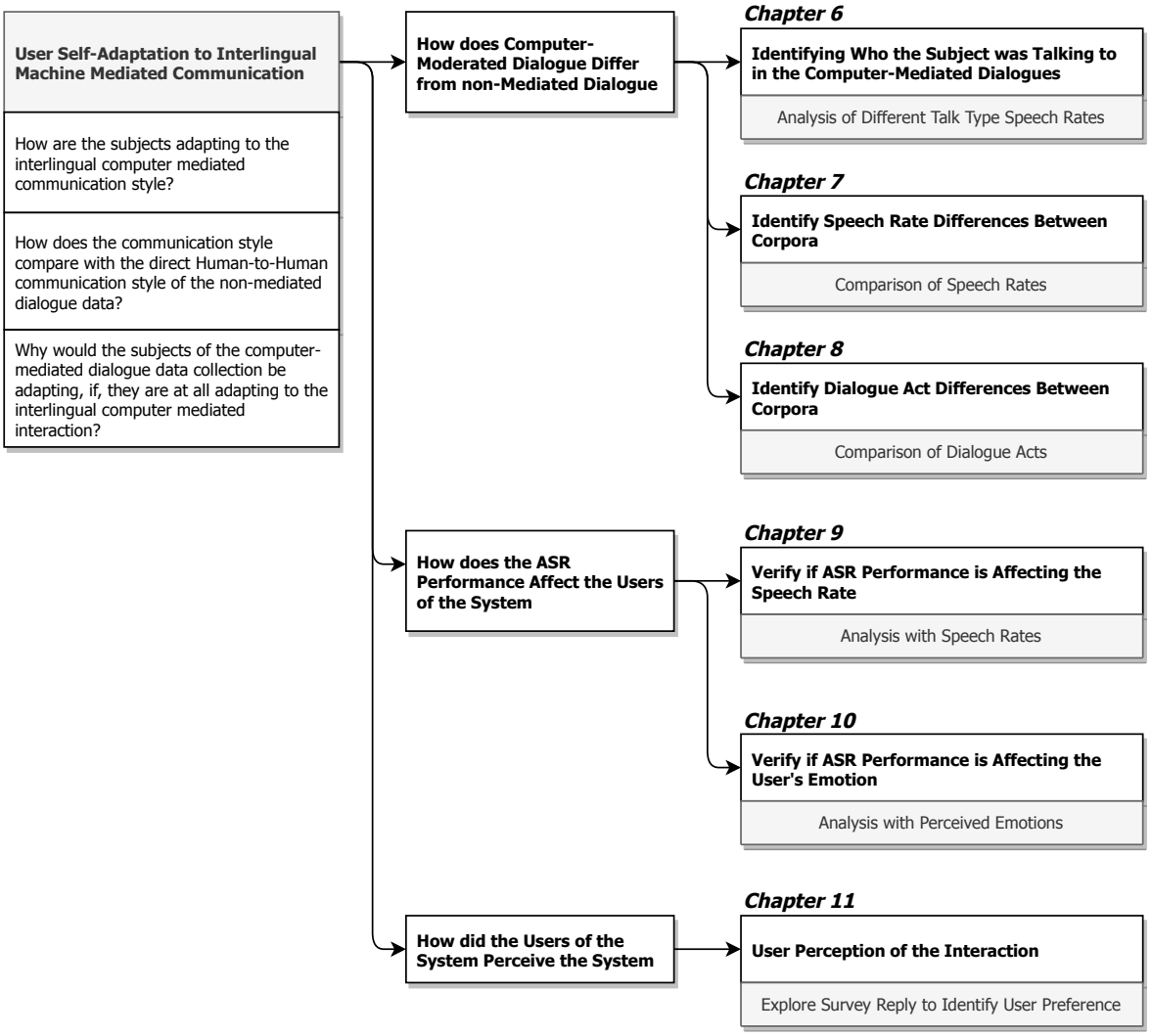


Figure 1.2: *Overview of the Analysis Conducted in this Study*

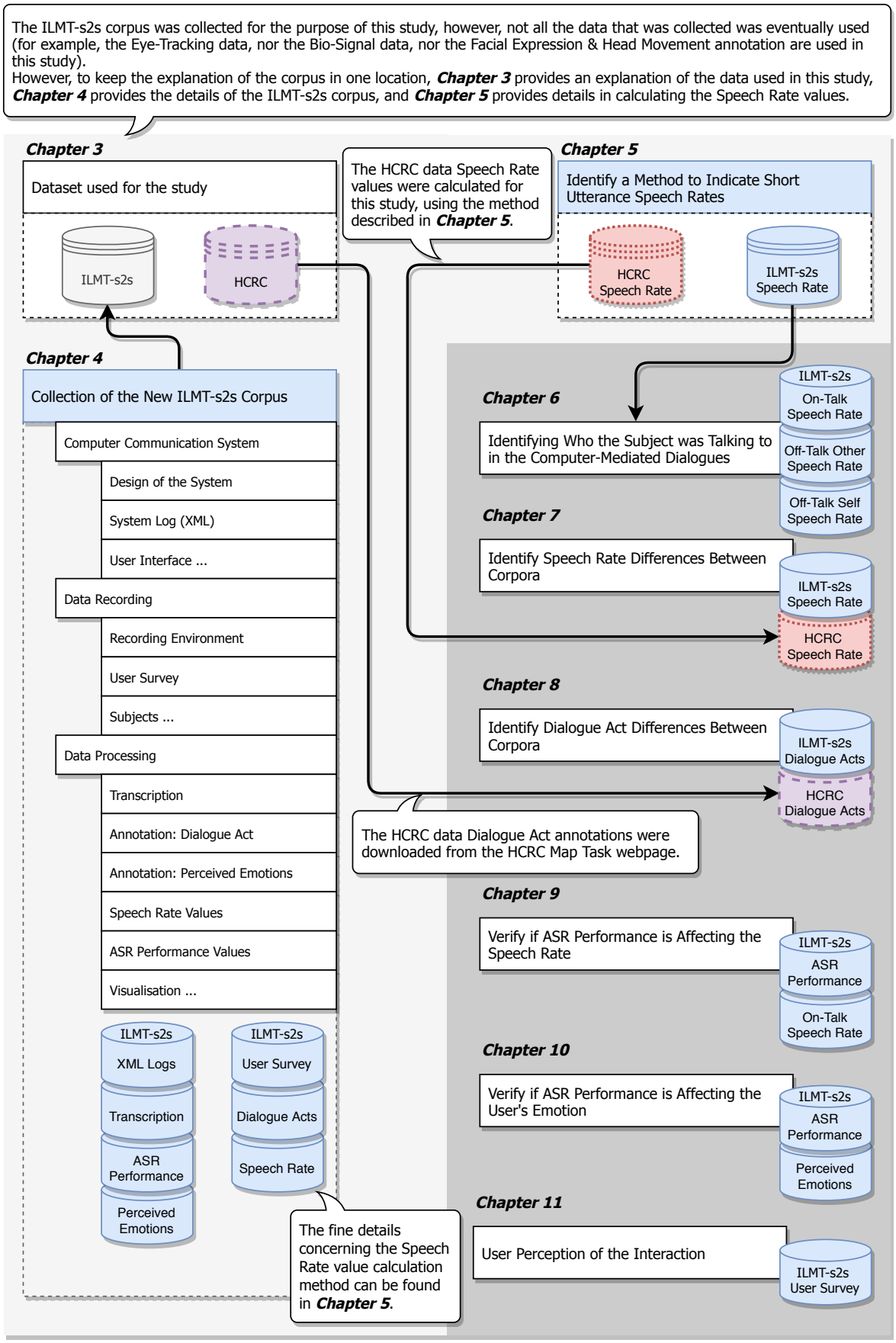


Figure 1.3: *Overview of the Data Used for the Analysis in this Study*

1.5.2 Structure of the Thesis

As mentioned above (§ 1.5.1), quantitive and qualitative analysis were conducted in this research. The quantitive analysis starts in Chapter 6 and continues to Chapter 10 with the qualitative analysis in Chapter 11. The following bullet points provide an introduction to the contents of the main chapters within this thesis.

- Chapter 3: The Datasets

  For the research in this thesis a dataset (the ILMT-s2s corpus, explained in Chapter 4), was collected. However not all of the data that was collected was used in the research of this thesis, and it was not the only dataset that was used — a subset of the HCRC Edinburgh Map Task corpus was used as a reference to compare the computer-mediated dialogue data with non-mediated dialogue data. For this reason, this chapter lists the data actually used in this research.

- Chapter 4: Data Created for this Study

  In this chapter, I explain the details of the data collected for this research. Starting with the initial purpose of the collection, details of how the speech-to-speech machine translation computer system (the ILMT-s2s System) that mediated the dialogues was designed, and details of how the data was collected and annotated.

  – A description of the collected data has been reported in my LREC 2016 (Hayakawa, Luz, Cerrato, & Campbell, 2016) paper.

- Chapter 5: Speech Rate Difference

  In this chapter, I explain one of the main contributions of this thesis — the speech rate calculation method used to identify if the subjects increased or decreased their speech rate. This speech rate difference uses the duration of an utterance and compares it with the duration of a text-to-speech synthesised output of the same utterance to identify if it is spoken faster or slower than the text-to-speech output reference. The speech rate difference, calculated with the method explained in this chapter was used in the analysis of Chapter 6, Chapter 7, and Chapter 9.

  – The method of calculating the speech rate has been reported in my LREC 2018 (Hayakawa, Vogel, Luz, & Campbell, 2018) paper.

- Chapter 6: The Three Talk Types

  In this chapter, I analyse the "*Talk Types*" labels explained in detail on page 90 (Chapter 4: "Data Created for this Study") and briefly on page 37 (Chapter 3: "The Datasets"). I look into the speech rate difference pattern between the three talk types of *On-Talk*,[5] *Off-Talk Self*,[6] and *Off-Talk Other*.[7] This resulted with *On-Talk* displaying the slowest mean value, followed

[5] Subject utterances to the interlocutor via the computer system.
[6] Subject utterances to themselves.
[7] Subject utterances face-to-face to someone else.

by *Off-Talk Self*, and *Off-Talk Other* displaying the fastest speech rate mean value, which indicates that the subjects use different speech rates depending on the situation surrounding the communication.

  - Results related to this analysis have been reported in my Speech Prosody 2016 (Hayakawa, Haider, Luz, Cerrato, & Campbell, 2016) paper and/or also mentioned in my LREC 2016 (Hayakawa, Luz, Cerrato, & Campbell, 2016), INTERSPEECH'16 (Hayakawa, Vogel, Luz, & Campbell, 2017b), and CogInfoCom 2017 (Hayakawa, Vogel, Luz, & Campbell, 2017a) papers.

- Chapter 7: Speech Rate Differences Between Datasets

  In this chapter, I compare the speech rate of the utterances in the computer-mediated dialogue dataset and the non-mediated dialogue dataset. This comparison is to identify if, (i) the subjects of both datasets use a similar speech rate when talking directly to a fellow human, (ii) what patterns, if any, differ between subsets of the data such as gender, role, and setting and (iii) how the patterns identified in (ii), if any, differ between the two datasets. The results indicated that the speech rate of *Off-Talk Other* in the computer-mediated dialogue dataset and the utterances of the non-mediated dialogue dataset are similar. This signaled that in similar situations (talking directly to a fellow human), the subjects of the two dataset used similar speech rates and strengthened the findings of Chapter 6 by anchoring one of the talk type speech rates. Furthermore, similarities in speech rate patterns were identified in the two dataset subsets, but also a difference was identified in the subset of gender — a difference that is not found so strongly in the non-mediated dialogue dataset, where female subjects talking significantly slower than male subjects.

  - Results related to this analysis have been reported in my INTERSPEECH'17 (Hayakawa et al., 2017b) paper and/or mentioned in my CogInfoCom 2017 (Hayakawa et al., 2017a) paper.

- Chapter 8: Dialogue Acts Differences Between Datasets

  In this chapter, I compare the communication style using the dialogue act annotations of the direct H2H dialogues in the non-mediated dialogue dataset with the computer-mediated dialogue dataset. A difference in the speech rate patterns were identified in Chapter 7 and this comparison is to identify if, (i) the subjects structure their dialogues differently with the addition of a speech-to-speech machine-translation system in between the subjects and (ii) what patterns, if any, differ between datasets. Results from this comparison indicate that the the communication style used with the ILMT-s2s System has characteristics similar to that of high latency video conference meetings — greater effort is required to communicate and hence communication items become richer in contents.

  - Results related to this analysis have been reported in my INTERSPEECH'16 (Hayakawa, Luz, & Campbell, 2016) paper and/or mentioned in my CogInfoCom 2017 (Hayakawa et al., 2017a) paper.

- Chapter 9: ASR Performance & Speech Rate

  In this chapter, I study the *On-Talk* speech rate of the subjects, to see if the variation in the speech rate is related to Automatic Speech Recognition (ASR) performance. This analysis is to identify if, (i) the cause of the slower speech rate of *On-Talk* in the computer-mediated dialogue data is originating from the speech-to-text component of the computer system and (ii) what patterns, if any, differ between subsets of the data such as gender, role, and setting. Results from this analysis are inconclusive in identifying a correlation between the speech rate of the subject and the ASR performance of the computer system. However, one finding of interest was that subjects increased their speech rate after a system reboot. This indicated that the subjects reset their expectation, in other words, they think that it is possible to use a speech rate they would use towards a fellow human, which indicates that the performance of the system, or expectation of the performance of the system is affecting the adaptation strategies that the subjects use. This directs me to the conclusion that a deeper, more individual analysis of each value is needed to obtain a stronger convincing statement of the results. For example, not only better understanding of how the speech rate changed with each ASR performance value, but also a more detailed understanding of other errors of communication related to the map task that are simultaneously occurring during the dialogue.

  – Results related to this analysis have been reported in my ICPhS 2015 (Hayakawa, Cerrato, Campbell, & Luz, 2015) paper and/or mentioned in my LREC 2016 (Hayakawa, Luz, Cerrato, & Campbell, 2016), and CogInfoCom 2017 (Hayakawa et al., 2017a) papers.

- Chapter 10: ASR Performance & Perceived Emotions Analysis

  In this chapter, I analyse the "*perceived emotions*" labels explained in detail on page 82 (Chapter 4: "Data Created for this Study") and briefly on page 37 (Chapter 3: "The Datasets") and its relation with the ASR performance. The performance of the ASR system was known to return erroneous results, so the assumption was that these ASR errors would change the perceived emotion of the subject. This analysis is to identify if, (i) the ASR performance causes the a change to the subject's emotion or if the perceived emotion of the subject is affecting the ASR performance and (ii) what patterns, if any, differ between subsets of the data such as language, and setting. Results from this analysis indicate that the perceived emotions of "*Surprised*", "*Frustrated*", and "*Amused*" are an effect of the ASR performance. Furthermore, a difference in how the subject displayed their emotion to various ASR performance ranges differed by the language spoken by the subject. The the clearest being the usage of "*Amused*". The Portuguese speakers showed amusement at a lower WER than the mean neutral state while the English speakers showed it at a higher mean to the neutral state. It would be interesting to see if the same order is maintained if the experiment was conducted with Portuguese speakers living in their own country (Brazil or Portugal) and not residing in foreign country where they might

have unconsciously obtained an inferiority complex and hence acting more conservatively.

  - Results related to this analysis have been reported in my INTERSPEECH'15 (Hayakawa, Haider, Cerrato, Campbell, & Luz, 2015) and CogInfoCom 2017 (Hayakawa et al., 2017a) papers and/or mentioned in my LREC 2016 (Hayakawa, Luz, Cerrato, & Campbell, 2016) paper.

- Chapter 11: User Survey Study

  In this chapter, a qualitative look at the experience of the subjects using the computer-mediated dialogue system. Using fifteen statements in a 7 point Likert scale, this analysis is to identify if, (i) the subjects overall have a certain sentiment towards the computer system and (ii) a certain subset of the subjects have a preference towards the computer system. This is based on the assumption that the subjects using the different settings and playing the different roles would have a different appreciation of the computer system. The results of this analysis indicated that (i) the overall experience with the computer system was positive, (ii) the subjects the could not see their counterparts had a better experience, (iii) the subjects with the role of Information Giver had a better experience and (iv) once combined, the Portuguese speakers had a better experience except for those with the role of Information Follower in the setting where they could see their counterpart. As mentioned in the introduction to Chapter 10 above, it would be interesting to see if the Portuguese speaking subjects with the role of Information Follower and the setting where they could see their English speaking counterpart would have had an equally negative experience if they were participating in this study from their native country.

  - Results related to this analysis have been mentioned in my CogInfoCom 2017 (Hayakawa et al., 2017a) paper.

- Chapter 12: Discussions

  In this chapter I report how the results from the six studies of § 6 to § 10 are connected to the research question proposed in § 1.2 and list future work related to this thesis.

- Chapter 13: Conclusions

  In this chapter I concludes the findings of this thesis.

- The appendices start on page 224.

  They are included for completeness. Appendix A to Appendix C lists supplementary information that the author thinks will help anchor the discussion of this thesis.

Chapter 2

Literature Review

This chapter presents the general framework into which this research took place, introduces the current state of research around Speech-to-Speech Translation Systems, some of the previous projects and exploring their possibilities, and finally details which are the factors explored that require attention to answer the proposed research questions.

2.1 Speech-to-Speech Translation Systems

Hara & Iqbal (2015) introduced the problems that the language barrier is presenting in education (Morita, 2004), health care (Kale & Syed, 2010) and business collaboration (Feely & Harzing, 2003). They indicate that a cost effective countermeasure would be the usage of S2S-MT systems, since the service of a human translator or interpreter would not be feasible for all required occurrences. The authors present the findings of S2S-MT based communication in German and English over a platform very similar to the Skype Translator system (W. D. Lewis, 2015). The authors also mention that there is little published research in this domain — the domain of how users interact and adapt to S2S-MT.

Earlier, Yamashita & Ishida (2006b) study how users adapt to Text-to-Text Machine Translation (T2T-MT) mediated communication and also compare it with L2 (second language) mediated communication. To be more precise, they look into lexical entrainment (Brennan & Clark, 1996; Isaacs & Clark, 1987) of the subjects' dialogue while using a T2T-MT system. They report how Machine Translation (MT) can be fine tuned to create more than adequate oneway translation, but that MT has difficulty mirroring the terms and language usage in twoway translation as is required in dialogue interaction. Here too, the authors mention that there was little published research in the domain of MT mediated communication. An interesting aspect of the findings of Yamashita & Ishida (2006b) is that they seem to present the next step in subject adaptation to an S2S-MT communication system when the results from Schneider & Luz (2011) are taken into account. Schneider & Luz (2011) conclude that subjects of S2S-MT mediated communication will adjust their choice of words and mirror even incorrect, yet understandable, words that an S2S-MT system uses. Schneider & Luz (2011) go on

to discuss how software designers would need to incorporate this behaviour and log the usage of terms used in the translation of one direction so that it can be reused in the opposite direction. Though the findings of Yamashita & Ishida (2006b) are not related to the mirroring of erroneous MT output that is reported by Schneider & Luz (2011), but both papers recognise the importance of subject entrainment and MT in a dialogue situation and not only oneway transactions. With the findings of Yamashita & Ishida (2006b) it is predictable that users of S2S-MT communication systems will repeat[1] the words of the system initially, but as soon as reliable understandable results are unattainable, communication fails with the interlocutor, due to confusion from mistranslation: "since the Chinese to Japanese translation translated 'dance' into 'jump,' the Japanese participant got confused" (Yamashita & Ishida, 2006b, p. 520) — the subjects will change the structure of their replies to minimise the possibility of error — changing their communication to the minimum required. This would change the structure of the dialogues from a collaborative effort (Clark & Wilkes-Gibbs, 1986; Clark & Brennan, 1991) to a somewhat more Gricean (maxim theory) interaction structure (Davies, 2000) where the subject places emphasis on the "Cooperative Principle" of communication (H. P. Grice, 1975).

The above is one example of the how research in H2H communication was tested in T2T-MT mediated communication and then onto structured S2S-MT mediated communication that should influence the design of S2S-MT communication systems. However, there is little published research on this subject (Hara & Iqbal, 2015; Yamashita & Ishida, 2006b).

2.1.1 S2S-MT Projects to Date

To date, there have been several major S2S-MT system projects. e.g., (i) The *interpreting telephone* (Abe, Shikano, & Kuwabara, 1990), (ii) VERBMOBIL (Bub, Wahlster, & Waibel, 1997; Nöth, Batliner, Kießling, Kompe, & Niemann, 2000), (iii) C-STAR (Blanchon & Boitet, 2000), (iv) TC-STAR (Höge, 2002), (v) The Tongues Speech-to-Speech Machine Translation System (Frederking, Black, Brown, Moody, & Steinbrecher, 2002), (vi) NESPOLE! (Lavie, Pianesi, & Levin, 2006), (vii) The ATR Mulitilingual Speech-to-Speech Translation System (Nakamura et al., 2006), (viii) The 2006 BBN English/Iraqi Two-way Speech Translation System (Stallard et al., 2006), (ix) The EMIME Project,[2] and (x) VoiceTra (Matsuda et al., 2013).

Of these ten major projects, few study user interaction with an S2S-MT system. Frederking et al. (2002) and Stallard et al. (2006) present a description of their system created for the US military with a user study analysis that evaluates the system presented as part of the result. The NESPOLE! project picks up on the usability of an S2S-MT system based on a multimodal communication Wizard-of-Oz (WOZ) setup, that used two forms of input, verbal and textual input and reported the user selection of more reliable methods for tasks that the user perceived to be more efficient (Oviatt, 1997; Oviatt, DeAngeli, & Kuhn, 1997; Oviatt, 1999). In a separate study, Batliner also investigates how the user

[1] Literature describes this phenomenon in various terms (e.g., alignment, entrainment, priming, lexical convergence, mirroring, etc.).

[2] The EMIME project: http://www.emime.org (accessed on 2017/04/07)

interacts with the system and specifically, when the subject talks to the system and when the subject is not talking to the system and if this can be automatically detected — "*On-Talk*" and "*Off-Talk*" (Batliner, Hacker, & Nöth, 2006, 2009). It is difficult to think that these major projects do not research how users use and adapt to their systems, but to the best of my knowledge there seem to be little published research in this direction. However, as mentioned in the next § 1.2, once the domain is changed, numerous references can be found — adaptation to the interlocutor, agent or system.

2.1.2 Map Task Corpora

To elicit dialogue the map task technique was used, and rather than creating both a computer mediated version and a non-mediated version of the dialogues, the HCRC Edinburgh Map Task corpus (Anderson et al., 1991) was used as the reference dialogues to the computer-mediated dialogues. The HCRC Edinburgh Map Task corpus has been widely studied and replicated in a number of languages including Dutch (Lickley, Schepman, & Ladd, 2005, p.174), Italian (M. Grice & Savino, 2003, p.155), German (Schmidt, Hedeland, Lehmberg, & Wörner, 2010), Japanese (Horiuchi et al., 1999), Swedish (Helgason, 2006), Portuguese (Trancoso, Viana, Duarte, & Matos, 1998)[3] to name a few, and a variety of English accents such as the MIT American English Map Task, the Australian Map Task. Though papers of the HCRC Edinburgh Map Task corpus were originally published in 1991, this map task continues to be used today, as it still constitutes a particularly fit object of study, given all the controlled factors it includes.

2.2 Aspects Influencing and Being Influenced by S2S-MT Systems

In this section, the focus is on the possible factors that influence how the subject and interlocutor adjust their communication style, and how it compares with direct Human-to-Human communication, namely: the speech rate of the subjects, the ASR Performance (e.g., word error rate) of the ILMT-s2s System, the dialogue acts used within the dialogues of the dataset and the perceived emotions of the subjects. These factors are additionally explored in interaction to controlled sociological factors: gender, eye-contact (either face-to-face or through a video medium) and task role. In other words, how the subjects adjust their speaking style and speaking rate to adapt to the ILMT-s2s System which mediates their communication, and also how different or similar these adaptations are to direct Human-to-Human communication.

[3]The CORAL corpus ISLRN: 499-311-025-331-2

2.2.1 Speech Rate

Spoken dialogue can be seen as a joint action in which interlocutors coordinate their verbal and non-verbal behaviour. By doing so, they adapt their linguistic choices to each other which often results in the phenomenon referred to as convergence or alignment. As reported by Giles et al. (1987), convergence or alignment is a tendency of the subject to adopt and re-use the interlocutor's language pattern in the course of interaction (dialogue). According to the Interactive Alignment Model, originally proposed by Pickering & Garrod (2004) this linguistic coordination in dialogue occurs at many levels such as lexicon, syntax, and pronunciation which represents one way in which interlocutors achieve understanding in Human-to-Human dialogue. Previous studies have suggested that this is a phenomenon of human dialogue that persists even when one of the interlocutors is replaced by a computer interface. L. Bell et al., (2003, p. 2453ff.) reported how adult subjects adapted their speech rate to that of the varying Text-to-Speech (TTS) output of an artificial agent as follows:

> "The experiment confirmed that the users adapted to the speaking rate of the system, and no subjects afterwards seemed to be aware they had been affected in this way. Another finding was that speakers varied their speaking rate substantially in the course of the dialogue. In particular, problematic sequences where subjects had to repeat or rephrase the same utterance several times elicited slower speech. [...] That is, they use the same means to increase intelligibility as they would in dialogue with other humans." (L. Bell et al., 2003, p. 2453ff.)

Oviatt, Darves, & Coulston (2004, p. 320) reported how children vary their speech rate depending on that of their interlocutor they were interacting with, which in this case were different animated agents with varying personas:

> "[...] 7 to 10-year-old children adapted the amplitude and durational features of their speech by a substantial 10–50% when interacting with animated characters representing different TTS voice styles. These changes in their speech patterns were also bidirectional. That is, when exposed to an extrovert voice, child speakers increased their amplitude and decreased utterance duration and dialogue response latencies. Conversely, when interacting with an introvert voice, they decreased amplitude and increased utterance duration and response latencies. This change observed in children's speech represented a substantial convergence toward their computer partner's TTS output." (Oviatt, Darves, & Coulston, 2004, p. 320)

Also, Suzuki & Katagiri (2007, p. 139) tried to see if they could manipulate the subject to talk at a certain rate (faster or slower) by changing the speech rate of the TTS system:

> "As a result, we found that participants' speech during interaction with the system aligned

at least unidirectionally, even though it was not clearly adapted bidirectionally. While participants produced a louder voice in response to slight increases in the loudness of the system's voice, they did not produce a softer voice in response to slight decreases in loudness of the system's voice. Moreover, while participants produced a shorter response latency in response to a slight decrease in the response latency in the system's voice, they did not produce a longer response latency based on slight increases in the response latency in the system's voice." (Suzuki & Katagiri, 2007, p. 139)

Given that there are reports that interlocutors adapt their linguistic choices to each other and that humans tend to adapt their speech even to that of a conversational interface, the initial task was to verify if the same occurred when the subject and interlocutor are placed in a setting where their interaction is mediated by a translation system (the ILMT-s2s System).

2.2.2 Dialogue Acts and Communication Medium

As mentioned in the previous section § 2.2.1: "Speech Rate", spoken dialogue can be seen as a joint action in which interlocutors coordinate their verbal and non-verbal behaviour. These joint actions can be identified and distinguished from each other, and labelled into categories, to obtain a better understanding of how a dialogue unfold, and observe its structure. Among the possible coding schemes used to segment and label verbal communication, I choose to focus on the Dialogue Structure Coding scheme developed by Carletta et al.(1997), as it was specifically created for a use on the HCRC Map Task. To summarise the scheme, the dialogues are divided into *transactions*, that are sub-dialogues constituted of the major steps into the accomplishment of the task. These transactions are divided into *games* that are characterised by their purpose into the discourse, and differentiate between the *initiations* and the *responses*, and finally the conversational games are divided into *moves*. This coding scheme labels utterances with conversational move categories and in the case of this thesis, these units are called dialogue acts (namely: instruct, explain, check, align, query-yn, query-w, acknowledge, reply-n, reply-w, and clarify).[4]

These dialogue acts are important tools to understand the structure of a task-based interaction, however, in order to answer the previously given research questions, this thesis examines how this dialogue acts are influenced by the medium of communication. In a previous study by O'Conaill, Whittaker, & Wilbur (1993, p. 420) which compared high speed high quality video conference communication with low speed low quality video conference communication and Face-to-Face communication, it has been reported that the latency added to the communication flow by a system affects the communication style. High latency video conference systems produce fewer, but longer utterances.

[4]Changes have been added to the Dialogue Structure Coding scheme with modifications to "Acknowledgement" to differentiate the simple acknowledgement (CP) of the utterance from the acknowledgement with the actual understanding of the utterance (CPU) which is defined under the MUMIN coding scheme (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). Also, a "Solo" variant of the dialogue acts, and "Interjection" were added due to the "*Off-Talk*" characteristics (Oppermann, Schiel, Steininger, & Beringer, 2001) of the collected data.

Lower latency systems, shorter but more utterances, a style closer to Face-to-Face communication as extracted below:

> **"**Our results showed that, compared with face-to-face meetings, spoken conversation patterns are disrupted over ISDN with its half-duplex line, transmission lags, and poor image quality
>
> - Listeners produced fewer backchannels and interrupted less often.
> - Listeners were also less likely to anticipate turn endings.
> - Speakers also alter their behavior, being more likely to hand over turns formally using question or naming the next speaker. They were also less likely to hold the floor with redundant phrases.
> - The result of listeners reducing interruptions and speaker feedback, combined with the general difficulty of switching speakers, was a formal lecture style of interaction, with long turns, handed over by a very deliberate process.
>
> In LIVE-NET, even when there is a full-duplex line, immediate transmission, and broadcast-quality image, the properties of the spoken communication still differ from face-to-face interaction.
>
> - Although listeners interrupt as frequently as in face-to-face, they are less likely to give backchannels.
> - Speakers use questions to formally hand over the floor more frequently, and they are also less likely to hold the conversation floor with redundant information.**"** (O'Conaill et al., 1993, p. 420)

The point of concern here is that users of the video-conferencing devices (i) used fewer backchannels, (ii) used more formal language structures, and (iii) the latency of the device changed the style of communication. Another interesting point concerning communication style was reported in a reference to the publication in Japanese of Yoshida and Kakuta (Morikawa & Maesako, 1998, p. 149), that users change their communication style depending on the medium:

> **"**A study by Yoshida et al. showed that college students in and around Kyoto–Osaka (effective response 549) cited the following reasons for the phone's popularity: speed, available anytime, available anywhere, no visual information about the other end, and easier way to say what we want than in face-to-face conversation. They also cited the following disadvantages: no visual information about the other end and difficulty in conveying subtle emotion. How interesting that no visual information about the other end becomes both an advantage *and* disadvantage !

> The above study suggests that the telephone provides its own conversation environment rather than a substitute for face-to-face conversation – in short, people have found a new way to communicate using voice alone." (Morikawa & Maesako, 1998, p. 149)

The expectation from these two studies is that the communication style of the computer-mediated dialogue data will differ from that of the non-mediated dialogue data. While the direct H2H communication, like that of the non-mediated dialogue dataset, enables speakers to benefit from non-verbal acts, gestures and facial expressions. Machine mediated conversation on the other hand, like that of the computer-mediated dialogue dataset, is more complex for the subject and interlocutor. Due to latency from the system, errors made by Automatic Speech Recognition (ASR) system, the subjects adapt their speaking style and also apply repair strategies in order to accomplish the tasks successfully. However, even with improvement to ASR performance and other system latency problems, the characteristics of interpreting one language to another will continue to provide a latency that is uncommon in monolingual communication.

2.2.3 Perceived Emotion (Cognitive State)

The computer-mediated dialogue data was collected with the use of two main tools, which were the ILMT-s2s System and two maps used in the HCRC Edinburgh Map Task corpus. In other words, a S2S-MT system to facilitate the communication, and the map task technique to elicit communication. The maps of the HCRC Edinburgh Map Task corpus where not designed to be easy. They were designed to cause miscommunication at identified locations with the subjects tasked to overcoming the obstacles as Anderson et al. (1991, p. 353) explains;

> "These skills have to do with speakers' ability to establish sufficient "mutual knowledge" (Clark & Marshall, 1981) to understand one another's contributions to the conversation. Any mismatch between landmarks on Giver's and Follower's maps makes this requirement particularly salient and gives rise to different strategies in more and less successful communicators." (Anderson et al., 1991, p. 353)

Furthermore, the limitations of the S2S-MT system, with ASR errors, Machine Translation (MT) that is based on erroneous ASR results, and the latency that is inherent to interpretation (translation) would cause an effect on the subjects. The assumption was that this would bring out cognitive characteristic of "*Surprised*", "*Frustrated*", and "*Amused*" in the interaction with the ILMT-s2s System. Batliner, Fischer, Huber, Spiker, & Nöth (2000, p. 195) provide an extremely pragmatic reason to study this: Cost.

> "The longer the communication is performed automatically with a pleased user/costumer the cheaper it is; if the user becomes annoyed and irritated, however, such that he or she breaks off the communication (hangs up) the costs are fatal — one more costumer gone. It

is therefore desirable to be able to find the beginning of the critical phase in the dialogue well before the point of no return.**"** (Batliner et al., 2000, p. 195)

Picard (1995, p. 14) uses a slightly more academic approach of reasoning about the various scenarios where understanding the emotion/cognitive state of the subject will affect the resulting outcome of the interaction and concludes that

> **"**Emotions have a major impact on essential cognitive processes; neurological evidence indicates they are not a luxury. I have highlighted several results from the neurological literature which indicate that emotions play a necessary role not only in human creativity and intelligence, but also in rational human thinking and decision-making. Computers that will interact naturally and intelligently with humans need the ability to at least recognize and express affect.**"** (Picard, 1995, p. 14)

My motivation for looking into the three perceived emotions of "*Surprised*", "*Frustrated*", and "*Amused*" in the data is related more to that of Batliner et al. (2000). Not only the financial cost of losing a customer, but also the emotional cost for the communication breakdown; both from the manufacturing side and the consumer's. As Botherel & Maffiolo (2006, p. 3) explains, from a study of two call-centres "emotional attitudes mainly detected in the customers' speech are negative, linked up with difficulties they have" which from the customers' point of view could derive from the "when the individual customer's lack of sovereignty becomes starkly apparent. [...] This anger is vented at the closest symbol of customer's painful journey from enchantment to disillusionment" (Korczynski, 2003, p. 57). This "lack of sovereignty" could easily be replicated in communication with a S2S-MT system.

2.2.4 Different Types of Talk

A phenomenon that requires attention when one studies the usage of a speech-to-speech machine translation systems is the type of talk, where it can be distinguished by: to whom the speech is addressed to. In other words, the phenomenon of *Off-Talk*. Batliner et al. (2006) open their article with an example from Shakespeare's *Hamlet*, where Hamlet seems to change his speaking style when addressing his interlocutor, to that of utterances which are spoken, but not directed towards his interlocutor. This shows that this is not a new phenomenon, but part of human nature that even Shakespeare expressed with his characters. However, the dialogues of the computer-mediated dialogue dataset are not that of a play like Shakespeare's *Hamlet*. It is a collection of task oriented computer mediated interlingual dialogues. The *Off-Talk* by the subjects of the computer-mediated dialogue dataset are fundamentally the same as the definition of *Off-Talk* provided by Oppermann et al. (2001, p. 1):

> **"We define Off-Talk to be every utterance that is not directed to the system as a question, a feedback utterance or as an instruction.**

The following is an open list of possible utterances which could be considered as Off-Talk in human-machine-interaction:

- soliloquy / thinking aloud
- swearing
- reading from displayed text aloud
- conversation with other person(s) present
- telephone conversation (e.g., with cellular phone)
- extrinsic speech (e.g., video player, TV set, etc.)" (Oppermann et al., 2001, p. 1)

Oppermann et al. (2001) look into the type of *Off-Talk* by differentiating *Off-Talk* into two categories, "ROT" (Read Off-Talk) and "OOT" (Other forms of Off-Talk) which looks at the source of the utterance — Is the *Off-Talk* something that is read, or some other type of communication? For the computer-mediated dialogue data collection, *Off-Talk* is differentiated by the direction of the utterance — Is the subject talking to themselves or to someone else other than themselves and their interlocutor through the ILMT-s2s System?

2.3 Conclusion

In conclusion of this review, a number of aspects still remain unknown in the way users adapt to an S2S-MT system and how does it compares to Human-to-Human communication. By exploring variations in terms of Speech Rates, types of dialogues acts, labeled perceived emotions and types of talks (On and Off), the following chapters provide answers to the research questions given in 1.2.2.

Chapter 3

The Datasets

3.1 Introduction

This study uses two dataset of task oriented dialogue data. For the non-mediated dialogue data, which is used as a reference standard in this study, the HCRC Edinburgh Map Task corpus is used (§ 3.2). For the computer-mediated dialogue data, which is used as the core of this study, the ILMT-s2s corpus is used (§ 3.3).

Methodological Note

The computer-mediated dialogue data (the ILMT-s2s corpus) was collected for the purpose of this study. For details of the whole dataset and the method of collecting and creating the data, refer to Chapter 4, "Data Created for this Study".

3.2 The Non-Mediated Dialogue Dataset

For the non-mediated dialogue data the HCRC Edinburgh Map Task corpus (Anderson et al., 1991) is used. For the latest information about this corpus, refer to the HCRC Edinburgh Map Task corpus web-site http://groups.inf.ed.ac.uk/maptask/.

3.2.1 Introduction of the HCRC Edinburgh Map Task Corpus

The HCRC Edinburgh Map Task corpus consists of 128 dialogues of direct Human-to-Human task based interactions. There are 64 native English speaking subjects, mostly from Scotland, who were even split in gender and familiarity who participate in the map task 4 times each. As quoted in the journal paper (Anderson et al., 1991, pp. 360–361):

> "Every subject participated in the Map Task four times, twice as Instruction Giver, twice as Instruction Follower, once in each case with his or her Familiar partner, once with an Unfamiliar partner. As Instruction Giver, each Subject used the same map twice; as

> Follower s/he used a different map each time. Half of the Subjects gave instructions to a Familiar partner first, half to an Unfamiliar. Half the Subjects performed all four tasks while able to see the other participant's face (With Eye Contact), half while unable to do so (Without Eye Contact)." (Anderson et al., 1991, pp. 360–361)

The dialogues within the corpus was also graded with a score that calculated the success of communication called the deviation score.[1] This score was a measure of task success that quantified the deviation of the Information Follower's completed route to the the original route on the Information Giver's map (Newlands et al., 2003, p. 331).

> "[An] acetate was made of the original Instruction Givers map which showed the intended route, this was overlaid with a grid of one centimetre squares. The acetate was placed on top of the Instruction Follower's completed map, and the area (in square centimetres) between the two routes computed." (Newlands et al., 2003, p. 331)

This totaled in 14 h 24 m 19 s (51,859 seconds) of data for the 128 dialogues with the relevant success of each dialogue quantified in the deviation score. A summary of the gender divide between roles (Information Giver as IG, and Information Follower as IF), the duration of the dialogues and the mean deviation score of the subjects per map type[2] and per map are listed in Table 3.1.

	Gender by Role				Duration (seconds)				Dialog	*M* Dev.
	IG – ♀	IG – ♂	IF – ♀	IF – ♂	*Mdn*	*M*	*SD*	Total	Count	Score
All Dialogues	64	64	64	64	354.0	405.1	185.73	51,859	128	71.8
Map Type 1	8	24	18	14	418.0	449.7	199.93	14,390	32	75.7
Map Type 2	20	12	16	16	337.0	402.1	202.93	12,867	32	85.7
Map Type 3	14	18	14	18	337.0	358.1	121.37	11,458	32	72.0
Map Type 4	22	10	16	16	351.0	410.8	202.65	13,144	32	53.9
Map 00 – Type 1	0	8	4	4	508.0	462.6	164.92	3,701	8	55.4
Map 04 – Type 1	4	4	4	4	363.5	359.8	80.51	2,878	8	54.0
Map 08 – Type 1	0	8	4	4	437.5	489.6	211.29	3,917	8	76.6
Map 12 – Type 1	4	4	6	2	433.5	486.8	293.03	3,894	8	116.6
Map 01 – Type 2	2	6	4	4	344.5	418.4	314.19	3,347	8	64.0
Map 05 – Type 2	6	2	3	5	326.0	372.9	115.49	2,983	8	94.0
Map 09 – Type 2	6	2	6	2	333.5	413.2	221.87	3,306	8	122.1
Map 13 – Type 2	6	2	3	5	390.5	403.9	140.26	3,231	8	62.6
Map 02 – Type 3	2	6	3	5	337.0	361.0	121.47	2,888	8	82.6
Map 06 – Type 3	6	2	5	3	310.0	353.0	127.86	2,824	8	50.5
Map 10 – Type 3	2	6	3	5	403.0	391.6	144.56	3,133	8	70.6
Map 14 – Type 3	4	4	3	5	305.0	326.6	103.80	2,613	8	84.4
Map 03 – Type 4	6	2	4	4	273.0	299.4	110.51	2,395	8	59.9
Map 07 – Type 4	6	2	4	4	466.0	499.4	199.56	3,995	8	32.9
Map 11 – Type 4	4	4	3	5	454.5	513.2	277.35	4,106	8	67.3
Map 15 – Type 4	6	2	4	4	325.0	331.0	102.37	2,648	8	55.8

Table 3.1: *Summary of HCRC Edinburgh Map Task corpus dialogue duration by maps*

[1] The deviation scores are publicly available at http://groups.inf.ed.ac.uk/maptask/deviation-scores.txt (accessed on 2017/04/07).

[2] Refer to Figure 3.1 of § 3.2.3 for an illustration of the 4 map types.

3.2.2 The Selection of the Two Maps

A subset of 16 dialogues were used from the total of 128. The 16 dialogues are those that were recorded using Map 01 and Map 07. The 16 dialogues totaled in 2 h 2 m 22 s (7,342 seconds) of data as summarized in Table 3.2.

	Gender by Role				Duration (seconds)				Dialog	*M* Dev.
	IG – ♀	IG – ♂	IF – ♀	IF – ♂	*Mdn*	*M*	*SD*	Total	Count	Score
Map 01 & Map 07	8	8	8	8	406.0	458.9	257.69	7,342	16	48.4
Map 01 – Type 2	2	6	4	4	344.5	418.4	314.19	3,347	8	64.0
Map 07 – Type 4	6	2	4	4	466.0	499.4	199.56	3,995	8	32.9

Table 3.2: *Summary of HCRC Edinburgh Map Task corpus dialogue used*

The two maps have been selected for the following reasons:

(i) The ease of the route, derived from a low mean deviation score of the map.

(ii) A longer mean duration of the map, so that the dialogues will not be too short.

(iii) Opposing directions of the route, with the START located in the top or bottom half of the map.

(iv) A close to even gender balance.

	Gender Balance		Gender by Role				*M* Duration		*M* Dev.		Rank
	Female	Male	IG – ♀	IG – ♂	IF – ♀	IF – ♂	Sec.	Rank	Score	Rank	Total
Map 07 – Type 4	10	6	6	2	4	4	499.4	2	32.9	1	3
Map 00 – Type 1	4	12	0	8	4	4	462.6	5	55.4	4	9
Map 11 – Type 4	7	9	4	4	3	5	513.2	1	67.3	9	10
Map 01 – Type 2	6	10	2	6	4	4	418.4	6	64.0	8	14
Map 08 – Type 1	4	12	0	8	4	4	489.6	3	76.6	11	14
Map 04 – Type 1	8	8	4	4	4	4	359.8	12	54.0	3	15
Map 13 – Type 2	9	7	6	2	3	5	403.9	8	62.6	7	15
Map 06 – Type 3	11	5	6	2	5	3	353.0	13	50.5	2	15
Map 12 – Type 1	10	6	4	4	6	2	486.8	4	116.6	15	19
Map 15 – Type 4	10	6	6	2	4	4	331.0	14	55.8	5	19
Map 10 – Type 3	5	11	2	6	3	5	391.6	9	70.6	10	19
Map 03 – Type 4	10	6	6	2	4	4	299.4	16	59.9	6	22
Map 02 – Type 3	5	11	2	6	3	5	361.0	11	82.6	12	23
Map 09 – Type 2	12	4	6	2	6	2	413.2	7	122.1	16	23
Map 05 – Type 2	9	7	6	2	3	5	372.9	10	94.0	14	24
Map 14 – Type 3	7	9	4	4	3	5	326.6	15	84.4	13	28

Table 3.3: *Ranking of HCRC Edinburgh Map Task corpus dialogues by maps*

To determine the maps that are used, the mean deviation score, the mean duration and the gender balance of each map was calculated. Of these, the mean deviation score was ranked from lowest to highest, and the duration was ranked from longest to shortest. These two rankings were added together to make a list that takes both rankings into account as presented in Table 3.3. With Map 07 selected for having the best ranking, other maps starting in the bottom half were removed from the selection (indicated as "Type 3" and "Type 4" in Table 3.3). Of the remaining maps, Map 01 was selected since it provided an even gender balance when combined with Map 07, when compared to the higher ranked map of Map 00.

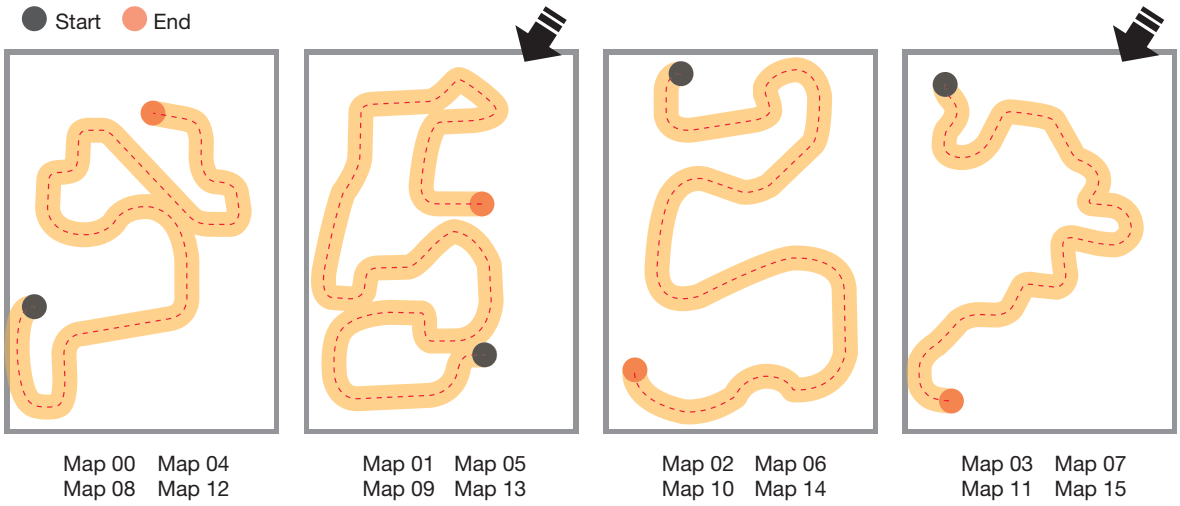


Figure 3.1: *The 4 route patterns of the HCRC Edinburgh Map Task corpus maps*

3.2.3 The Subjects of Map 01 & Map 07

As explained in the previous §3.2.2, the two maps were selected for their low mean deviation score, high mean duration, opposing direction of the routes and a close to even gender balance.

As illustrated in Figure 3.2, twenty four subjects participated in the collection of the sixteen dialogues, with each Information Giver taking part in two dialogues. Once with an interlocutor that they are familiar with, and again with an interlocutor they are not familiar with. The dialogues of Map 01 provided five female subjects and seven male subjects, while Map 07 provided seven female subjects and five male subjects. This resulted in the overall balance being an even twelve female subjects and twelve male subjects. The gender balance was also carried over to the roles of the subject, with four female subjects and four male subjects given the role of Information Giver and eight female subjects and eight male subjects given the role of Information Follower.

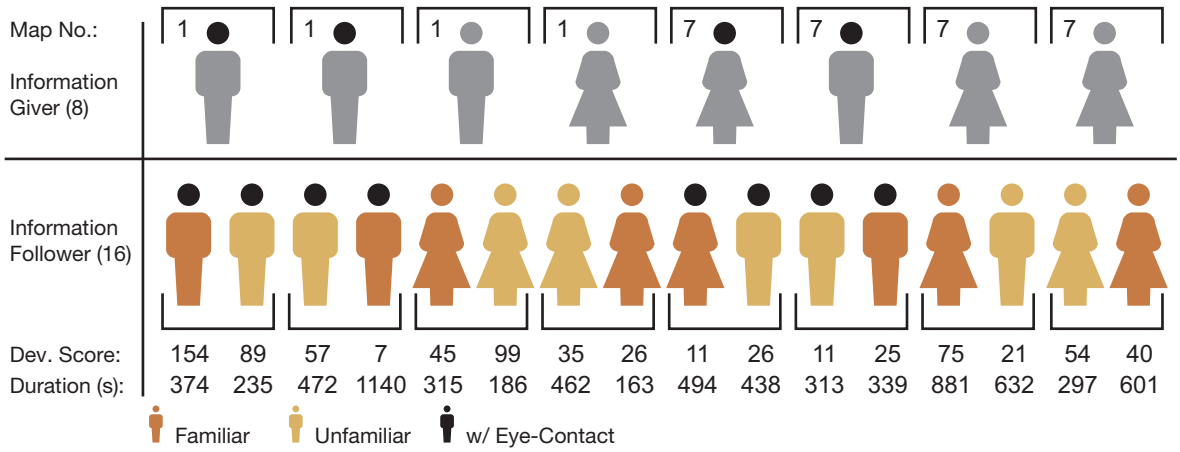


Figure 3.2: *Image of subject pairs and recording situation with deviation score and duration of dialogues*

The subjects are native English speakers, with twenty-three of the twenty-four subjects from Scotland and one from the USA. The setting of the recordings (with or without eye-contact) was evenly split so twelve subjects could see their interlocutor while the remaining twelve couldn't. The

deviation score and duration (in seconds) of the sixteen dialogues are included in Figure 3.2.

3.2.4 The Dialogues of Map 01 & Map 07

The data of the computer-mediated dialogues consist of sixteen stereo recordings as summarised in Table 3.4. The sixteen recordings total to 2h 02m 22s of data. The orthographically transcription and annotation data has be obtain as explained in the following § 3.2.5 "The Transcription & Annotation Data".

Dialogue	Duration	Utt. Count	(IG : IF [% Diff.])	Word Count	(IG : IF [% Diff.])
m1 1 w/ EC	00:06:14	140	(77 : 63 [−18%])	1,020	(744 : 276 [−63%])
m1 2 w/ EC	00:03:55	72	(40 : 32 [−20%])	587	(463 : 124 [−73%])
m1 3	00:05:15	143	(77 : 66 [−14%])	825	(601 : 224 [−63%])
m1 4	00:03:06	91	(45 : 46 [+2%])	536	(426 : 110 [−71%])
m1 5 w/ EC	00:07:52	151	(81 : 70 [−14%])	925	(788 : 137 [−83%])
m1 6 w/ EC	00:19:00	478	(257 : 221 [−14%])	3,004	(1,999 : 1,005 [−50%])
m1 7	00:07:42	175	(97 : 78 [−20%])	1,586	(1,188 : 398 [−66%])
m1 8	00:02:43	84	(45 : 39 [−13%])	595	(518 : 77 [−85%])
m7 9 w/ EC	00:08:14	199	(114 : 85 [−25%])	1,326	(1,007 : 319 [−68%])
m7 10 w/ EC	00:07:18	151	(78 : 73 [−6%])	1,294	(849 : 445 [−48%])
m7 11	00:14:41	449	(239 : 210 [−12%])	2,916	(1,597 : 1,319 [−17%])
m7 12	00:10:32	361	(196 : 165 [−16%])	2,071	(1,165 : 906 [−22%])
m7 13 w/ EC	00:05:13	171	(94 : 77 [−18%])	870	(657 : 213 [−68%])
m7 14 w/ EC	00:05:39	193	(109 : 84 [−23%])	948	(708 : 240 [−66%])
m7 15	00:04:57	103	(61 : 42 [−31%])	834	(758 : 76 [−90%])
m7 16	00:10:01	265	(146 : 119 [−18%])	1,751	(1,238 : 513 [−59%])
Total	02:02:22	3,226	(1,756 : 1,470 [−16%])	21,088	(14,706 : 6,382 [−57%])

Table 3.4: *HCRC Edinburgh Map Task corpus data summary — Dialogues*

3.2.5 The Transcription & Annotation Data

The data of the HCRC Edinburgh Map Task corpus has already been transcribed and annotated extensively for disfluencies, gaze, moves, prosody, syntax, etc. In this study, I use the segmentation time stamps, transcribed text, and dialogue act annotations from release version 2.1.[3] As a precaution, the extracted segmentation and transcription was verified using the dedicated annotation tool ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) after converting the format for ELAN.

3.2.5.1 The Dialogue Act Labels

The dialogue act labels available in the HCRC Edinburgh Map Task corpus are based upon the Dialogue Structure Coding scheme (Carletta et al., 1997). For the sixteen dialogues of Map 01 and Map 07, a total of 3,790 items were labelled with a dialogue act. The Figure 3.3 illustrates the percentage and the number of occurrences of each dialogue act within this subset of data. Furthermore, the distribution of dialogue acts between the Information Giver and the Information Follower have also been added to the figure and details provided in Table 3.5.

[3]Data downloaded from http://groups.inf.ed.ac.uk/maptask/maptasknxt.html (last accessed on 2017/06/05)

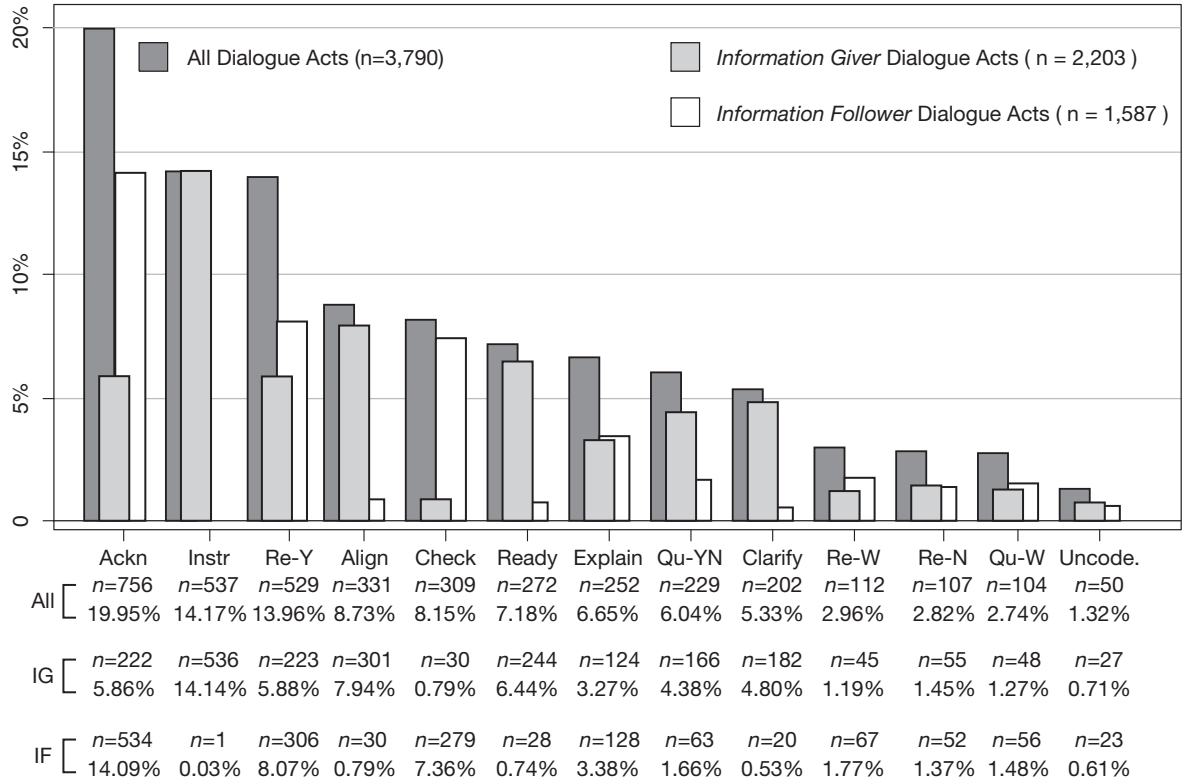


Figure 3.3: *Bar chart of dialogue act occurance within the HCRC Edinburgh Map Task corpus*

Dialogue Acts	Total		Information Giver w/ Eye-Contact				Information Giver w/o Eye-Contact			
	Count	Sub (%)	Rank	Count	All (%)	Sub (%)	Rank	Count	All (%)	Sub (%)
IG dialogue acts	2,203	58.13	—	1,042	100.00	27.49	—	1,161	100.00	30.63
Instruct	536	14.14	1	284	27.26	7.49	1	252	21.71	6.65
Align	301	7.94	2	152	14.59	4.01	3	149	12.83	3.93
Ready	244	6.44	3	115	11.04	3.03	4	129	11.11	3.40
Reply Y	223	5.88	7	72	6.91	1.90	2	151	13.01	3.98
Acknowledgement	222	5.86	4	98	9.40	2.59	5	124	10.68	3.27
Clarify	182	4.80	5	96	9.21	2.53	7	86	7.41	2.27
Query Y-N	166	4.38	6	75	7.20	1.98	6	91	7.84	2.40
Explain	124	3.27	8	63	6.05	1.66	8	61	5.25	1.61
Reply N	55	1.45	9	22	2.11	0.58	9	33	2.84	0.87
Query W	48	1.27	11	18	1.73	0.47	10	30	2.58	0.79
Reply W	45	1.19	9	22	2.11	0.58	11	23	1.98	0.61
Check	30	0.79	13	11	1.06	0.29	12	19	1.64	0.50
Uncodable	27	0.71	12	14	1.34	0.37	13	13	1.12	0.34
IF dialogue acts	1,587	41.87	—	745	100.00	19.66	—	842	100.00	22.22
Acknowledgement	534	14.09	1	275	36.91	7.26	1	259	30.76	6.83
Reply Y	306	8.07	2	157	21.07	4.14	3	149	17.70	3.93
Check	279	7.36	3	127	17.05	3.35	2	152	18.05	4.01
Explain	128	3.38	4	60	8.05	1.58	4	68	8.08	1.79
Reply W	67	1.77	5	30	4.03	0.79	6	37	4.39	0.98
Query Y-N	63	1.66	8	21	2.82	0.55	5	42	4.99	1.11
Query W	56	1.48	6	24	3.22	0.63	7	32	3.80	0.84
Reply N	52	1.37	7	22	2.95	0.58	8	30	3.56	0.79
Align	30	0.79	12	4	0.54	0.11	9	26	3.09	0.69
Ready	28	0.74	11	6	0.81	0.16	10	22	2.61	0.58
Uncodable	23	0.61	10	7	0.94	0.18	11	16	1.90	0.42
Clarify	20	0.53	9	12	1.61	0.32	12	8	0.95	0.21
Instruct	1	0.03	13	0	0.00	0.00	13	1	0.12	0.03

Table 3.5: *Summary of dialogue acts annotated in the HCRC Edinburgh Map Task corpus*

3.2.5.2 Speech Rate Labels (Trinity Dataset)

Methodological Note

For details on the reasoning to use this calculation method, refer to Chapter 5 "Speech Rate Difference".

The speech rate used on the data of the non-mediated dialogue data is the same as that used for the computer-mediated dialogue data. The formula used to indicate speech rate is as follows:

"$Speech\ Rate = 1 - ((Subject\ Utterance\ Duration)/(TTS\ Utterance\ Duration))$"

This results in a value that indicates the difference between the subject's utterance speech rate and the TTS system's speech rate that is set at 180 wpm.

■ Summary of the Speech Rate.

The speech rate difference was calculated on the ≈3,800 utterances of the non-mediated dialogue dataset using the formula listed above. A summary of the speech rate difference is listed in Table 3.6 and illustrated in Figure 3.4.

Speech Rate (%)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	−547.200	−8.940	15.650	6.059	32.690	82.250	43.73	3,790
HCRC (IG)	−438.700	−17.710	11.830	1.113	30.800	82.250	46.28	2,203
HCRC (IF)	−547.200	1.473	21.040	12.920	34.060	76.300	38.90	1,587

Table 3.6: *Summary of subject utterance duration divided by TTS output duration* — "$1 - (S/T)$"

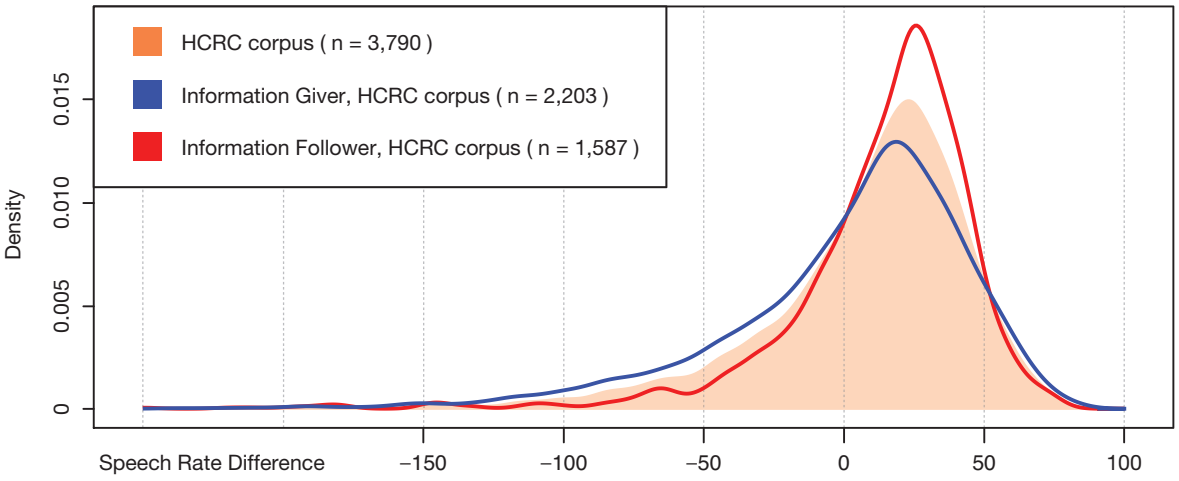


Figure 3.4: *Speech rate difference density plot of Information Giver and Information Follower dialogue data*

3.2.6 Summary of the Data

This subset of the HCRC Edinburgh Map Task corpus is used as the reference non-mediated dialogue dataset that is used to compare with the computer-mediated dialogue data explained in § 3.3. All the data except the speech rate difference mentioned in § 3.2.5.2 is downloadable from http://groups.inf.ed.ac.uk/maptask/.

3.3 The Computer-Mediated Dialogue Dataset

Methodological Note

The computer-mediated dialogue data was collected specifically to be used as data for this thesis. This section only describes a simple overview of the data. For details on the collection and creation of the computer-mediated dialogue dataset (the ILMT-s2s corpus), refer to Chapter 4, "Data Created for this Study".

3.3.1 The Subjects of the Data

The subjects (aged 18–45) of the ILMT-s2s corpus data collection were recruited from Trinity College Dublin, The University of Dublin's digital noticeboard and personal connections. Fifteen recordings of fifteen native English speakers (♀ 5, ♂ 10), and fifteen native Portuguese speakers (♀ 11, ♂ 4), were collected. The subjects are paired as illustrated in Figure 3.5.

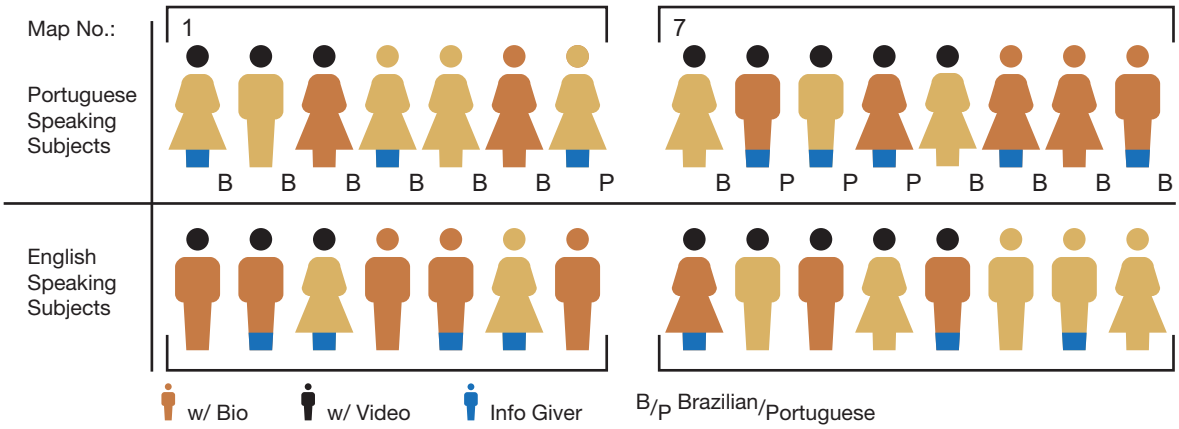


Figure 3.5: *Image of subject pairs and recording situation*

Methodological Note

Within the fifteen dialogues, five English dialogues and one Portuguese dialogue was conducted by an acquaintance of mine (Grace Cowderoy, Shane Sheehan, Kevin Doherty, and Sheila Castiho), while I too participated in two dialogues since the scheduled subject nor their backup arrived as expected.

3.3.2 The Dialogues

Methodological Note

For details about The Dialogues, refer to § 4.4.1 "Synchronisation".

The data of the computer-mediated dialogues consist of fifteen stereo recordings as summarised in Table 3.7. The fifteen recording total to 9h 39m 27s of data that has been orthographically transcribed and annotated, as explained in the following § 3.3.3 "The Transcription & Annotation of the Data".

Dialogue	Duration	*On-Talk* Utterance count			*Off-Talk* Utterance count			Total Utterance count		
(Setting, Map used)		Combined	IG	IF	Combined	IG	IF	Combined	IG	IF
01 (w/ Vid, Map 01)	00:23:02	88	55	33	31	6	25	119	61	58
02 (w/ Vid, Map 01)	00:48:57	217	100	117	104	59	45	321	159	162
03 (w/ Vid, Map 01)	01:14:20	209	106	103	197	81	116	406	187	219
04 (w/o Vid, Map 01)	00:22:31	132	59	73	79	33	46	211	92	119
05 (w/o Vid, Map 01)	00:40:20	262	201	61	2	2	—	264	203	61
06 (w/o Vid, Map 01)	00:44:07	171	93	78	144	100	44	315	193	122
07 (w/o Vid, Map 01)	00:37:55	183	94	89	11	4	7	194	98	96
08 (w/ Vid, Map 07)	00:59:29	174	121	53	128	41	87	302	162	140
09 (w/ Vid, Map 07)	00:29:13	202	98	104	29	5	24	231	103	128
10 (w/ Vid, Map 07)	00:36:07	168	76	92	130	52	78	298	128	170
11 (w/ Vid, Map 07)	00:43:17	157	101	56	75	2	73	232	103	129
12 (w/ Vid, Map 07)	00:48:19	176	100	76	66	54	12	242	154	88
13 (w/o Vid, Map 07)	00:26:13	98	49	49	108	106	2	206	155	51
14 (w/o Vid, Map 07)	00:25:59	126	69	57	69	59	10	195	128	67
15 (w/o Vid, Map 07)	00:19:38	76	34	42	16	15	1	92	49	43
Total	09:39:27	2,439	1,356	1,083	1,189	619	570	3,628	1,975	1,653

Table 3.7: *Summary of the ILMT-s2s corpus duration and utterance count*

3.3.3 The Transcription & Annotation of the Data

Methodological Note

For details about the transcription of the dialogue data, refer to § 4.4.2 "Transcription".

Two students transcribed the thirty dialogue files (fifteen each in their native language).[4] In total, 19h 18 m 54s of audio files (9h 39m 27s each of English and Portuguese dialogues) have been transcribed to create 2h 15m 41s of orthographically transcribed utterances (1h 3m 59s of English utterances and 1h 11m 41s of Portuguese utterances) as summarised in Table 3.8.

	Total	With Eye-Contact	Without Eye-Contact
No. of dialogues transcribed	15	8	7
Duration of transcribed utterances	2h 15m 41s	1h 19m 30s	56m 10s
English speaker	1h 03m 59s	38m 54s	25m 05s
Portuguese speaker	1h 11m 41s	40m 36s	31m 05s
No. of word tokens	18,678	11,101	7,577
English	9,728	5,854	3,874
Portuguese	8,950	5,247	3,703

Table 3.8: *ILMT-s2s corpus data transcription summary*

[4]Both students that transcribed and annotated the dialogue files were participants of the data collection.

3.3.3.1 Dialogue Act Labels

Methodological Note

For details about the Dialogue Act Labels, refer to § 4.4.3.2 "Annotation of Dialogue Acts [25 labels]:".

In total there are 3,850 labeled dialogue acts summarised as shown in Table 3.9 and distributed as illustrated in Figure 3.6. Since the dialogue act scheme used for this ILMT-s2s corpus are fundamentally the same as that used by the HCRC Edinburgh Map Task corpus, this annotation data can be used in a direct comparison with the dialogues of the HCRC Edinburgh Map Task corpus (Chapter 8).

Dialogue Acts	Total			With Eye-Contact			Without Eye-Contact		
	Count	All (%)	Sub (%)	Count	All (%)	Sub (%)	Count	All (%)	Sub (%)
Utterance count	3,628	—	—	2,151	—	—	1,477	—	—
Acknowledgement CP	10	0.26	0.39	1	0.04	0.07	9	0.57	0.80
Acknowledgement CPU	97	2.52	3.77	42	1.84	2.91	55	3.51	4.87
Align	206	5.35	8.00	91	3.99	6.30	115	7.34	10.18
Check	110	2.86	4.27	66	2.89	4.57	44	2.81	3.89
Clarify	146	3.79	5.67	64	2.80	4.43	82	5.23	7.26
Explain	609	15.82	23.66	428	18.75	29.64	181	11.55	16.02
Instruct	510	13.25	19.81	276	12.09	19.11	234	14.93	20.71
Query Y-N	262	6.81	10.18	142	6.22	9.83	120	7.66	10.62
Query W	287	7.45	11.15	145	6.35	10.04	142	9.06	12.57
Reply Y	137	3.56	5.32	84	3.68	5.82	53	3.38	4.69
Reply N	43	1.12	1.67	29	1.27	2.01	14	0.89	1.24
Reply W	157	4.08	6.10	76	3.33	5.26	81	5.17	7.17
Interjection	343	8.91	—	240	10.51	—	103	6.57	—
Ackn. CP (Solo)	85	2.21	9.11	57	2.50	9.52	28	1.79	8.38
Ackn. CPU (Solo)	405	10.52	43.41	258	11.30	43.07	147	9.38	44.01
Align (Solo)	60	1.56	6.43	32	1.40	5.34	28	1.79	8.38
Check (Solo)	16	0.42	1.71	3	0.13	0.50	13	0.83	3.89
Clarify (Solo)	75	1.95	8.04	33	1.45	5.51	42	2.68	12.57
Explain (Solo)	142	3.69	15.22	128	5.61	21.37	14	0.89	4.19
Instruct (Solo)	31	0.81	3.32	9	0.39	1.50	22	1.40	6.59
Query Y-N (Solo)	23	0.60	2.47	20	0.88	3.34	3	0.19	0.90
Query W (Solo)	57	1.48	6.11	27	1.18	4.51	30	1.91	8.98
Reply Y (Solo)	21	0.55	2.25	18	0.79	3.01	3	0.19	0.90
Reply N (Solo)	8	0.21	0.86	6	0.26	1.00	2	0.13	0.60
Reply W (Solo)	10	0.26	1.07	8	0.35	1.34	2	0.13	0.60

Table 3.9: *Summary of dialogue acts annotated in the ILMT-s2s corpus*

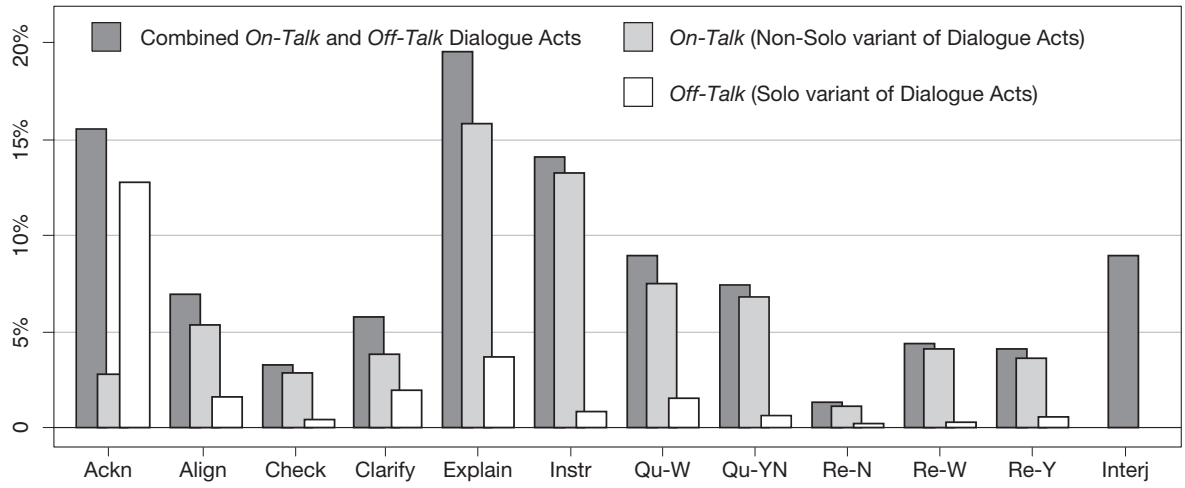


Figure 3.6: *Bar chart of dialogue acts: Total of On-Talk and Off-Talk variant's combined as well as separated — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item*

3.3.3.2 Perceived Emotion Labels

Methodological Note

For details about the Perceived Emotion Labels, refer to § 4.4.3.1 "Perceived Emotions [3 labels]:".

There are a total of 1,835 labeled perceived emotions summarised as shown in Table 3.10.

Perceived Emotions	Count	Total duration (approx.)
Amused	812	1 h 33 m
Frustrated	670	51 m
Surprised	353	17 m
Total	1,835	2 h 41 m

Table 3.10: *Summary of perceived emotions annotated in the ILMT-s2s corpus*

The annotators were allowed to freely segment the dialogue so that the label covers the area where the subject visually expressed the given state.

The distribution of the perceived emotion labels within the major groupings of the ILMT-s2s corpus are illustrated in Figure 3.7. The state of "*Surprised*" being the least frequent in all settings, as can

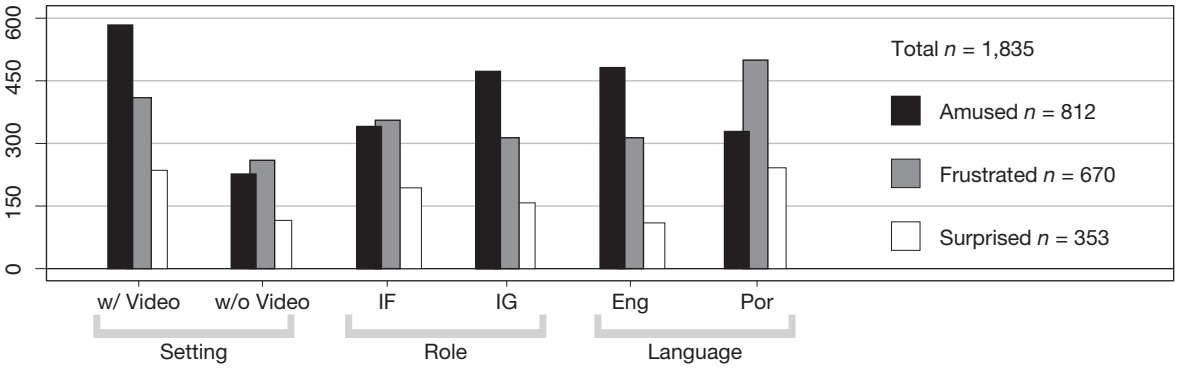


Figure 3.7: *Distribution of perceived emotions within the ILMT-s2s corpus — Categorised by group. The "n" indicates the number of labels within each grouping of "Setting", "Role", and "Language".*

be predicted from the overall low total, but interestingly the frequency of the states "*Amused*" and "*Frustrated*" interchange within each grouping.

3.3.3.3 Talk Type Labels

Methodological Note

For details about the Talk Type Labels, refer to § 4.4.3.4 "Talk Types [4 lables]:".

During the recording of the dialogues, the subjects directed their utterances at three different audiences. Themselves, directly to a fellow human, and to their interlocutor via the computer system. These three directions were labelled as *Off-Talk Self*, *Off-Talk Other* and *On-Talk* respectively. There are a total of 3,628 labeled talk types summarised as shown in Table 3.11. Since the talk type labels

are linked one to one with the utterances of the ILMT-s2s corpus, the values in Table 3.11 also refer to the number of utterances in the dataset.

Talk Type	All	English	Portuguese
On-Talk	2,439	1,328	1,111
Off-Talk	1,189	652	537
Off-Talk Self	848	483	365
Off-Talk Other	341	169	172
Total	3,628	1,980	1,648

Table 3.11: *Summary of ILMT-s2s corpus Talk Types*

3.3.3.4 Speech Rate Labels

Methodological Note

For details about the speech rate calculation method, refer to Chapter 5 "Speech Rate Difference".

The speech rate used on the data of the non-mediated dialogue data is the same as that used for the computer-mediated dialogue data. The formula used to indicate speech rate is as follows:

"$Speech\ Rate = 1 - ((Subject\ Utterance\ Duration)/(TTS\ Utterance\ Duration))$"

This results in a value that indicates the difference between the subject's utterance speech rate and the TTS system's speech rate that is set at 180 wpm.

■ Summary of the Speech Rate.

The speech rate difference was calculated on the ≈3,600 utterances of the non-mediated dialogue dataset using the formula listed above. A summary is listed in Table 3.12 and illustrated in Figure 3.8.

Speech Rate (%)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
ILMT-s2s	−1539.00	−58.76	−23.39	−37.92	1.86	83.30	74.14	3,628
ILMT-s2s (En)	−878.70	−52.43	−18.80	−28.71	7.56	83.30	58.04	1,980
ILMT-s2s (Pt)	−1539.00	−66.95	−29.12	−48.98	−4.41	58.64	88.50	1,648

Table 3.12: *Summary of subject utterance duration divided by TTS output duration* — "$1 - (S/T)$"

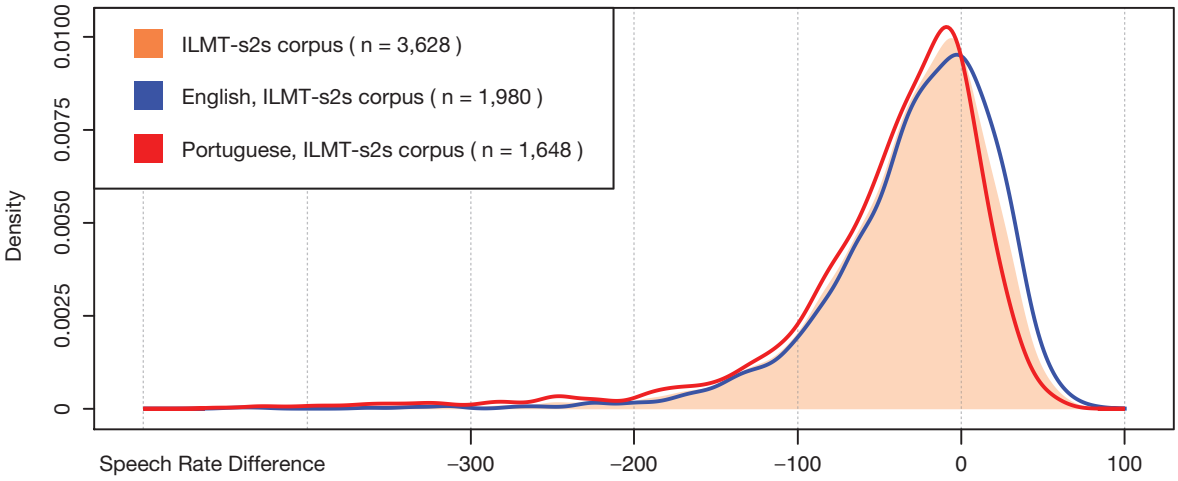


Figure 3.8: *Speech rate difference density plot of English and Portuguese dialogue data*

3.3.3.5 ASR Performance Labels

Methodological Note

For details about the ASR performance calculation method, refer to § 4.4.5 "Calculating the ASR Performance (WER & *F*-Score)".

A total of 2,370 *On-Talk*[5] utterances are labeled with an ASR performance label. The method used to calculate this value is explained in § 4.4.5. A summary of the WER and the Reverse *F*-Score are listed in Table 3.13 with the English speakers showing a higher WER variance and average.

ASR Performance		Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
WER	All utterances	0.00	0.00	0.25	0.41	0.67	10.00	0.55	2,370
	English utterances	0.00	0.00	0.29	0.43	0.75	10.00	0.60	1,302
	Portuguese utterances	0.00	0.00	0.25	0.38	0.54	8.00	0.48	1,068
Reverse *F*-Score	All utterances	0.00	0.00	0.22	0.34	0.56	1.00	0.36	2,370
	English utterances	0.00	0.00	0.25	0.37	0.67	1.00	0.38	1,302
	Portuguese utterances	0.00	0.00	0.20	0.31	0.43	1.00	0.34	1,068

Table 3.13: *Summary of computer-mediated dialogue WER & Reverse F-Score values*

The distribution of the ASR performance values for WER are illustrated as a density plot in Figure 3.9 and the *Reverse F*-Score values are illustrated as a density plot in Figure 3.10.

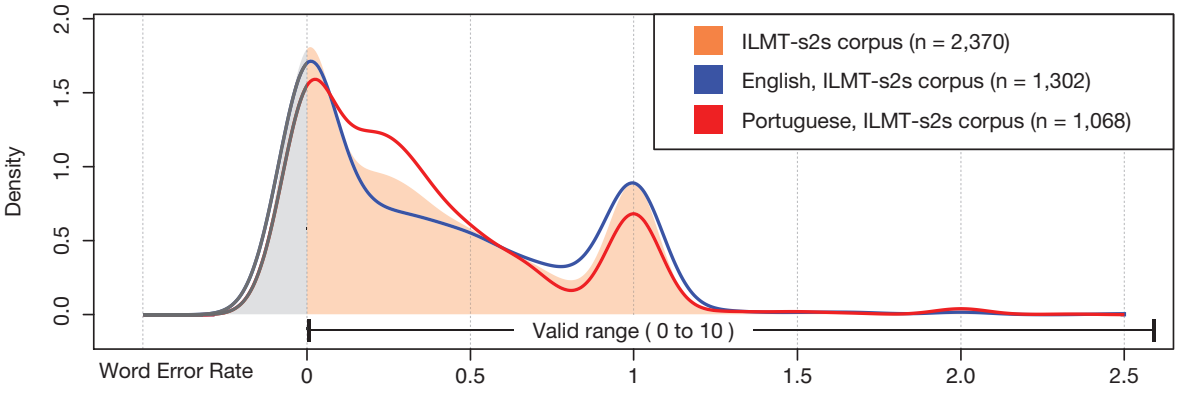


Figure 3.9: *WER: ASR performance density plot of Eng. and Portuguese dialogue data*

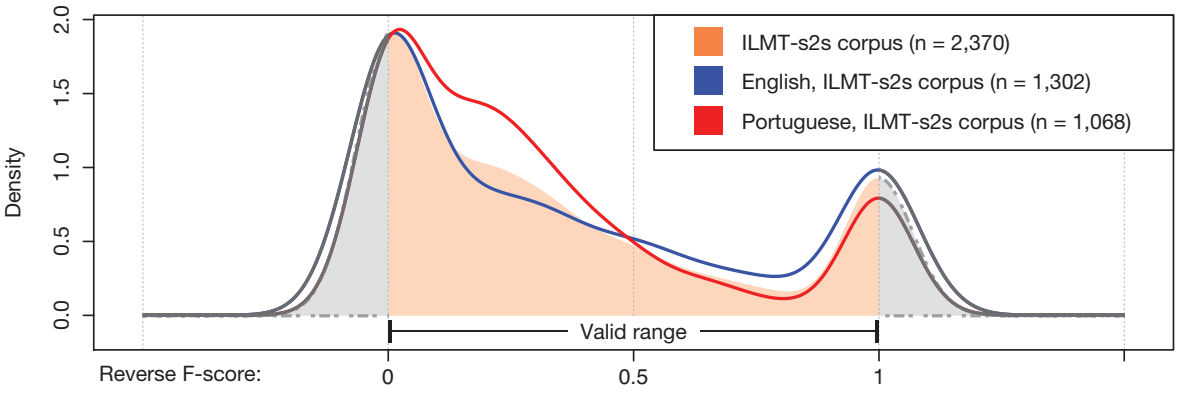


Figure 3.10: *Reverse F-Score: ASR performance density plot of Eng. and Portuguese dialogue data*

[5] "*On-Talk*" are utterances directed to the subject's interlocutor via the ILMT-s2s System as explained in § 4.4.3.4.

3.3.4 The Subject Survey

Methodological Note

For details about the subject survey, refer to § 4.3.6 "ILMT-s2s System User Survey".

A 7 point Likert scale was used to gather a qualitative assessment of the computer system used to collect the computer-mediated dialogues. The subjects of the data collection were asked to complete a survey to express their disagreement or agreement on fifteen statements and also answer five questions after each recording.[6]

3.3.4.1 The Fifteen 7 Point Likert Scale Statements

The fifteen statements used in the survey are listed in Table 3.14. Thirteen of the fifteen statements (S.01 – S.12 & S.15 of Table 3.14) were taken from "The Post-Study System Usability Questionnaire (PSSUQ)" (J. Lewis, 1993, pp. 28–33) and two (S.13 & S.14 of Table 3.14) were added specifically for characteristics of the computer-system — the system's TTS voice and the system's output.

7 point Likert scale statements	
S.01:	Overall, I am satisfied with how easy it was to use this system.
S.02:	It was simple to use this system.
S.03:	I could effectively complete the tasks using this system.
S.04:	I was able to complete the task quickly using this system.
S.05:	I was able to efficiently complete the task using this system.
S.06:	I felt comfortable using this system.
S.07:	It was easy to learn to use this system.
S.08:	I believe I could become productive quickly using this system.
S.09:	Whenever I made a mistake using the system, I could recover easily and quickly.
S.10:	The interface of the system was pleasant.
S.11:	I liked using the interface of this system.
S.12:	This system has all the functions and capabilities I expected it to have.
S.13:	I was satisfied with the voice of this system.
S.14:	I was satisfied with the output of this system.
S.15:	Overall, I am satisfied with this system.

Table 3.14: *List of survey statements about the ILMT-s2s System*

3.3.4.2 The replies to the 7 Point Likert Scale Statements

The thirty replies to the fifteen statements (450 replies) of the subject survey are listed in Table A.14 of Appendix A. The stacked bar charts and density plots of the subject survey results is illustrated in Figure 3.11. The overview of the thirty replies to the subject survey indicate that all statements return a median value of a neutral four or better than neutral as indicated in the Table 3.15.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	4.0	4.5	6.0	6.0	5.0	6.0	4.5	4.0

Table 3.15: *Median value per statement of subject survey Likert results*

[6] Only the Likert scale results of the fifteen statements are analysed in this study.

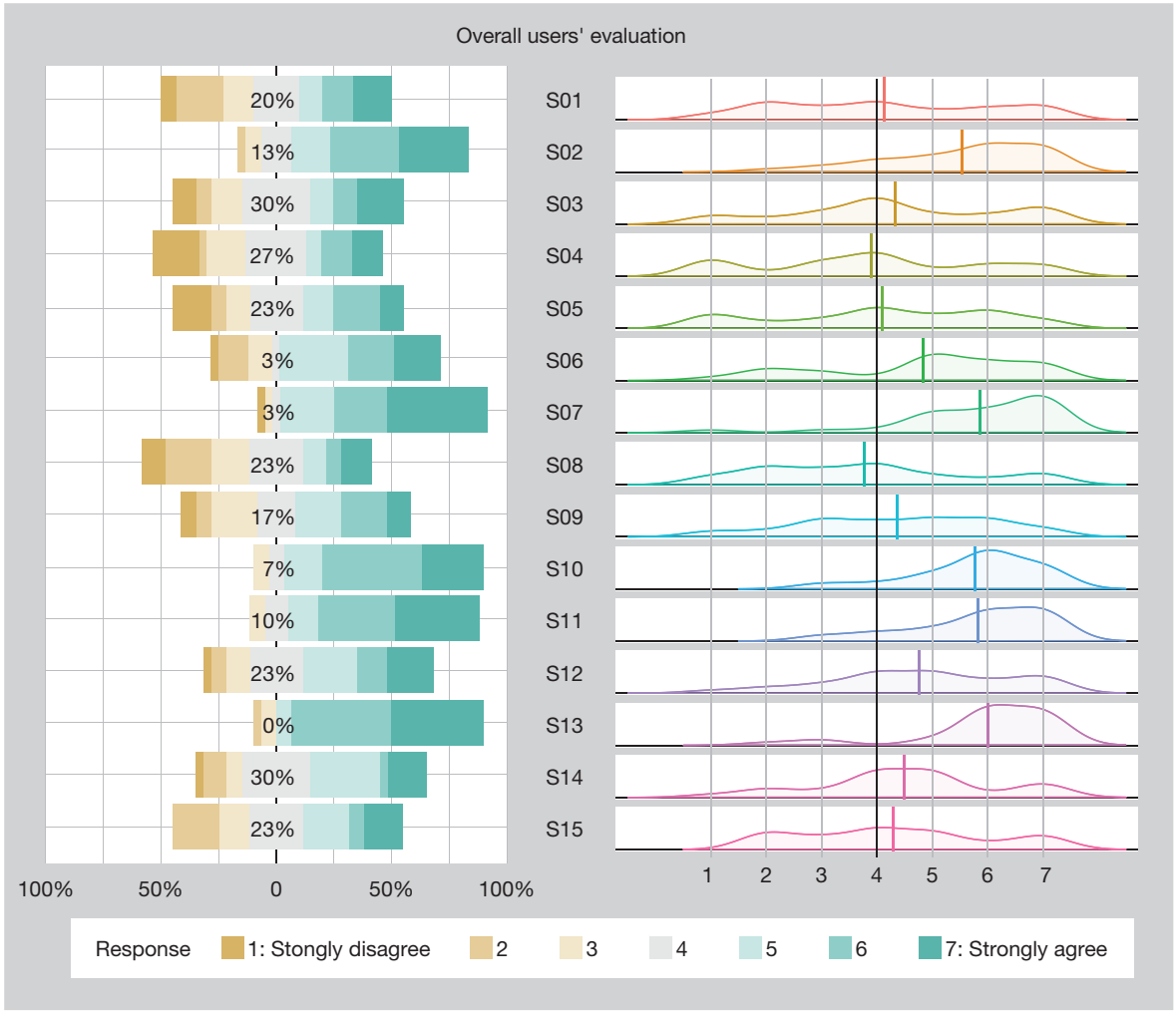


Figure 3.11: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale replies*

3.3.5 Summary of the Data

The computer-mediated dialogue data is used as the core of this study. The data within this section is used in the following chapters;

- Chapter 5 "Speech Rate Difference":

  The reason why the speech rate difference of § 3.3.3.4 is used instead of words-per-minute or syllables-per-second.

- Chapter 6 "The Three Talk Types":

  How the speech rate of § 3.3.3.4 differs when the subject is talking to themselves, the computer system and directly to a fellow human.

- Chapter 7 "Speech Rate Differences Between Datasets":

  A comparison between the speech rate of the computer-mediated dialogues of § 3.3.3.4 and the speech rate of the non-mediated dialogues of § 3.2.5.2.

- Chapter 8 "Dialogue Acts Differences Between Datasets":

  A comparison between the dialogue acts of the computer-mediated dialogues of § 3.3.3.1 and the

dialogue acts of the non-mediated dialogues of § 3.2.5.1.

- Chapter 9 "ASR Performance & Speech Rate":
  How the ASR performance of § 3.3.3.5 affects the speech rate difference of § 3.3.3.4.

- Chapter 10 "ASR Performance & Perceived Emotions Analysis":
  How the ASR performance of § 3.3.3.5 affects the perceived emotion of the subject of § 3.3.3.2.

- Chapter 11 "User Survey Study":
  Using the replies to the fifteen 7 point Likert scale statements of § 3.3.4 , a qualitative study of the computer-mediated dialogues is reported to identify how the subjects interacted with the computer system and how it may have affected the subjects of the computer-mediated dialogues' data collection.

Methodological Note

As mentioned earlier on page 27, the computer-mediated dialogue data (the ILMT-s2s corpus) used as the dataset of this study and explained in this section was collected for the purpose of this study. The details of the whole dataset and the method of collecting and creating the data is explained in further detail in Chapter 4, "Data Created for this Study".

3.4 Summary

The data from two datasets is used in this thesis — a subset of the HCRC Edinburgh Map Task corpus and a subset of the ILMT-s2s corpus. The core of this study rotates around the computer-mediated dialogue data from the ILMT-s2s corpus with the non-mediated dialogue data from the HCRC Edinburgh Map Task corpus used as a gold standard reference of task-based dialogue speech. In other words, the non-mediated dialogue data is used to identify differences and similarities in the speech rate (Chapter 7), utterances word count and utterance intention (as in dialogue acts in Chapter 8), with the computer-mediated dialogues as a means to identify unique characteristics of the computer-mediated task-based communication (Chapter 6, Chapter 9, Chapter 10, and Chapter 11).

Chapter 4

Data Created for this Study

Methodological Note

This chapter contains repetition from the previous Chapter 3 "The Datasets", as this chapter explains the collection of the dataset explained in § 3.3 "The Computer-Mediated Dialogue Dataset".

The data for the ILMT-s2s corpus was collected at Trinity College Dublin, The University of Dublin, between September 2014 and December 2014, by myself and Dr. Loredana Cerrato, with the guidance of Dr. Saturnino Luz.

4.1 Introduction

The corpus[1] presented in this chapter was designed to support the examination of how participants adapt to interlingual language-technology mediated task oriented communication and to provide a starting point for future data collections.

Fifteen dialogues between English and Brazilian/Portuguese subjects talking to each other cross-lingually but each in their native language were collected with the map task (Brown, Anderson, Shillcock, & Yule, 1985; Anderson et al., 1991) elicitation technique in a setting where the interaction is mediated by a Speech-to-Speech Machine Translation (S2S-MT) system. The system was implemented using off-the-shelf Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech (TTS) synthesis technologies to enable speakers of different languages to communicate with each other, remotely over the network, each in their native language. This prototype Speech-to-Speech Machine Translation system (ILMT-s2s System) adds three elements of ASR, MT and TTS synthesis into the communication that presumably creates a different communication style to the Human-to-Human communication style that the subjects are well adapted to (Newlands et al., 2003).

In this corpus a variety of synchronised and finely time-stamped data streams, including high

[1]The ILMT-s2s corpus data collection obtained the approval of the Trinity College Dublin, The University of Dublin, School of Computer Science and Statistics ethics committee in August 2014, under ethical approval No. 133/14.

quality audio of the subjects' utterances, video and ASR, MT, TTS events have been collected. In addition, using biosignal recording devices and eye tracking glasses, the Heart Rate (HR), Skin Conductance (SC), Blood Volume Pressure (BVP), Electroencephelography (EEG) signals, and the eye gaze of one subject in each dialogue has also been collected.

This ILMT-s2s corpus is a new source of data for analysing the different repair strategies and cooperative behaviour of the subjects, but since it uses the same maps to elicit the dialogue as the direct Human-to-Human (H2H) communication of the HCRC Edinburgh Map Task corpus (Anderson et al., 1991), it can also be used to compare aspects of adaptation to the S2S-MT technology.

Methodological Note

The ILMT-s2s corpus was not only collected for this thesis work. Biosignals, such as Skin Conductance (SC), Heart Rate (HR) using the Blood-Volume Pulse (BVP) readings, and the brains' electrical activity through Electroencephelography (EEG) where also collected,[a] as were Eye-tracking data that recorded the subject's eyes and pupil movement.

[a]Publications related to the collected biosignal data can be found in the following: INTERSPEECH'15 (Hayakawa, Haider, et al., 2015), Speech Prosody 2016 (Hayakawa, Haider, et al., 2016), ICASSP 2018 (Haider, Hayakawa, Luz, Vogel, & Campbell, 2018), the journal Speech Communication (Haider, Hayakawa, Vogel, Campbell, & Luz, 2019) and the PhD thesis of Fasih Haider (Haider, 2018).

4.1.1 Main Design Concept of the Corpus Collection

The main design concept of the corpus was based upon part of the initial CNGL II[2] showcase scenario, where a potential customer calls a manufacturer's call centre for advice on a product. The agent in the call-centre has the information about a product, situation, procedure or task. The customer calling the call-centre agent wishes to acquire that information from the agent. As humans, we have a skill of understanding information, but a computer doesn't. The scenario was to provide the information to a human wherever they were in the world. The design concept of ILMT-s2s corpus was to present a scenario of one subject with most of the information and the other without equal information, attempting a shared task where success depends on pooling this information while using the computer mediated interlingual Speech-to-Speech system and understand the difficulties that the human user encounters during the process of the communication.

4.1.2 Initial Objectives of the Corpus

Following the main concept of the corpus collection, where one subject with the information guides another without the same amount of information, the two initial objectives of collecting this corpus were:

(i) To see if it is possible to detect the occurrence of surprise and frustration in the interactions of the subjects of a interlingual language-technology mediated task (Picard, Vyzas, & Healey,

[2]CNGL II was a research consortium of academia and industry that is mainly funded by the Science Foundation Ireland.

2001). The electroencephelography (EEG) signal data was selected to detect the N400 effect that is linked to neurological activities of meaning processing (van Berkum, Hagoort, & Brown, 1999; Kutas & Federmeier, 2011; O'Sullivan et al., 2015) and the other signals of skin conductance (SC) (Picard et al., 2001; Takahashi, 2004), heart rate (HR) or blood volume pressure (BVP) (Takahashi, 2004; Rainville, Bechara, Naqvi, & Damasio, 2006), and pupil size from the eye gaze recorded by eye tracking glasses (Kahneman, 1973; Beatty, 1982; Just, Carpenter, & Miyake, 2003; Palinko, Kun, Shyrokov, & Heeman, 2010) were expected to provide robust backing to the findings from the EEG signal data with regard to the cognitive load of the subjects.

(ii) To compare interlingual language-technology mediated communication with direct monolingual Human-to-Human (H2H) communication to see if adaptation methods of H2H (Oviatt et al., 1998) and Human-to-Computer (H2C) (L. Bell et al., 2003; Suzuki & Katagiri, 2007; Schneider & Luz, 2011) persist in interlingual language-technology mediated task oriented communication or differ from direct H2H task oriented communication.

As mentioned under the "Methodological Note" on page 44, research into the biosignals was not conducted by myself and is therefore not reported in this thesis. With regards to the biosignals, I collected the data and for this reason it is mentioned in this thesis. For details about the research of the biosignals collected in this corpus, please refer to the thesis and publications of my colleague Fasih Haider. The data used in the studies reported in this thesis are related to objective '(ii)' (mentioned above) and used the metadata (utterance timing segmentation, perceived emotion annotation, dialogue act annotation, ASR performance values, and the talk types of the subject) of this corpus.

4.1.3 Structure of this Chapter

The remainder of this chapter is structured as follows:

- Section 4.2 "The Communication Tool" describes the design of the interlingual computer system used to mediate the communication between the interlocutors.
- Section 4.3 "The Data Collection" describes how the data was collected.
- Section 4.4 "Processing the Data" describes how the separate data files are synchronised, transcribed and annotated.
- Section 4.5 "Summary" summarises the ILMT-s2s corpus.

4.2 The Communication Tool

This section explains the general design and the components of the ILMT-s2s System.

The ILMT-s2s System is a 700 line Python script[3] that combines three off-the-shelf technologies in a clean simple User Interface (UI) as illustrated in Figure 4.1. The ILMT-s2s System runs on both Mac OS X and Linux[4] with the four main components being:

1. "Click-and-Hold-to-Talk" for the recognition of system directed utterance (*On-Talk*);
2. The Google Speech API for Automatic Speech Recognition (ASR);
3. The Microsoft Bing translation system for Machine Translation (MT);
4. For Text-to-Speech (TTS):

   The Apple TTS system voices provided with Mac OS X computers;[5]

   The eSpeak speech synthesiser for Linux computers.

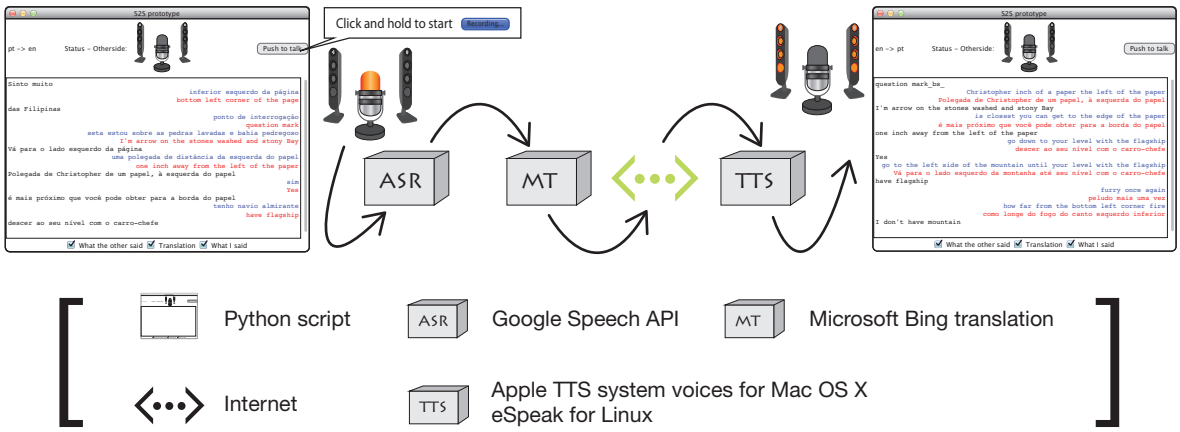


Figure 4.1: *ILMT-s2s System used to collect the data*

The ILMT-s2s System takes the speech uttered to the system (*On-Talk*) and provides Speech-to-Text (S2T/ASR)[6] to the producer, and Speech-to-Text Machine Translation (S2T-MT)[7] and Speech-to-Speech Machine Translation (S2S-MT) to the receiver situated in a different room.

4.2.1 Basic Structure of the System

The ILMT-s2s System is activated by a "Push to talk" button that the subject is required to click-and-hold for the duration of the utterance, and releases once the subject has finished the utterance. For each utterance the ILMT-s2s System executes the following processes and logs the five actions "Click-and-hold", "ASR", "MT", "TTS send", and "TTS received", into an XML file.

[3]My contribution to the ILMT-s2s System was the initial C script version of this system that didn't include a UI. As for the Python script, my contribution is limited to the final layout of the UI, the commands and keys to obtain the ASR, MT and TTS outputs, the selection of the voices to be used, the listing of the languages to be used, and making it compatible for use on an Apple computer. The communication between two computers, the creation of the initial UI, the time-stamping to XML and the initial Python script were created by Dr. Saturnino Luz.

[4]For the data collection, only Mac OS X computers were used.

[5]For the data collection, the voices of Kate (English), Joana (Portuguese) and Luciana (Brazilian) were used.

[6]For the data collection, the S2T result was not displayed to the receiver, but only to the producer.

[7]For the data collection, the S2T-MT result was not displayed to the producer, but only to the receiver.

1. Start recording audio at 96 kHz, 24 bit PCM format with the *sox*[8] command when the "Push to talk" button is first pressed.

2. Stop recording the audio when the "Push to talk" button is released.

3. Convert the audio file recorded in step 1 – 2 to a Google Speech API compatible 16 kHz, 8 bit FLAC format using the *sox* command.[9]

4. Send the FLAC format audio file created in step 3 to the Google Speech API via the internet with the *wget* command and receive the ASR result.[10]

5. Send the text result of the ASR (Google Speech API) to the Microsoft Bing translation system via the internet with the *wget* command and receive the MT result.[11]

6. Send the text result of the Microsoft Bing translation system to the interlocutor's computer via the internet.

7. The TTS component on the interlocutor's computer converts the text to speech with the TTS command and outputs the translated text as synthesised speech in the target language.

4.2.2 The Recognition Method

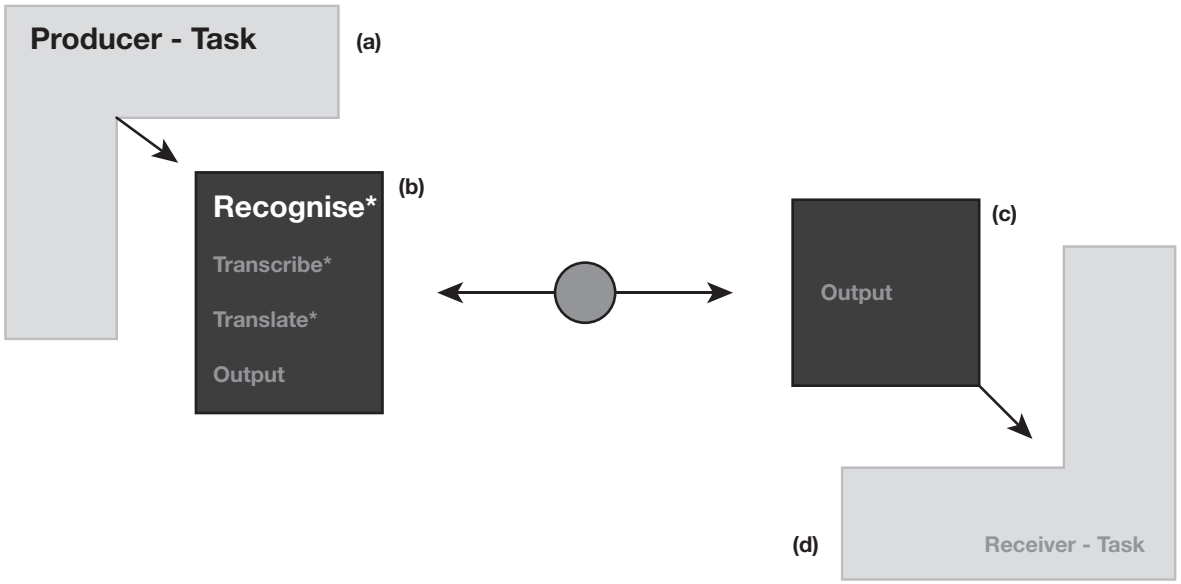


Figure 4.2: *Turn flow diagram when using the ILMT-s2s System. Items (a & b) related to this section are highlighted.*

[8]Sound eXchange (SoX): A command line based sound processing utility programme: http://sox.sourceforge.net (Accessed on 2017/11/16)

[9]The specification of 16 kHz, 8 bit FLAC format was a requirement of the Google Speech API at the time of the data collection (Autumn 2014).

[10]Refer to the portion between "<raw_output>" and "</raw_output>" in Listing (ii) of § 4.2.6 for the exact result received from the Google Speech API, and the portion between "<clean_output>" and "</clean_output>" for the result that was used by the ILMT-s2s System.

[11]Refer to the portion between "<raw_output>" and "</raw_output>" in Listing (iii) of § 4.2.6 for the exact result received from the Microsoft Bing translation system, and the portion between "<clean_output>" and "</clean_output>" for the result that was used by the ILMT-s2s System.

The communication tool (the ILMT-s2s System) uses the "Click-and-Hold-to-Talk" method to recognise that the utterance of the producer is a utterance to be transmitted to the receiver.

A problem that occurs with uncontrolled dialogues is the identification of utterances directed to the system. In other words, where exactly is the start and end of each utterance and also, which utterances are to be recognised by the system and processed for the interlocutor. Three interaction methods were thought of while planning the ILMT-s2s System;

(i) "Always-Listening"

(ii) "Click-and-Talk"

(iii) "Click-and-Hold-to-Talk".

Popular S2S voice command systems use a wake-word, such as Siri of Apple with "Hey Siri", Google voice command with "Ok Google", and Alexa of Amazon with "Alexa", "Echo", "Amazon", "Computer", to indicate to the "Always-Listening" S2S system that the subject would like to direct an utterance to the system. From the wake-word the system would determine the completion of the utterance and act upon the command depending on the structure of the underlying dialogue management system. Since the objective of the study was to study the subject's adaptation to a interlingual computer mediated communication, the ILMT-s2s System did not include a dialogue management system so that the system dialogue would not be interpreted as another interlocutor interfering with the "Human-to-Human" dialogue in the data collection.

The "Click-to-Talk" method would eliminate the necessity of the wake-word, but completion of the utterance would still need to be determined. Limiting the utterance end identification to the audio signal, fixed duration and pause duration were considered, but as previous systems and pilot test-runs indicated, this method is difficult to optimise. The UI design used in the DARPA funded DIPLOMAT project (Frederking, Rudnicky, Hogan, & Lenzo, 2000) that underlined the Tongues Speech-to-Speech Machine Translation System (Black et al., 2002), used a "Listen" and "Stop Listen" button to segment the part that the system was to process. A publication for VoiceTra (Matsuda et al., 2013, p. 230) used the mobility of the system being installed onto a mobile phone to their advantage by using the movement of bringing the mobile phone up to the subject's ear to indicate the start of the listening period and the movement of the mobile phone being placed down and away from the ear to stop the listening period.[12] Unfortunately neither of these methods were appropriate for the ILMT-s2s System. The design used in the DIPLOMAT project was designed to be operated by military personnel, so the operator of the system could be trained not to forget the secondary step of clicking the button for "Stop Listen". Furthermore, since the ILMT-s2s System was installed into a desktop computer, the method published for VoiceTra was not feasible. Therefore, I implemented the "Click-and Hold-to-Talk" method.

[12]The latest version (Ver. 6.0) of the VoiceTra iOS application is using the "Click-and-Hold-to-Talk" method which is the same as the method chosen for the ILMT-s2s System (verified on 2018/01/07).

4.2.2.1 After Note

Though objection was raised by other members of the faculty, the "Click-and-Hold-to-Talk" was selected for the ILMT-s2s System to provide the user of the system the capability to determine when their utterance to the system started and ended and not let the system control or interfere with the dialogue. The objection is rooted to the fact that during direct Human-to-Human communication, there is no need to press an item for the duration of one's utterance (turn) to indicate that one is talking to the interlocutor, and that a button being pressed for the duration of the utterances would be intrusive to the act of communicating, unnatural and counterintuitive.

This criticism of the "Click-and-Hold-to-Talk" does not take into account the different communication style of task based dialogues or S2S-MT, and also the lack of control the user feels when adapting to the usage of ASR. The expectation that the subjects with the role of Information Giver will need more time to articulate an instruction that they think is adequate will create a slower utterance with longer pauses and more frequent restructuring of phrases. In addition, the new modeling the subject will be adjusting to is not common nature to the subject as mentioned by Branigan et al. (2010, p. 2366) where "[speakers] are unlikely to have a good model of what computers are likely to know, presumably in part because natural language systems tend to be tailored to particular and quite specific uses". Adding a third learning curve of creating phrases for a limited time frame (a system listening cut-off-time) is a cognitive load that I did not feel easy about adding to the system. Especially after taking into account that CB radio transceiver and walkie-talkies still use this "Click-and-Hold-to-Talk" method of communication.

Methodological Note

With regards to this activation method, four of the thirty subjects of the ILMT-s2s corpus collection brought up the "Click-and-Hold-to-Talk" button in the user survey (further details of this is provided later in this chapter; see §11.3, "Methods of the Analysis") even though it was not included as a question. Of the four comments, one was a negative opinion of the method — "*This system is very useful, but a bit slow to process, I also didn't like that you have to push the button to talk, it would be easier just talking and that is it, then you can talk while doing something else*". The other three subjects commented positively about the method by indicating the ease of just pressing a button and speaking — "*Simple to click a button and speak*", "*I think the sistem was easy because its only press a button and speak*", and "*Easy to click the button - very straightforward to learn and explain to others*".

4.2.3 The Transcription and Translation Method

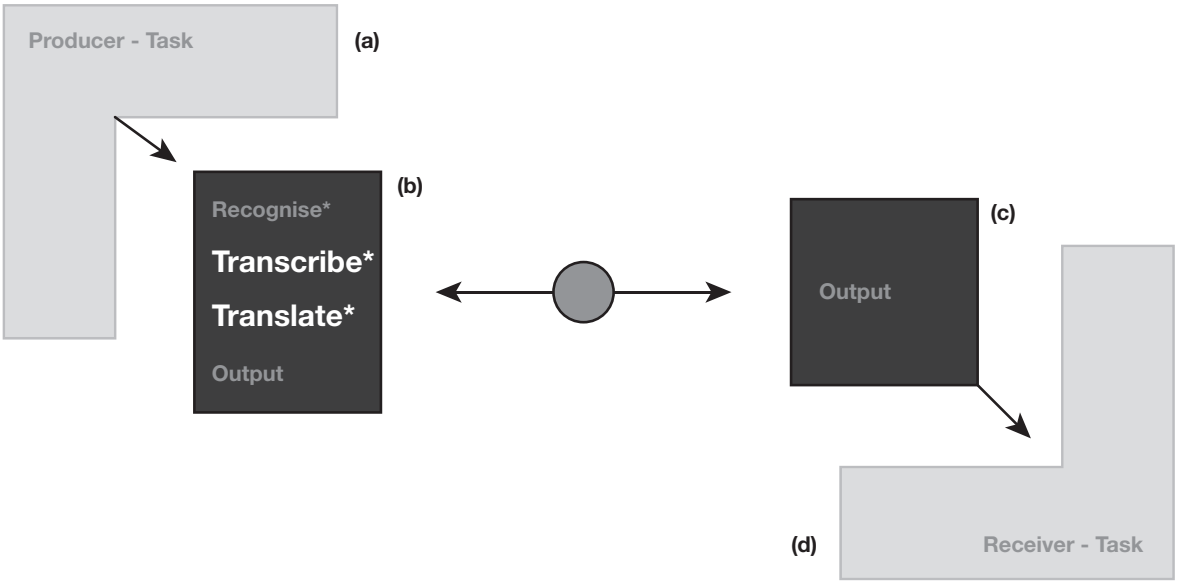


Figure 4.3: *Turn flow diagram when using the ILMT-s2s System. Items (b) related to this section are highlighted.*

The communication tool (the ILMT-s2s System) used off-the-shelf technologies to provide the Speech-to-Text (S2T/ASR) and the Machine Translation (MT) components of the system. The ILMT-s2s System used the Google Speech API for S2T/ASR and the Microsoft Bing translation API for the MT of the text from the source language to the target language. Since the ILMT-s2s System used off-the-shelf technologies, the system was capable of using any language available by the technologies and was setup for the following languages: (i) English, (ii) German, (iii) French, (iv) Spanish, (v) Portuguese, and (vi) Japanese.

Depending on the source and target language setup of the ILMT-s2s System, the system would;

(i) **Transcribe:**

Send the recognised utterance audio file to the S2T/ASR API and retrieve the transcription of the audio file, as specified as the source language.

(ii) **Translate:**

Send the transcribed text to the MT API and retrieve the translation of the text, as specified as the target language.

4.2.4 The Output Method

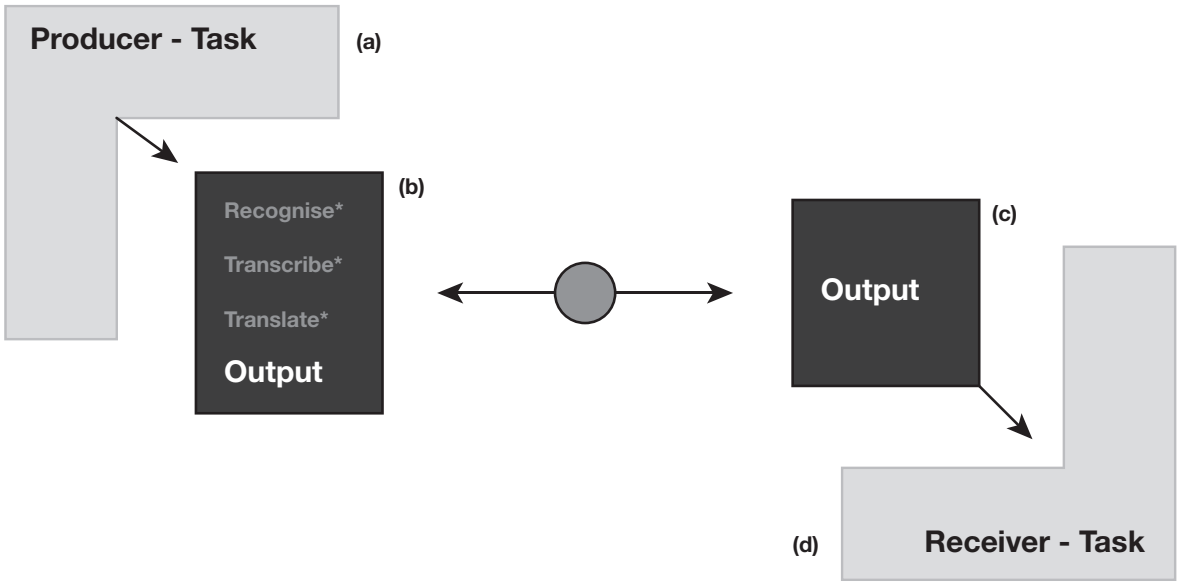


Figure 4.4: *Turn flow diagram when using the ILMT-s2s System. Items (a, b, c & d) related to this section are highlighted.*

The communication tool (the ILMT-s2s System) uses two means of output; (i) verbal, and (ii) textual. The verbal output of the ILMT-s2s System was used to output the translated utterance of the producer on the receiver's system. The textual output, on the other hand, was used for output on both the producer's and receiver's system.

As with the Transcription and Translation processes of the system (§ 4.2.3), the verbal output, Text-to-Speech (TTS), used off-the-shelf technologies to output the translated text.

- **Producer's system:** Display the S2T/ASR result utterance text.
- **Receiver's system:** Verbal output of MT result & Display the MT result utterance text.

Since a number of past S2S-MT systems such as VERBMOBIL (Nöth et al., 2000), JANUS III (Levin, Lavie, et al., 2000), the Tongues Speech-to-Speech Machine Translation System (Frederking et al., 2002), NESPOLE! (Lavie et al., 2006), the 2006 BBN English/Iraqi Two-way Speech Translation System (Stallard et al., 2006), and VoiceTra (Matsuda et al., 2013) generally included a text field to display the following:

(i) ASR results

(ii) MT results

(iii) MT back-translation results.

The inclusion of these three information channels was initially deemed mandatory, but as explained in the following § 4.2.4.1, "Concerning the Back-Translation Channel", the MT back-translation display was not included in the system. The reason for the system to display the two communication items of "ASR" results and "MT" results are as follows:

(i) The ASR results would provide the producer with a visible output of the system's recognition result (Lavie, Metze, & Pianesi, 2002), and apply as an ASR error indicator by enabling the producer to compare what they said with the textual display of the ASR. The visualisation of the ASR result also applies as a backup memory for the producer (Oviatt, Coulston, & Lunsford, 2004), by providing a textual indication of the previous utterances.

(ii) The MT results that is displayed on the receiver's computer provides the receiver with a textual representation of the verbal output and also applies as a backup memory feature as mentioned above in (i).

4.2.4.1 Concerning the Back-Translation Channel

As mentioned in § 4.2.4, back-translation is a common feature in S2S-MT systems, but it was not added to the ILMT-s2s System. The back-translation has been considered as a method to determine the quality of human translation (Brislin, 1970) and machine translation (Flournoy & Callison-Burch, 2001) for some time now. It has also been used within a S2S-MT system, where concerning the MT back-translation results Lavie et al. (2006, p. 1667) mentioned that "[by] monitoring the "System Understands" field, the user can verify that the meaning of the utterance was correctly captured by the analyzer within the translation server (by judging whether the paraphrase back into their own language reflects the same meaning as the originally spoken utterance)".

However, it must be noted that the NESPOLE! system converted the subject's speech utterance ASR results using an Interchange Format (Levin, Gates, Lavie, & Waibel, 1998; Levin, Gates, et al., 2000; Levin et al., 2002) and outputs the result of the utterance's Interchange Format into the target language. Therefore, this feedback field was not a verification field of the translation, but a verification of the system's understanding of the subject's utterance — the decoding result of the Interchange Format that is based on the ASR component input.[13] Then it is questionable as to why the experiment was designed to use two feedback fields, the subject's utterance ASR result and the back-translation of the interlingua, and not only the interlingua "System Understands" field, since the "System Understands" field is of sole importance to the translation output. Furthermore, Frederking et al. (2002, p. 162) raise a concern about the back-translation feedback after observing a phenomena in a field study of an S2S-MT system where the feedback is hindering progress: "[a] bilingual team member in fact observed that often an English-to-Croatian translation that was basically correct would be rejected by an English-speaker because the back-translation was seriously garbled".

This relates to the observation by Shigenobu, (2007) that the quality of the translation should be equal in both directions for the back-translation to be of reliable usability. Also, this feedback channel does not provide the subject with information on how to modify the erroneous utterances and as demonstrated in a field study by Frederking et al. (2002, p. 163), "if speech recognition produced

[13] The translation quality difference between the target language and the source languages from the interlingua does not seem to be in question in their study.

and displayed a correct transcript, and then the translation produced an unacceptable result, they would usually *respeak* the same utterance using the same words !" It is true that Yamashita & Ishida (2006a) presented a method of comparing the ASR result with the back-translation to indicate where the error of the translated text may be, and highlighted the estimated cause of error, but adding this feature into the ILMT-s2s System would be adding an addition filter to the current "ASR", "MT" and "TTS" which would likely deviate the subject's attention by adding another item for the subject to be concerned about — the concern, "is the translation correct".

My understanding is that the back-translation was not necessary in the ILMT-s2s System for the following reasons:

(a) it was not possible to guarantee the translation quality of both directions to be of equal (source to target, and target to source) during the data collection

(b) the back-translation does not provide information of what was causing the erroneous translation and therefore would not provide adequate feedback to the subjects

(c) the back-translation would provide an additional adaptation trigger that I do not want to consider

(d) in a real-life scenario it is very unlikely that one has two interpreters to verify the communication quality — the quality of the translation is verified from the interlocutor's reply.

For these reasons, the back-translation channel was not added to the ILMT-s2s System.

4.2.5 The User Interface

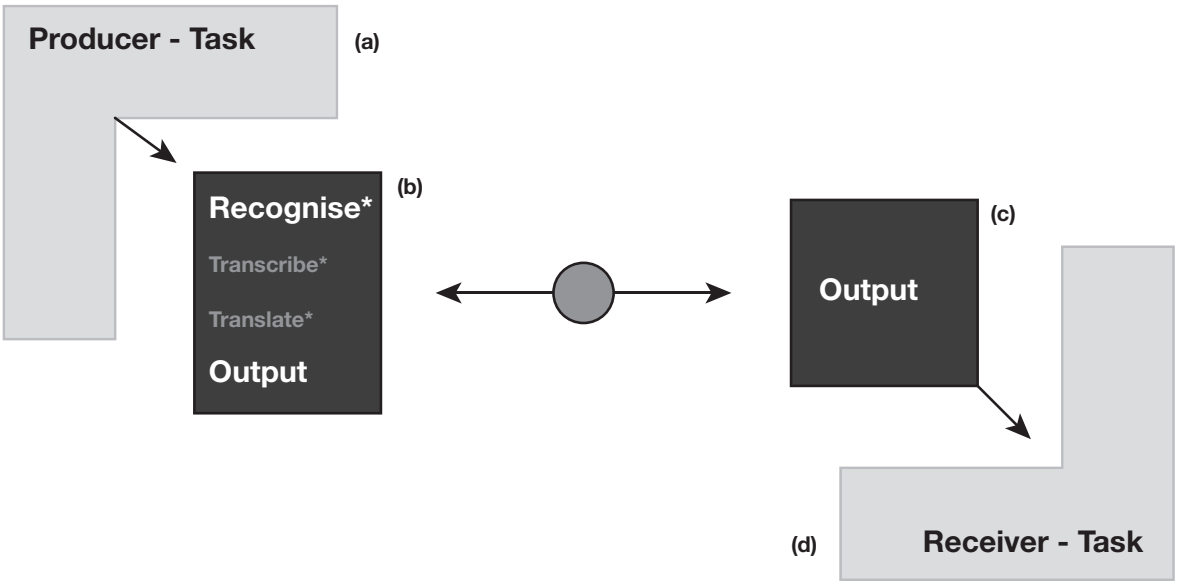


Figure 4.5: *Turn flow diagram when using the ILMT-s2s System. Items (a, b, c & d) related to this section are highlighted.*

The principle of the User Interface (UI) design is to be obvious as "Well-designed objects are easy to interpret and understand" (Norman, 2002, p. 2), and also to inform the subject what the system is doing by providing feedback of the current status; "Feedback and communication encompass far more than merely displaying alerts when something goes wrong. Instead, it involves keeping subjects informed about what's happening by providing appropriate feedback and enabling communication with your app" (Apple, 2013, p. 31). With this in mind, the UI was specifically designed to look minimalistic so that emphasis could be placed on the TTS output and the displayed text. Feedback indicators were also added to provide the subject information about the actions of the system and the interlocutor as illustrated in Figure 4.6.

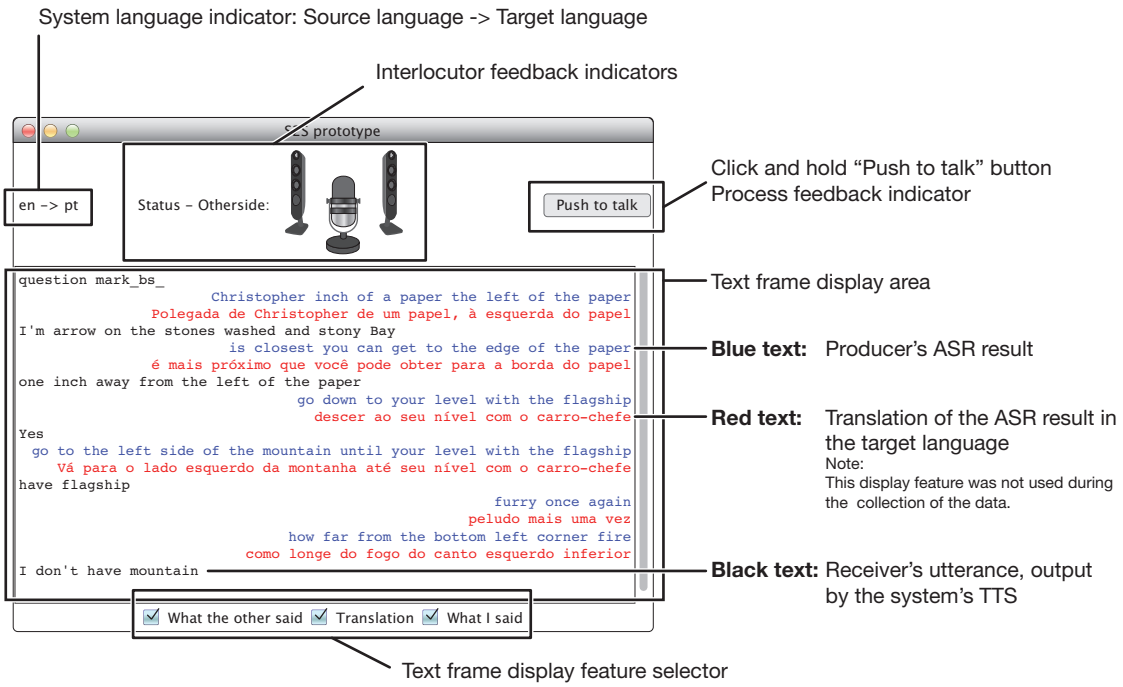


Figure 4.6: *User Interface of the ILMT-s2s System*

4.2.5.1 Dialogue Display

The design of the textual display is taken from the basic structure of popular texting software. This is based on the assumption that due to the popularity of texting applications on smartphones and computers, the subjects would be familiar with the right-side displaying the producer's text and the left-side displaying the utterances from the receiver. The text is timed to be displayed after all the system's internal processes are completed. For the producer this means that the ASR and MT (indicated as "Blue text" and "Red text" in Figure 4.6) are displayed immediately after the MT is sent to the receiver's computer. For the receiver, the producer's translated utterance (indicated as "Black text" in Figure 4.6) is displayed on the UI and output by the TTS as soon as it is received.

4.2.5.2 Process Display

As for the feedback indicators, two systems were included in the UI.

(i) An indicator of the process being performed by the ILMT-s2s System in the click-and-hold "Push to talk" button (Figure 4.7).

(ii) An indicator that displays the *Recognising* or *Outputting* action of the interlocutor's system at the top-centre of the UI (Figure 4.8).

System Process Indicator

Text within the click and hold "Push to talk" button was used to display the current process that was being performed by the ILMT-s2s System as explained in Figure 4.7.

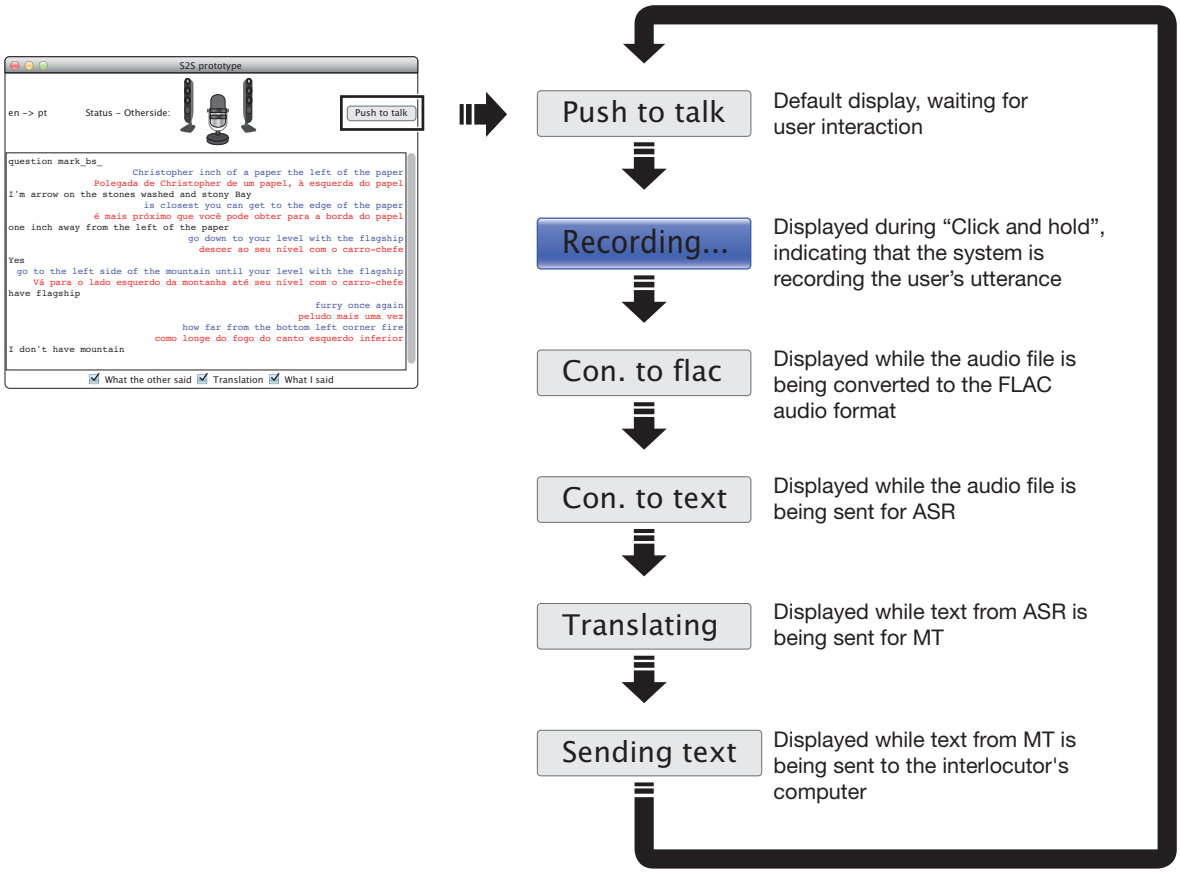


Figure 4.7: *System process indicator*

Interlocutor action indicator

The microphone and loud speakers placed at the top-centre of the UI illustrated the action of the interlocutor as explained in Figure 4.8.

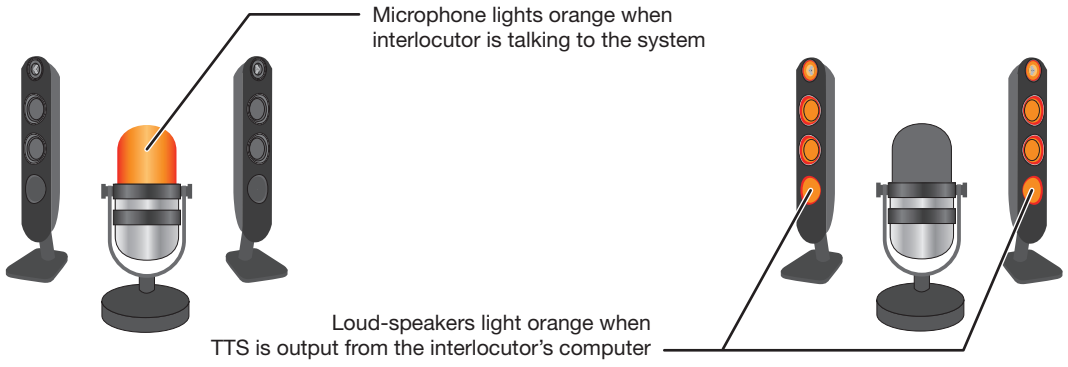


Figure 4.8: *Interlocutor action indicators*

4.2.6 System Log (XML)

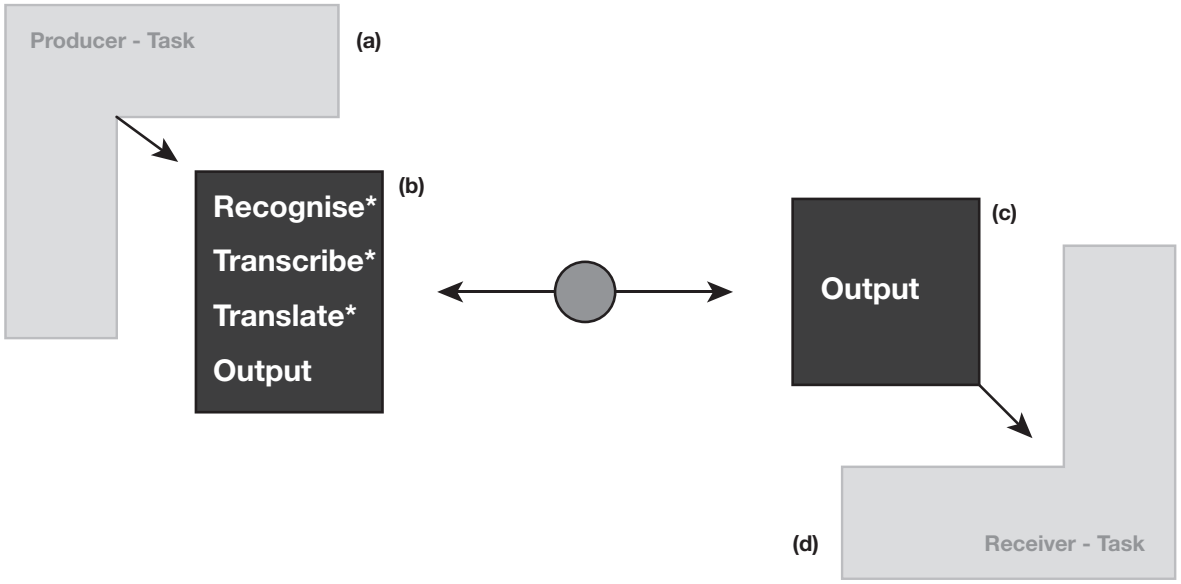


Figure 4.9: *Turn flow diagram when using the ILMT-s2s System. Items (b & c) related to this section are highlighted.*

To record the actions of the subjects, the ILMT-s2s System logged time-stamps of the four principal processes of "Recognition", "Transcription", "Translation", and "Output" to facilitate analysis of the dialogue. Each time the producer clicks and holds the "Push to talk" button, the ILMT-s2s System proceeds to record a log of the following four actions (sample of English utterance "Hi"):

(i) **Recognition:** Recorded utterance — Utterance origin, start time, and file name

```
<utterance speaker="me" st_time="1410259104.03" audio="u
    -1410259104.0-0.0330650806427.wav">
```

(ii) **Transcription:** ASR (Google Speech API) — Start and end time, command, and results

```
    <asr st_time="1.36742997169" end_time="2.23733305931">
      <command>
wget -q -U "Mozilla/5.0" --post-file toasr_tmp.flac --header="Content-
    Type: audio/x-flac; rate=16000" -O - "https://www.google.com/speech-api
    /v2/recognize?client=chromium&amp;lang=en&amp;key=xxxxx"
      </command>
      <raw_output>
{"result":[]}
{"result":[{"alternative":[{"transcript":"hi","confidence":0.82278818},{"
    transcript":"high"},{"transcript":"I"},{"transcript":"hey"},{"
    transcript":"eye"}],"final":true}],"result_index":0}

      </raw_output>
      <clean_output>
hi
      </clean_output>
    </asr>
```

(iii) **Translation:** MT (Microsoft Bing translation system) — Start and end time, command, and results

```
    <mt st_time="2.28379511833" end_time="2.79934000969">
      <command>
wget -q -O - -U "Mozilla/5.0" "http://api.microsofttranslator.com/V2/Http.
    svc/Translate?appid=xxxxx&amp;from=en&amp;to=pt&amp;text=hi"
      </command>
      <raw_output>
&lt;string xmlns="http://schemas.microsoft.com/2003/10/Serialization/"&gt;
    Oi&lt;/string&gt;
      </raw_output>
      <clean_output>
Oi
      </clean_output>
    </mt>
```

(iv) **Output:** TTS send — Start and end time, and send destination

```
    <tts st_time="2.81845903397" end_time="2.81848907471">
      <command>
"Oi" sent to ('192.168.1.9', 49740) for TTS
      </command>
    </tts>
```

When the receiver's system receives an utterance from the producer's computer, the receiver's system logs details of the TTS output command to the log file.

(i) **Output:** TTS output (Mac OS X TTS system) — Utterance origin, start and end time, and command

```
  <utterance speaker="other" st_time="1410259102.08">
    <tts st_time="1410259102.08" end_time="1410259103.23">
      <text>Hello</text>
      <command>
say -v Kate "Hello"
      </command>
    </tts>
```

4.3 The Data Collection

The data for the computer mediated interlingual corpus (the ILMT-s2s corpus)[14] was collected to explore and examine the adaptation of users to Speech-To-Speech Machine Translation (S2S-MT) systems, and provide a starting point for future data collections. For the data of the ILMT-s2s corpus to be a reference, the design of the corpus imitated parts of the HCRC Edinburgh Map Task corpus (Anderson et al., 1991) that used the map task technique (Brown et al., 1985) to elicit the dialogue.

4.3.1 Separation of Subjects

The data collection of the computer mediated interlingual corpus was conducted in Dublin, Ireland, where English is the dominant language. It was presumed that the subjects of the other language would understand English to some degree. This would cause a bias where the *other language* speaker would understand the utterances of the English speaking subject, whereas the English speaker would probably not understand the utterances of the *other language* speaker. To eliminate this bias, the subjects of the data collection were separated so they could not hear the utterances of their interlocutor and therefore rely on the output of the communication tool (§ 4.2).

■ With or Without Eye Contact.

The subjects of the data collection were placed into two different rooms for reasons mentioned in § 4.3.1, so they were not physically co-present and couldn't see each other. With this, two situations were created for the data collection.

(i) The subjects of the dialogue *can* see each other using a video feed ("w/ Video").

(ii) The subjects of the dialogue *cannot* see each other ("w/o Video").

This differentiation followed the setup of the HCRC Edinburgh Map Task corpus where half the subjects can see their interlocutor's face, while the other half are separated with a panel placed between them. In a previous study comparing face-to-face and video-mediated interactions using the map task (O'Malley, Langton, Anderson, Doherty-Sneddon, & Bruce, 1996a, p. 177), it was suggested that communication differed as "when speakers are not physically co-present, they are less confident in general that they have mutual understanding [...], and therefore over-compensate by increasing the level of both verbal and non-verbal information".

4.3.2 The Subjects

The subjects (aged 18–45) of the ILMT-s2s corpus data collection were recruited from Trinity College Dublin, The University of Dublin's digital noticeboard and personal connections. Fifteen recordings of fifteen native English speakers (♀ 5, ♂ 10), and fifteen native Portuguese speakers (♀ 11, ♂ 4), were

[14] The ILMT-s2s corpus data collection obtained the approval of The University of Dublin, Trinity College Dublin, School of Computer Science and Statistics ethics committee in August 2014, under ethical approval No. 133/14.

collected. One subject during each recording session was fitted with a biosignal recording device and an eye-tracking device, while the other subject was not.

Methodological Note

As mentioned under the "Methodological Note" on page 61, data from the biosignal recordings are not explained within the thesis since they are not used in this study.

The subject role (Information Giver or Information Follower) and which subject was to wear the biosignal recording devices were both randomly selected and changed upon request from the subject.[15] The subjects are paired as illustrated in Figure 4.10.

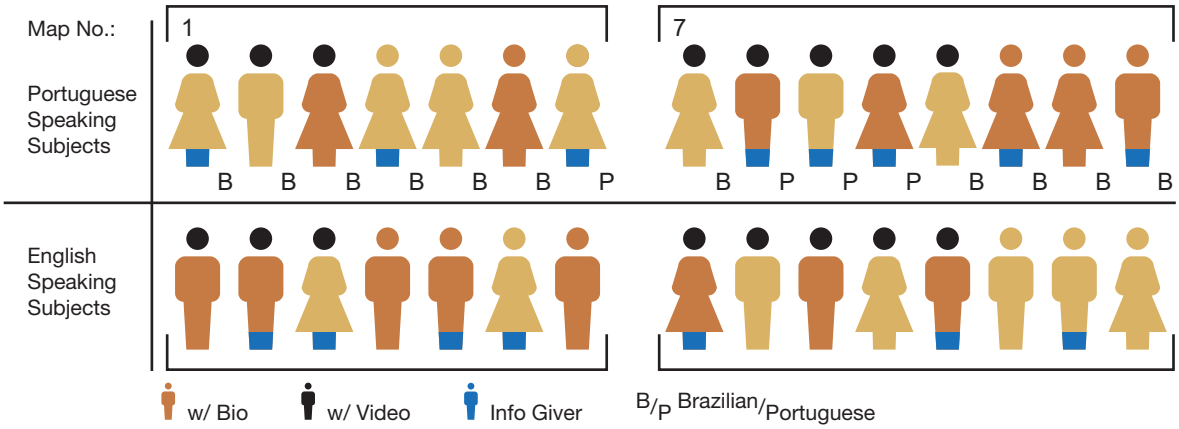


Figure 4.10: *Image of subject pairs and recording situation*

Methodological Note

Within the fifteen dialogues, five English dialogues and one Portuguese dialogue was conducted by an acquaintance of mine, while I too participated in two dialogues since the scheduled subject nor their backup arrived as expected.

The subjects of the ILMT-s2s System data collection were informed about the details of the experiment through an information sheet which was sent to them at least 24 hours before the expected time of the data collection. Immediately before the data collection, the two subjects were together given a short verbal explanation on how to use the ILMT-s2s System (page 46), how it is expected to operate and also a simple explanation of their roles as Information Giver and Information Follower of the map task.

4.3.3 Recording Environment

The recording of the data was conducted in two working offices inside the University of Dublin, Trinity College. Due to the location of Trinity College Dublin being in the centre of Dublin, environmen-

[15] One subject requested not to have the biosignals recorded during the explanation before the recordings, and therefore the biosignal recording devices were worn by their interlocutor.

tal noise from the street outside and students walking in the corridor was also picked up by the microphones and are unfortunately included in the recordings. However, since the findings in this thesis, in § 1.5: "Overview of the Research" are not related to signal processing or signal analysis, the environmental noise is of no concern.

4.3.4 Recording Devices

Methodological Note

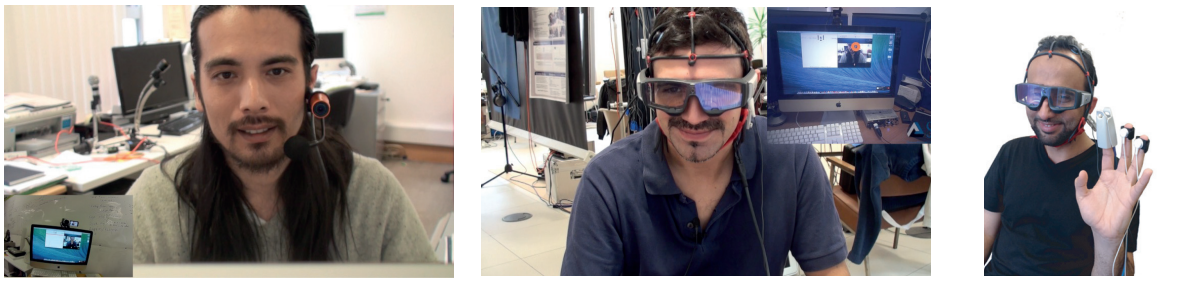
The ILMT-s2s corpus was not only collected for the study in this thesis. Biosignals, such as Skin Conductance (SC), Heart Rate (HR) using the Blood-Volume Pulse (BVP) readings, and the brains' electrical activity through Electroencephelography (EEG) where also collected,[a] as were Eye-tracking data that recorded the subject's eyes and pupil movement.

[a]Publications related to the collected biosignal data can be found in the following: INTERSPEECH'15 (Hayakawa, Haider, et al., 2015), Speech Prosody 2016 (Hayakawa, Haider, et al., 2016), ICASSP 2018 (Haider et al., 2018), the journal Speech Communication (Haider et al., 2019) and the PhD thesis of Fasih Haider (Haider, 2018).

Video, audio, biosignals and eye gaze[16] were recorded during the collection of the ILMT-s2s corpus data. However the setup for subjects that recorded the biosignal (w/ Bio) and those that did not (w/o Bio) differed as indicated in Table 4.1 and illustrated in Figure 4.11. The SONY HDR-XR500, that

Data Type	Subject w/o Biosignal Recording (w/o Bio)	Subject w/ Biosignal Recording (w/ Bio)
Video / Audio	SONY HDR-XR500 for subject front view, recording at 1080i, 29.97 fps	SONY HDR-XR500 for subject front view, recording at 1080i, 29.97 fps
Audio only	Audio-Technica HYP-190H for utterances to the ILMT-s2s System recorded at 96 kHz 24 bit PCM	Sennheiser MKE104 for utterances to the ILMT-s2s System recorded at 96 kHz 24 bit PCM
Video only	Panasonic HX-A100 for subject view range, recording at 1080p, 29.97 fps	—
EyeTracking / Video	—	SMI Eye Tracking Glasses 1.1 recording at 960p, 30 fps (eye tracking data recorded at 24 sps)
Biosignal only	—	Mind Media Nexus-4, for bio sensor recording (upto 1024 sps)

Table 4.1: *Recording devices per subject*



w/o biosignal recording (during recording) w/ biosignal recording (during recording) w/ biosignal recording

Figure 4.11: *Subjects during recordings*

recored the main video and audio of the subjects was the same in both setups, recording the subject's

[16]As mentioned under the "Methodological Note" on page 44, data from the biosignal and eye gaze recordings were not used in this study.

upper torso with the video set at 1080i, 29.97 fps[17] and the audio set at 44.1 kHz 16 bit PCM (ACVHD standard). The other setup that was the same for both subjects were the audio utterance recorded by the ILMT-s2s System (as mentioned on page 46). The audio was recorded at 96 kHz 24 bit PCM, though differences did exist in the microphone that was used. The microphone for the subject without the biosignal (w/o Bio) recording used a head-mounted Audio-Technica HYP-190H microphone, while the subject that did record the biosignal (w/ Bio) used a clip-on Sennheiser MKE104 microphone, since there was little room for a head-mounted microphone to be placed on the subject once the biosignal headset and eye-tracking glasses were worn (refer to the centre image in Figure 4.11).[18]

The following three items are the devices that differed between the subjects that had their biosignals recorded and not. The subject that recorded the biosignals (w/ Bio) used a Mind Media Nexus-4 to collect Skin Conductance (SC), Heart Rate (HR) using the Blood-Volume Pulse (BVP) readings, and the brain's electrical activity through Electroencephelography (EEG). The same subject that used the Mind Media Nexus-4 also wore a SMI Eye Tracking Glasses 1.1 to record pupil dilation, and gaze. A video image was also recorded from the SMI Eye Tracking Glasses 1.1 so that the gaze could be visually tracked. The video recorded from the eye tracking glasses was formatted as 960p[19] and 30 fps[20] while the data from the eye tracking glasses was recorded at 24 samples per second (sps). Since the SMI Eye Tracking Glasses 1.1 recorded video, a counter balance was added for the subject that didn't record the biosignal (w/o Bio). A Panasonic HX-A100 was worn to the left of the left cheek of the subject (as illustrated in the leftmost image in Figure 4.11), and recorded the direction that the subject's head was viewing at a quality of 1080p, 29.97 fps. A Picture-in-Picture image of the video recordings from the SMI Eye Tracking Glasses 1.1 and Panasonic HX-A100 can also be found in the corners of the left and centre images of Figure 4.11.

4.3.5 The Map Task

The recorded dialogues do not have a script, as they are elicited according to the map task scheme, in which the two subjects (dialogue pair) have a specific role designated before the recording starts: an Information Giver and an Information Follower. The Information Giver has a map with a route drawn on it with a START and a FINISH indicated on each side of a route, and has to guide the Information Follower to draw the same route on his/her unmarked copy of the map while neither subject can see the other's map.

[17] The 1080p format would have been more preferable due to the quality improvement for still images in progressive video, but the 1080i was the highest available for the SONY HDR-XR500 camcorder.

[18] The optimal setting would have been to use two clip-on Sennheiser MKE104 microphones, one on each subject, for standardisation purposes, but due to limitations to available equipment, different microphones were used.

[19] A quality of 1080p or even 1080i would have been preferable so as to keep all the video at the HD Video standard, but the 960p was the limit of the device.

[20] The frame rate of 30 fps is the actual 30 fps and not the 29.97 fps that is sometimes abbreviated as 30 fps in cinematography.

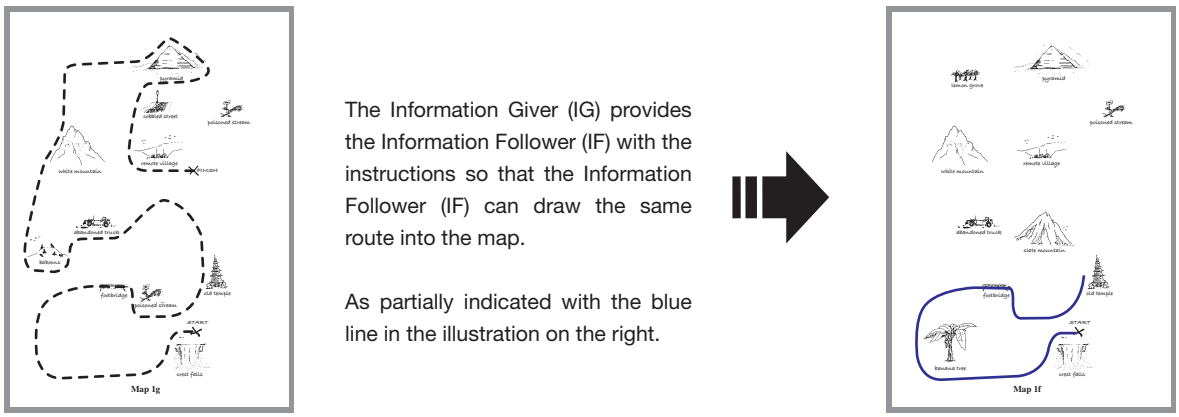


Figure 4.12: *An example of the Information Giver and Information Follower maps*

4.3.5.1 The HCRC Edinburgh Map Task corpus Maps

The maps used to elicit the conversation in this ILMT-s2s corpus data collection were taken from the HCRC Edinburgh Map Task corpus (Anderson et al., 1991).[21] The sixteen HCRC Edinburgh Map Task corpus maps are reported to be of "roughly equal complexity" (Anderson et al., 1991, p. 359) with complexities designed into them using the landmarks on the map. Quoting Anderson et al. (1991, p. 360), the complexities are described as follows:

> "*Landmark types.* All map routes began with a start point marked in the same way on Instruction Giver's and Instruction Follower's maps and ended with a finishing point marked only on the Giver's. Intermediate landmarks along the route alternated between those that were *common* to the Giver's and Follower's map, that is, identical in name, form, and location for both [...], and those that differed in some way. Of the landmarks not common to both, each map contained at least one of each of the following types.
>
> - *Absent/Present* landmarks were found on the Follower's map but not the Giver's [...].
> - *Name Change* landmarks have different names but identical forms and locations on the two maps [...].
> - The *2 : 1* landmarks appear twice on the Giver's map, once in a position close to the route and once far away [...]. The Follower has only the one far away from the route.

[21] As the author of this thesis, and the various other papers that I have published (page 7), when required, I reference the HCRC Edinburgh Map Task corpus with the journal paper instead of the website. This is due to my understanding that the peer reviewed journal paper has higher credibility than the website. However, since the publication of the journal paper more than 25 years ago, the detailed explanation of the maps has changed with the latest update to the website noted to have occured in 2008 (page 67). The difference between the journal paper and the website have been indicated here so that I do not need to add a new explanation of the HCRC Edinburgh Map Task corpus maps which will only be seen as less reliable as either.

- Finally, each basic plan was associated with a *Contrast Feature*, a pair of landmarks with similar names which might elicit contrastive pronunciations (*Green bay* and *Crane Bay*). Over the four map pairs for each basic plan, all the possible combinations of matching and mismatching contrast were represented: Giver's and Follower's maps have both members of the contrast, both maps have only one member, Giver's has both and Follower's one, Giver's has one and Follower's both.

In addition, each map contained an *Odd Man Out*, a landmark which was alien to the stereotypical locations to which all the other landmarks might easily belong [...], as well as a number of other landmarks located at a distance from the route." (Anderson et al., 1991, p. 360)

An example of the above mentioned differences has been highlighted in Figure 4.13[22] on the map that was used as an example in the quoted paper.

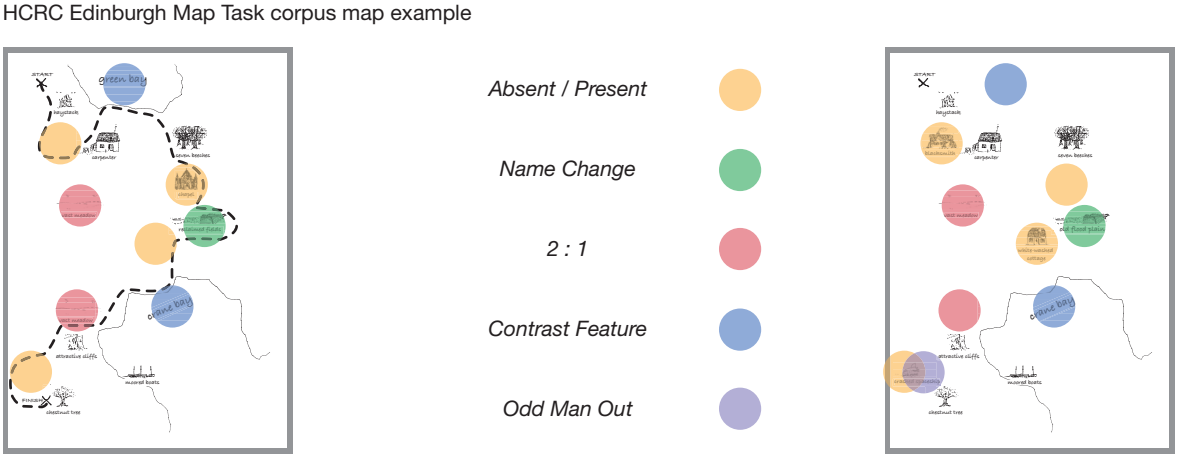


Figure 4.13: *An example of the differences found in the maps — Definition from "Language and Speech" Journal Paper Article: The HCRC Map Task Corpus. Anderson et al. (1991)*

Unfortunately, this explanation in the journal paper by Anderson et al. (1991, p. 360) raises a few questions.

(i) First, what is the meaning of "Intermediate landmarks along the route alternated between those that were *common* [...], and those that differed in some way"? Does this mean that every other landmark was different as in "common – different – common – different", or that the intermediate landmarks were randomly different?

If it does indeed mean that every other landmark was different, then Maps 00, 02, 03, 04, 05, 06, 08, 09, 12, and 15 do not follow this rule and only Maps 01, 07, 10, 11, 13, and 14 follow a "common – different – common – different" landmark pattern[23]

[22] The example map indicated as Figure 4.13 is not map 11 used in the HCRC Edinburgh Map Task corpus. Map 11 of the HCRC Edinburgh Map Task corpus does not include the *Name Change* difference highlighted in this example. Both items in map 11 are named as the same "reclaimed fields".

[23] Putting aside the uniqueness of the maps using route type 01 (Maps 00, 04, 08, and 12) and the landmark in the

(ii) Second, what is the meaning of "Of the landmarks not common to both, each map contained at least one of each of the following types"? The confusion originates from the wordings "of each". Does this mean that at all four of the indicated types of differences were present in each map at least once?

If so, "*Name Change*" is not present in Maps 00, 01, 04, 07, 10, 11, 13, and 14 which is half of the maps. Further more, depending on the definition of "both maps have only one member" which is indicated in the explanation for "*Contrast Features*" map 06 could be added to the list.

(iii) Third, why did the authors only mention the Information Follower's map in the explanation for "*Absent/Present*" as "landmarks were found on the Follower's map but not the Giver's"?

There are no errors in the explanation, but every single Information Giver's map contains one or more landmark that could be defined as a "*Absent/Present*" as listed in Table 4.2 so it is curious as to why they chose not to mention this.

(iv) Fourth, in the explanation for "*Contrast Feature*", there is a phrase that mentions "Over the four map pairs for each basic plan, all the possible combinations of matching and mismatching contrast were represented" with the explanation continuing as "Giver's and Follower's maps have both members of the contrast, both maps have only one member, Giver's has both and Follower's one, Giver's has one and Follower's both". Simply put, the authors have mentioned four combinations; "IG–AB : IF–AB", "IG–A : IF–A", "IG–AB : IF–A", and "IG–A : IF–AB" (where "A" is one member of the contrast and "B" is the other member of the contrast).

The authors have stated that this is "all the possible combinations", then if so, where is "IG–A : IF–B"? Also, what is the exact meaning of "both maps have only one member [of the contrast]"? Does this only mean "A : A", or could it also include "A : B"? Analysis of the provided maps indicate that it only means "A : A", but that is not obvious from the text. Since there are only four route patterns (Figure 4.17) and the authors mention "Over the four map pairs for each basic plan" they have limited the concept to have four combinations, which I cannot understand from the "all" that they indicate.

(v) Finally fifth, the example map used in the journal paper that presents the HCRC Edinburgh Map Task corpus is not a map that is actually used in the HCRC Edinburgh Map Task corpus data collection. The route, the images of the landmark and the difference between the Information Giver and Information Follower are the same as Map 11 of the HCRC Edinburgh Map Task corpus data collection, but the "*Name Change*" difference has been added to one landmark. This is not an error, since it is presented as a "sample" in the caption "Samples of maps used in the Map Task" and not as an actual map, but the similarities of this sample map and the actual

bottom left corner of map 05, the reason why the landmark differences do not follow the "common – different – common – different" pattern is that the "*Name Change* feature is creating changes to the otherwise common landmarks. However this will contradict the indication of "identical in name, form, and location for both".

Map 11, with the only difference being the name of one landmark on the Information Follower's map changing from "reclaimed fields" to "old flood plain", is too small a detail for it not to be misleading — providing the reader of the paper the impression that map 11 of the HCRC Edinburgh Map Task corpus includes all four difference types (including "*Name Change*").[24]

Map	No. of Landmarks	Landmark names
Map 00:	2	Youth Hostel, Thatched Mudhut
Map 01:	2	Baboons, Cobbled Street
Map 02:	2	Wagon Wheel, Swan Pool
Map 03:	2	Forge, Canal
Map 04:	2	Picnic site, LimeStone Cliffs
Map 05:	3	Gorillas, Stones, Soft Furnishing Shop
Map 06:	2	Graveyard, Cattle Ranch
Map 07:	2	Remote Village, Coconut Palm
Map 08:	1	Stone Circle
Map 09:	1	Roman Baths
Map 10:	1	Springboks
Map 11:	1	Chapel
Map 12:	2	Trig Point, Nuclear Test Site
Map 13:	2	Tribal Settlement, Pelicans
Map 14:	2	Graveyard, Cavalry
Map 15:	3	Well, Rockfall, Mountain Stream (away from the route)

Table 4.2: *List of HCRC Edinburgh Map Task corpus map Information Giver landmarks that could be defined as Absent/Present*

It is important to note that the wording provided in the paper by Anderson et al. (1991, p. 360) and the sample map have been modified over time on the HCRC Edinburgh Map Task corpus website[25] with a 2007 publication indicating the description as quoted below (Diderichsen, 2007, pp. 192–193):

- "*Shared landmarks:* Certain landmarks (the majority)[26] were represented on both maps;
- *Absence/presence of landmarks:* Certain landmarks were found on one map but not the other;
- *Name change of landmarks:* Certain landmarks were identical in form and location but had different labels on the two maps;
- *2 : 1 landmarks:* Certain landmarks appeared twice on the instruction giver's map (one in a position close to the route and one more remote), whereas the instruction follower had only the distant one." (Diderichsen, 2007, pp. 192–193)

[24] I understand that this difference is so small that the average reader will not be concerned about this. However, if someone is reading this paper after downloading all the maps, with the intention of using the maps, it will be misleading.

[25] Updated description of the design of the Maps: http://groups.inf.ed.ac.uk/maptask/maptask-description.html (Accessed on 2017/10/06)

[26] This is a personal opinion, but I feel that the authors of the HCRC Edinburgh Map Task corpus seem to want to emphasise the details of the corpus too strongly. With each emphasis of how structured, orderly, well thought the corpus is, they are creating opportunities for errors and counter interpretations. This indication of "the majority" can be replaced with "the minority" as someone creating a dialogue system to navigate the route would do. (There are 414 locations where a landmark can be placed between the START and FINISH of the route. Of the 414 locations, 318 include a landmark. Of the 318 landmarks, 200 are the same on both IG and IF maps, there are 114 landmarks locations that have different contents, and 96 locations that are empty. If the different 114 and empty 96 are added together, it becomes 210 locations that differ which is a value greater than the 200 locations that have the same item.) If this additional description that is thought to be only important enough to be in brackets was not included, this sentence would cause no opposition and it would not cause me to doubt the attention to detail of the authors.

This differs from the original journal paper in that "*Shared landmarks*" has been added, "*Contrast Feature*" has been removed, and *Absent/Present*" has included the map of the Information Giver. The latest indication on the HCRC Edinburgh Map Task corpus website under "Map Design" is as follows (indicated to have been last update on Fri Sep 12 15:04:13 BST 2008[27]):

> "All maps consist of landmarks — or *features* — portrayed as line drawings and labelled with their intended name. The differences in the maps result from the systematic manipulation of a design variable we refer to as **sharedness**: the extent to which features contrast or are shared between pairs of maps. Features were deemed as *common* if the identical form and label appeared in the identical location on both the Giver's and Follower's map. Features which were not common differed in one of three ways:
>
> - *Absent/Present* features were found on one map but not the other;
> - *Name Change* features were identical in form and location but had different labels on the two maps;
> - *2 : 1* features appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant irrelevant one.
>
> All map routes begin with a starting point, marked on both maps, and end with a finishing point marked only on the Instruction Giver's map. Both start and end points are adjacent to a common feature but landmarks between these points alternate in sharedness. This manipulation of mismatches between landmarks enables us to control the information initially shared by the participants."

HCRC Edinburgh Map Task corpus map example

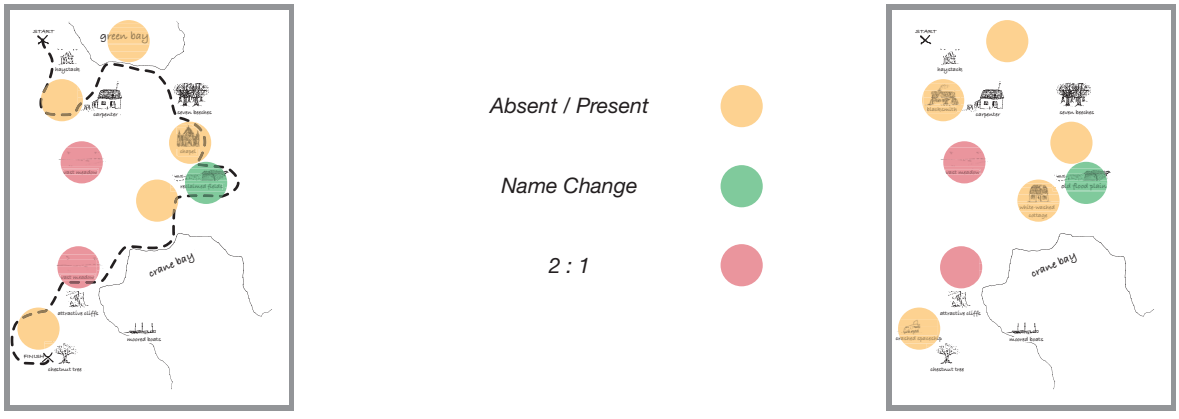


Figure 4.14: *An example of the differences found in the maps — Definition from the HCRC Map Task Corpus Website (Last verified: 2017/10/10)*

The modifications between the peer reviewed journal paper's indication of 1991 and the website of 2008 are subtle, but still substantial enough to change the intended difference of the maps. Defining

[27]Last verified: 2017/10/10 (http://groups.inf.ed.ac.uk/maptask/maptask-description.html)

phrases such as (i) "alternated between those that were *common* to the Giver's and Follower's map, that is, identical in name, form, and location for both [...], and those that differed in some way" has been modified to "alternate in sharedness" with the sharedness defined as "the extent to which features contrast or are shared between pairs of maps" which by changing the wording from "alternate between common – different" to "alternate between sharedness" makes the order in which the difference appear in the maps more ambiguous, (ii) "Of the landmarks not common to both, each map contained at least one of each of the following types" is replaced by "Features which were not common differed in one of three ways" which removes the definite minimum quantity of variance in the difference from four to a one, and (iii) "*Absent/Present* landmarks were found on the Follower's map but not the Giver's" is re-phrased as "*Absent/Present* features were found on one map but not the other" which adds the possibility of the the feature to be present on the Information Giver's map and missing on the Information Follower's. This possibility was not mentioned in the previous explanation, so technically speaking, the explanation was not incorrect, but only incomplete by providing the impression that additional landmarks were only present on the Information Follower's maps.

There are still questions about the depth of the explanation. Maps 00, 04, 08, and 12 use an *Absent/Present* difference that is unique to this route pattern[28] while map 05 uses another *Absent/Present* difference that is unique to this map only. The difference between the Information Giver and Information Follower's maps for all sixteen HCRC Edinburgh Map Task corpus maps can be found in Appendix B.

4.3.5.2 The ILMT-s2s corpus Maps

To make the HCRC Edinburgh Map Task corpus maps usable for the ILMT-s2s corpus data collection, the postscript files of the maps were downloaded[29] and modified using Adobe Illustrator. The original landmark illustrations, placement, routes and names were used but the landmark names where re-written with the "Bradley Hand" font so that all translations that use the latin alphabet could be displayed similarly.

All sixteen maps from the HCRC Edinburgh Map Task corpus have been translated into Portuguese (with a different map created if the translation for the landmarks in Portuguese spoken in Brazil and Portugal were different) and have been through a two step translation checking procedure.[30] The translation was performed using Google Translate and Wikipedia to find the relevant translation of the two hundred landmarks used in the maps and listed in a spreadsheet. Once all the landmarks were translated, the spreadsheet was checked by a native Portuguese speaker before being added to the maps manually as language specific layers[31] in the Adobe Illustrator .eps file. With all the translations

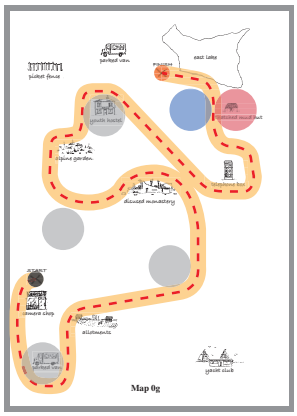
[28] See Figure 4.17 for all four route patterns.

[29] Link to download HCRC Edinburgh Map Task corpus maps: http://groups.inf.ed.ac.uk/maptask/maptasknxt.html (accessed on 2017/04/07).

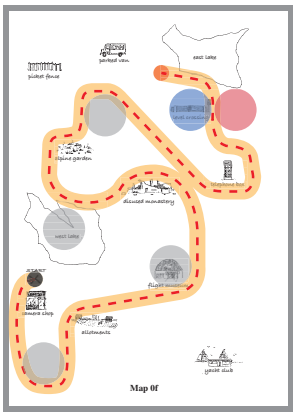
[30] Spanish, German and Japanese variants of the maps have also made with the same method, following the initial intention to create a larger corpus with different languages and also a Wizard-of-Oz variant where the Information Follower is guided by a dialogue system.

[31] By reverse engineering the .svg file format I am sure there might be a faster more appropriate automated way of

HCRC Edinburgh Map Task corpus Map 00 (Route Type 1)

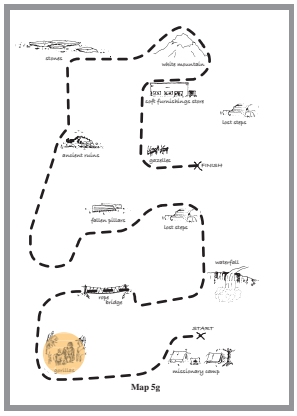


The 2 *Absent/Present* difference that are circled just below "east lake" in this map are unique with this route pattern in that either the differences are not categorised as occuring as an "alternate" (one after the other), or the definition of "identical location" has been increased for this landmark.

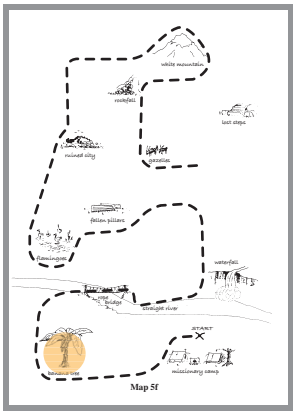


Map of Information Giver (IG) ***Map of Information Follower (IF)***

HCRC Edinburgh Map Task corpus Map 05



The *Absent/Present* difference highlighted in the bottom left corner is unique in the fact that it is the only occurance where a landmark occupying the same location is absent in both maps (in other words, replaced by a different landmark on the other map: Gorillas - Banana Tree).



Map of Information Giver (IG) ***Map of Information Follower (IF)***

Figure 4.15: *Redefining the definition of identical location*

inserted into the maps, the maps were sent to a second native speaker to verify the translation and the manual insertion of the translation. This method was used because of the high expectation that the translation from Google Translate and Wikipedia would be wrong, and the same landmarks were used in various maps. If the first native check was performed with the translation already inserted into the maps, there would have been a higher probability of error in correcting the wrong translation within all the affected maps.

4.3.5.3 Selecting the Maps Used for the Corpus Collection

Of the sixteen available maps, only two maps, Map 01 and Map 07, illustrated in Figure 4.16, were selected to be used for the ILMT-s2s data collection, due to the ease of the route but also the longer average duration of the dialogues created by the maps in the HCRC Edinburgh Map Task corpus. The easy route and long duration was used because the communication using the ILMT-s2s System was expected to be difficult, since the subjects would not be familiar with using a Speech-to-Speech Machine Translation (S2S-MT) system and it was preferred to have bigger dialogue data.

The HCRC Edinburgh Map Task corpus maps are made from 4 route patterns as illustrated in Figure 4.17, that then have 4 different combinations of landmarks following a basic theme on each map. The two maps have been selected for the following reasons:

(i) The ease of the route, derived from a low mean deviation score of the map.[32]

(ii) A longer mean duration of the map, so that the dialogues will not be too short.

(iii) Opposing directions of the route, with the START located in the top or bottom half of the map.

(iv) A close to even gender balance.

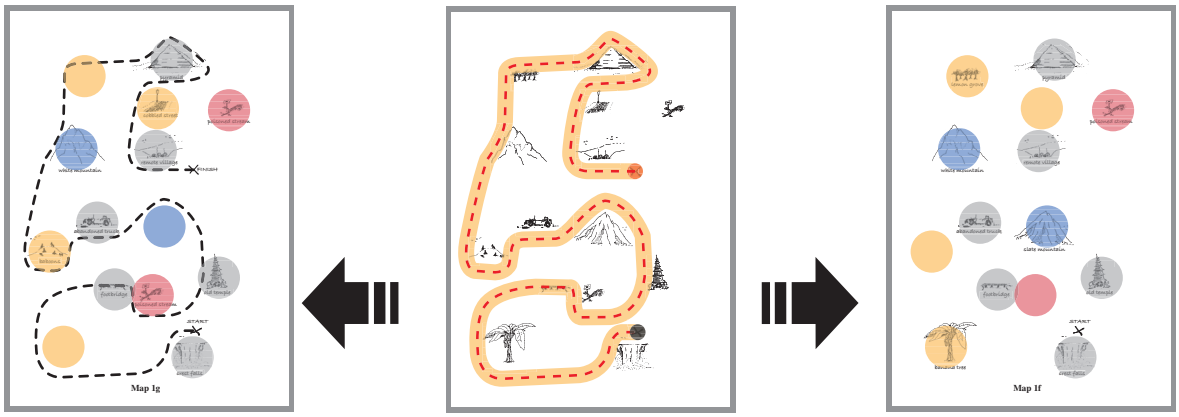
	Gender Balance		Gender by Role				*M* Duration		*M* Dev.		Rank
	Female	Male	IG – ♀	IG – ♂	IF – ♀	IF – ♂	Sec.	Rank	Score	Rank	Total
Map 07 – Type 4	10	6	6	2	4	4	499.4	2	32.9	1	3
Map 00 – Type 1	4	12	0	8	4	4	462.6	5	55.4	4	9
Map 11 – Type 4	7	9	4	4	3	5	513.2	1	67.3	9	10
Map 01 – Type 2	6	10	2	6	4	4	418.4	6	64.0	8	14
Map 08 – Type 1	4	12	0	8	4	4	489.6	3	76.6	11	14
Map 04 – Type 1	8	8	4	4	4	4	359.8	12	54.0	3	15
Map 13 – Type 2	9	7	6	2	3	5	403.9	8	62.6	7	15
Map 06 – Type 3	11	5	6	2	5	3	353.0	13	50.5	2	15
Map 12 – Type 1	10	6	4	4	6	2	486.8	4	116.6	15	19
Map 15 – Type 4	10	6	6	2	4	4	331.0	14	55.8	5	19
Map 10 – Type 3	5	11	2	6	3	5	391.6	9	70.6	10	19
Map 03 – Type 4	10	6	6	2	4	4	299.4	16	59.9	6	22
Map 02 – Type 3	5	11	2	6	3	5	361.0	11	82.6	12	23
Map 09 – Type 2	12	4	6	2	6	2	413.2	7	122.1	16	23
Map 05 – Type 2	9	7	6	2	3	5	372.9	10	94.0	14	24
Map 14 – Type 3	7	9	4	4	3	5	326.6	15	84.4	13	28

Table 4.3: *Ranking of HCRC Edinburgh Map Task corpus dialogues by maps*

doing this task, but for these maps it was performed manually (copy-&-paste).

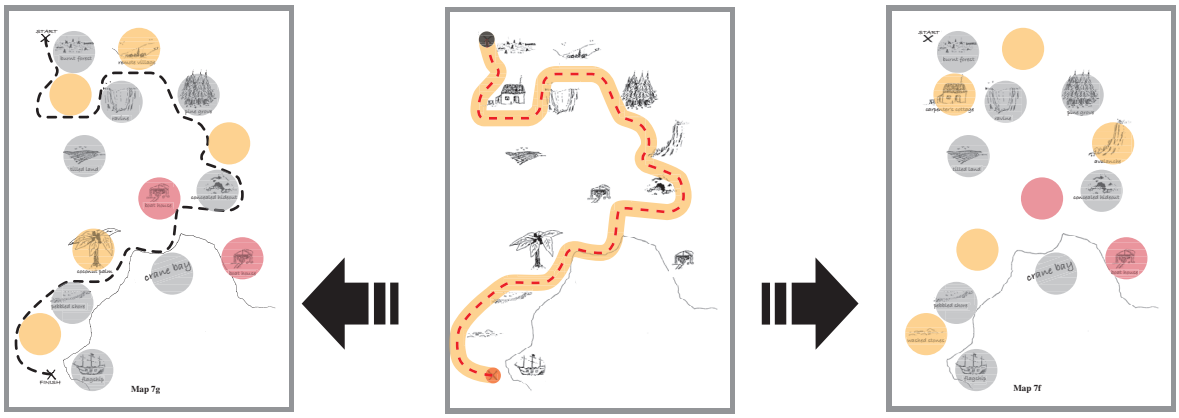
[32]The deviation scores are publicly available at http://groups.inf.ed.ac.uk/maptask/deviation-scores.txt (accessed on 2017/04/07).

HCRC Edinburgh Map Task corpus Map 01



Information Giver (IG) *Information Follower (IF)*

HCRC Edinburgh Map Task corpus Map 07



Information Giver (IG) *Information Follower (IF)*

Same Landmark *Difference Type: 2 : 1*

Difference Type: Absent / Present *Difference Type: Contrast Feature*

Figure 4.16: *Maps used in the ILMT-s2s Data Collection with differences highlighted*

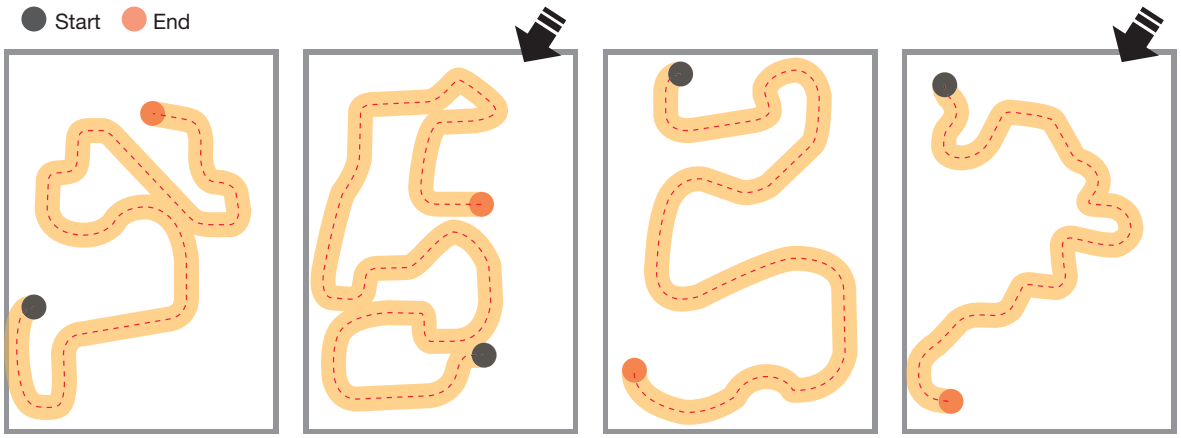


Figure 4.17: *The 4 route patterns of the HCRC Edinburgh Map Task corpus maps*

To determine the maps that are used, the mean deviation score, the mean duration and the gender balance of each map was calculated. Of these, the mean deviation score was ranked from lowest to highest, and the duration was ranked from longest to shortest. These two rankings were added together to make a list that takes both rankings into account as presented in Table 4.3. With Map 07 selected for having the best ranking, other maps starting in the bottom half were removed from the selection (indicated as "Type 3" and "Type 4" in Table 4.3). Of the remaining maps, Map 01[33] was selected since it provided an even gender balance when combined with Map 07,[34] when compared to the higher ranked map of Map 00.

4.3.6 ILMT-s2s System User Survey

To gather a qualitative assessment of the ILMT-s2s System usability, the subjects of the data collection were asked to complete a survey to express their disagreement or agreement on fifteen statements and also answer five questions after each recording. The aim of this survey was to gather a subjective assessment of the user's overall perception of the system's usability.

4.3.6.1 The Fifteen 7 Point Likert Scale Statements

Thirteen of the fifteen statements (S.01 – S.12 & S.15 of Table 4.4) were taken from "The Post-Study System Usability Questionnaire (PSSUQ)" (J. Lewis, 1993, pp. 28–33) and two (S.13 & S.14 of Table 4.4) were added specifically for characteristics of the ILMT-s2s System — the system's TTS voice and the system's output.

7 point Likert scale statements	
S.01:	Overall, I am satisfied with how easy it was to use this system.
S.02:	It was simple to use this system.
S.03:	I could effectively complete the tasks using this system.
S.04:	I was able to complete the task quickly using this system.
S.05:	I was able to efficiently complete the task using this system.
S.06:	I felt comfortable using this system.
S.07:	It was easy to learn to use this system.
S.08:	I believe I could become productive quickly using this system.
S.09:	Whenever I made a mistake using the system, I could recover easily and quickly.
S.10:	The interface of the system was pleasant.
S.11:	I liked using the interface of this system.
S.12:	This system has all the functions and capabilities I expected it to have.
S.13:	I was satisfied with the voice of this system.
S.14:	I was satisfied with the output of this system.
S.15:	Overall, I am satisfied with this system.

Table 4.4: *List of survey statements about the ILMT-s2s System*

The PSSUQ of J. Lewis (1993, p. 14), includes nineteen statements overall, with eight statements regarding the "System Usefulness", seven statements on "Information Quality", three statements on "Interface Quality", and finally one statement for the "Overall Satisfaction" of the system (J. Lewis,

[33]The second fastest (2m 48s) and also the longest duration (19m 00s) dialogue of the HCRC Edinburgh Map Task corpus were recorded with Map 01, but yet those same two dialogues were included in the lowest eleven deviation scores (eleventh and second).

[34]Map 07 has five dialogues out of eight, included within the lowest first quarter route deviation score.

1993, p. 34). For the fifteen statements used in the ILMT-s2s System user survey (Table 4.4), statements S.01 to S.08 were taken from "System Usefulness", statement S.09 from "Information Quality",[35] statements S.10 to S.12 from "Interface Quality" and statement S.15 was from the "Overall Satisfaction". The additional two statements of S.13 and S.14 were added to supplement the S2S-MT aspect of the "Interface Quality" statements.

Responses to the fifteen statements (Table 4.4) were given on a 7 point Likert scale (evenly spaced from 1 = Strongly disagree to 7 = Strongly agree), presented together with an open text field for possible further comments as illustrated in the example of S.01 in Figure 4.18. The subject replies to the 7 point Likert scale statements are listed in § A.5.3 of Section A.5.

Methodological Note

The five open text field questions have not been analysed in this study.

The subject replies to the comments entered into the open text field listed in § A.5.4 of Appendix A. Spelling and grammar of the open text field comments in § A.5.4 have not been corrected and remain as originally entered by the subject.

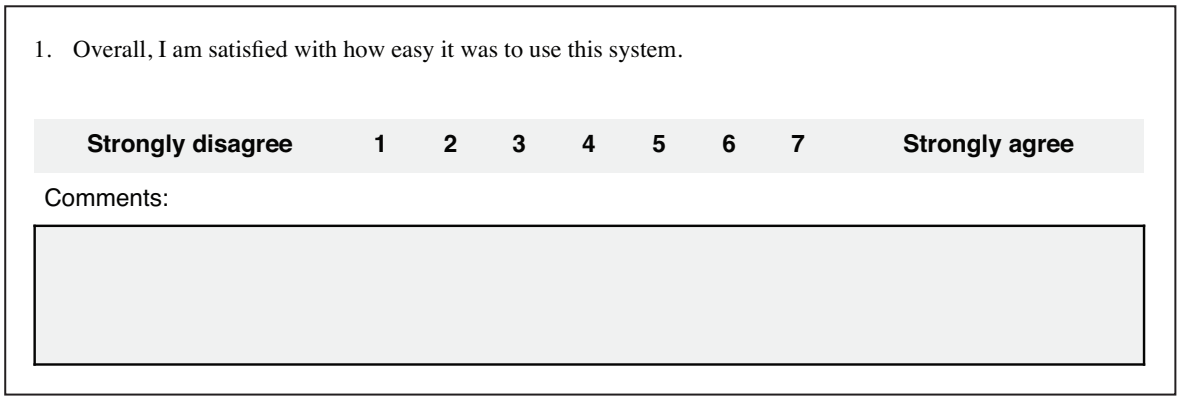


Figure 4.18: *7 point Likert scale user survey actual layout example*

4.3.6.2 The Five Open Text Field Questions

Five questions were added to the user survey to ask the subject how they felt they had adapted to the ILMT-s2s System. The questions, as listed in Table 4.5, were used following observations from the pilot run of the ILMT-s2s System where, for example, the tester subject would change his communication style from phrased utterances to simple word utterances.

Example phrased utterance: Okey I've got it.

Example simple word utterance: Understood.

Hence Q.01 of the five questions.

[35] Of the seven "Information Quality" statements, only one was used, because the other statements were related to the system's error feedback and information retrieval that was related to the usage of a system. Since the ILMT-s2s System did not have any help menu or detailed error messages, these statements were not added to the ILMT-s2s data collection survey.

Open text field questions	
Q.01:	Please indicate why you changed the style of communication.
Q.02:	Please indicate what made you give up clarifying the intention of the other participant.
Q.03:	Please indicate all the things that irritated you.
Q.04:	Please indicate all the things that pleased you.
Q.05:	Please indicate what you felt was most difficult.

Table 4.5: *List of survey questions about the ILMT-s2s System*

The main objective of the questions was to understand where the subjects thought the difficulties were, where the frustration points of using the system were, and also what aspects the subject appreciated. Replies to the questions were entered into the user survey PDF in text fields as illustrated in the example of Figure 4.19.

Methodological Note

The five open text field questions have not been analysed in this study.

The subject replies to the questions entered into the open text field are listed in § A.5.5 of Appendix A.

Spelling and grammar have not been corrected and remain as originally entered by the subject.

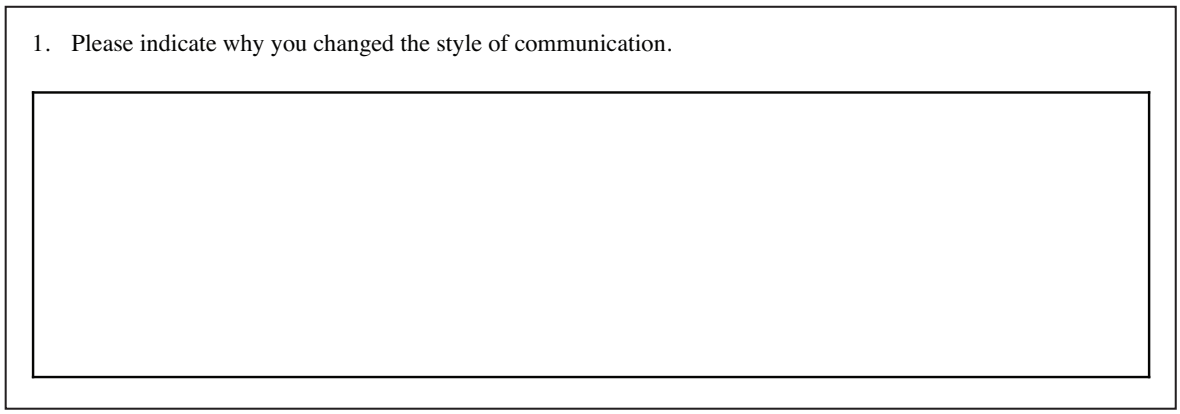


Figure 4.19: *Open text field question user survey actual layout example*

4.3.6.3 Summary of the ILMT-s2s System User Survey

Table 4.6 lists the number of comments and answers entered into the text field of each statement and question. The sample size of each group is indicated in the bottom row.

Item	All Subj.	IG	IF	Por	Eng	♀	♂	w/ Vid	w/o Vid	w/ Crash
S.01	20	10	10	8	12	8	12	11	9	7
S.02	17	7	10	7	10	8	9	9	8	4
S.03	19	7	12	7	12	9	10	10	9	5
S.04	15	8	7	7	8	8	7	9	6	5
S.05	11	6	5	6	5	4	7	6	5	4
S.06	11	4	7	6	5	5	6	5	6	2
S.07	11	6	5	5	6	5	6	7	4	3
S.08	14	7	7	7	7	7	7	9	5	4
S.09	12	7	5	5	7	8	4	7	5	3
S.10	8	5	3	4	4	5	3	4	4	2
S.11	3	2	1	0	3	0	3	2	1	1
S.12	10	4	6	6	4	6	4	6	4	2
S.13	7	3	4	4	3	3	4	4	3	2
S.14	5	2	3	1	4	1	4	2	3	0
S.15	5	2	3	2	3	3	2	3	2	0
Q.01	28	15	13	14	14	15	13	15	13	7
Q.02	21	10	11	10	11	11	10	11	10	4
Q.03	29	15	14	14	15	15	14	16	13	8
Q.04	26	15	11	13	13	14	12	14	12	6
Q.05	27	15	12	14	13	14	13	16	11	8
$n =$	30	15	15	15	15	16	14	16	14	8

Table 4.6: *Text reply count of ILMT-s2s corpus user survey*

4.3.7 Biosignal Recording

Methodological Note

Related publications using this data: INTERSPEECH'15 (Hayakawa, Haider, et al., 2015), Speech Prosody 2016 (Hayakawa, Haider, et al., 2016), ICASSP 2018 (Haider et al., 2018), the journal Speech Communication (Haider et al., 2019) and the PhD thesis of Fasih Haider (Haider, 2018).

As mentioned in Table 4.1, to record the biosignals, a Mind Media Nexus-4[36] was used to collect Skin Conductance (SC), Heart Rate (HR) using the Blood-Volume Pulse (BVP) readings, and the brains' electrical activity through Electroencephelography (EEG). The SC sensor was placed on the middle and ring finger and the BVP sensor was placed on the index finger. EEG sensors were placed in the F4, C4, P4 position with a ground channel placed at A1 of the 10 – 20 location system (Klem, Lüders, Jasper, & Elger, 1999; Kutas & Federmeier, 2011) illustrated in Figure 4.20. The sampling frequency for the SC, HR and EEG were 32 kHz, 32, kHz and 1,024 kHz respectively.

4.3.8 Eye Gaze Recording

Methodological Note

The collected eye tracking data has not been used in this study.

As mentioned in Table 4.1, to record the movement of the subject's eyes and pupil, a SensoMotoric Instruments GmbH (SMI), Eye Tracking Glasses 1.1 was used to record pupil dilation, and gaze.[37]

[36] Refer to http://www.mindmedia.info/CMS2014/products/systems/nexus-4 for details (accessed on 2017/04/07).
[37] The data recorded from the SMI Eye Tracking Glasses has not been used for the studies that I conducted.

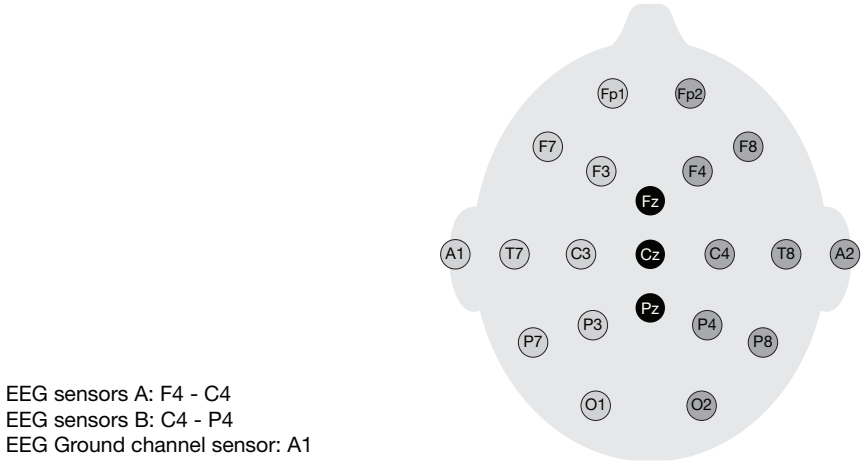


Figure 4.20: *10 – 20 location system layout map*

The video recorded from the eye tracking glasses was formatted as 960p and 30 fps[38] while the data from the eye tracking glasses was recorded at 24 sps.

4.4 Processing the Data

As mentioned in § 4.3: "The Data Collection", video, audio, biosignals and eye-gaze recordings were collected for the ILMT-s2s corpus. The data on its own, does not constitute a corpus, so the data needed to be packaged into a corpus. In this section, I explain how the data was linked together (§ 4.4.1: "Synchronisation", § 4.4.2: "Transcription", § 4.4.3: "Annotation", § 4.4.4: "Calculating the Speech Rate", § 4.4.5: "Calculating the ASR Performance (WER & F-Score)", and § 4.4.6: "20 Frame/Window Normalization").

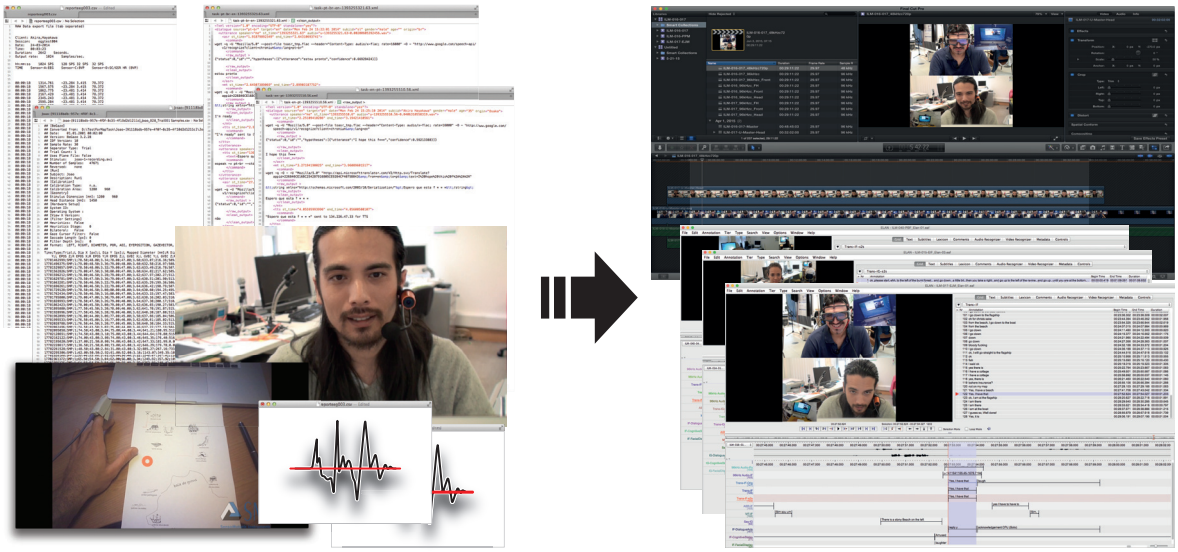


Figure 4.21: *Image of changing the data to a corpus*

[38]The frame rate of 30 fps is the actual 30 fps and not the 29.97 fps that is sometimes abbreviated as 30 fps in other hardware.

4.4.1 Synchronisation

Since all the devices used during the ILMT-s2s corpus data collection recorded video and audio, or audio only, a video editing software called Final Cut Pro X[39] of Apple Inc. was used for the synchronisation. The aim of the synchronisation was to align the start and end time of all the video, audio, biosignal, and eye gaze files of each recording session so that the segmentation time stamps of the transcription or annotation files did not need adjustments (a calculated offset).

The recordings of the data collection was started manually on each recording device, because the devices are not controlled by a single management tool. This method simplified the recording of the subjects, by using available consumer grade devices, but created the need to synchronise the recorded data. The video and audio files that were recorded are synchronised using the non-linear video editing software, Final Cut Pro X from Apple Inc. The synchronisation process arranged the recorded files to start and end at the same moment so that the time stamp and duration would be the same for each subject as illustrated in Figure 4.22.

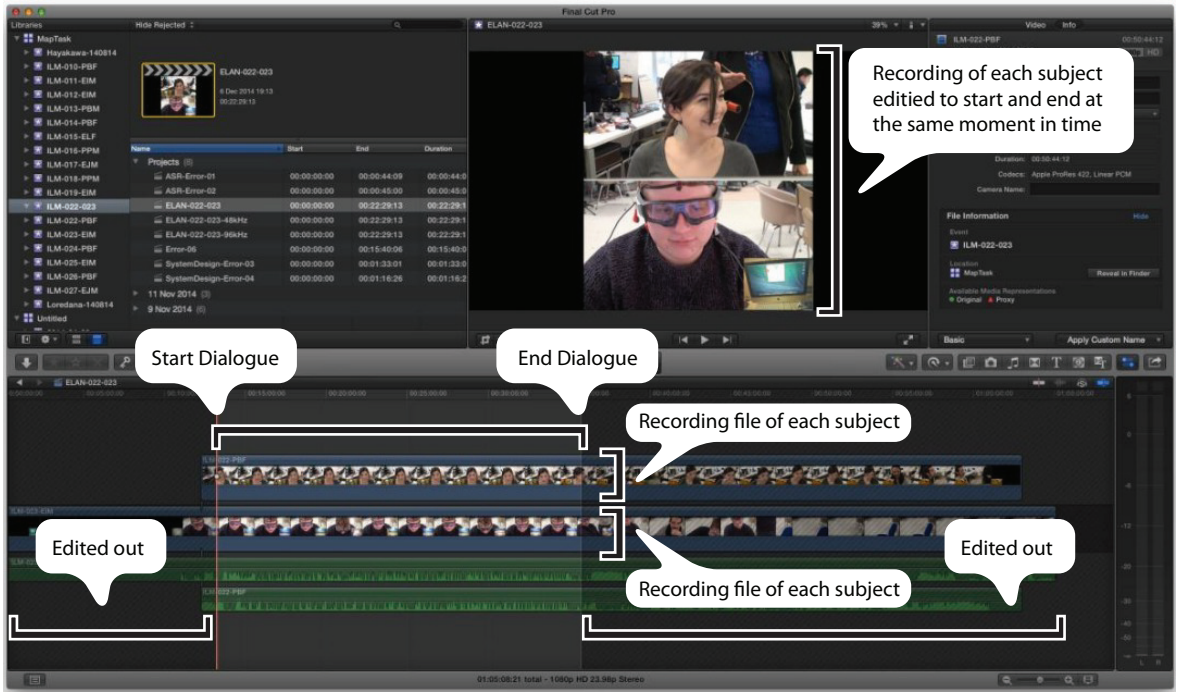


Figure 4.22: *Screenshot of data synchronisation process*

4.4.1.1 The Synchronisation Workflow

Each subject of the data collection had two video files and three or four audio channels (three for the subject w/o Bio and four for subjects w/ Bio) of their dialogue. Each dialogue had two subjects. The idea was to synchronise all the media files for one subject first, and then synchronise each subject

[39]Refer to http://www.apple.com/final-cut-pro/ for details (accessed on 2017/04/07). Final Cut Pro X and Adobe Premiere Pro are both commonly used high quality video editing software with similar capabilities to synchronise video and audio using the audio signal, but Final Cut Pro X was selected due to the single purchase method over the monthly subscription payment method of Adobe Premiere Pro.

to their interlocutor to create a synchronised complete dialogue. The steps for the workflow are as follows:

1. Create a Final Cut Pro X Library for the dialogue and within the Library, three Project folders are created. One Project for each subject and a Project for the dialogue.

2. All media files for one subject are imported into the Project folder of the said subject and repeated for the other subject.

3. All the subject utterance audio files recorded by the ILMT-s2s System and the video file from the SONY HDR-XR500 are selected and automatically synchronised using the Final Cut Pro X synchronisation function.

4. The synchronised file of step 3 is selected with the remaining media files and automatically synchronised to create a set of synchronised files for the subject.

5. Once the media files for one subject are synchronised, the file is then synchronised with their interlocutor's synchronised media files.

6. Once both subject media files are synchronised, the recordings before the task starts and the portion of after the task finishes are cut off and the middle section where the subjects are performing the map task is exported.

7. Finally an XML record of the modifications and synchronisation are exported for all three Projects to keep a detailed log of all the changes that were applied to the media data.

The details of the above synchronisation steps, are listed in § A.2 of Appendix A.

4.4.1.2 Summary of Synchronisation

Of the data recorded for the ILMT-s2s corpus, after the synchronisation and editing process, the fifteen dialogues have a total duration of ≈9h 40m. A summary of the edited data is listed in Table 4.7.

	Total	With Eye-Contact (w/ Video)	Without Eye-Contact (w/o Video)
Dialogue count	15	8	7
Total duration of dialogues	9h 39m 27s	6h 02m 44s	3h 36m 43s
Means	38m 38s	45m 21s	30m 58s
Median	37m 55s	45m 48s	26m 13s
Standard deviation	14m 57s	15m 57s	09m 15s
Front view video (x2)	9h 39m 27s	6h 02m 44s	3h 36m 43s
Eye tracking video (x1)	9h 39m 27s	6h 02m 44s	3h 36m 43s
Head mounted video (x1)	6h 57m 43s	4h 05m 07s	2h 52m 36s
Biosignal signal recording (x1)	9h 39m 27s	6h 02m 44s	3h 36m 43s

Table 4.7: *ILMT-s2s corpus data statistics*

The recordings of the dialogues have been edited and exported in a variety of display formats for varying analysis. A display format showing both subjects in the same video that is listed as "Combined" in Table 4.8, the frontal view of a single subject listed as "Front", a view of the head

mounted cameras listed as “Head” and “Eye Track” and finally the audio of both subjects combined into a stereo file and also of the stereo audio file of a single subject with the recordings from the camcorder on one channel and the ILMT-s2s System utterance recordings in the other channel.

Dialogue (Setting, Map used)	Duration	Combined	Front	Head	Eye Track	Audio	Offset (seconds)	
							Biosignal	Eye Track
01 (w/ Vid, Map 01)	00:23:02	✓	✓	✓ IG	✓ IF	✓	1,508.0250	2,151.4827
02 (w/ Vid, Map 01)	00:48:57	✓	✓	✓ IF	✓ IG	✓	168.3897	695.3280
03 (w/ Vid, Map 01)	01:14:20	✓	✓	✗	✓ IF	✓	316.5269	485.5184
04 (w/o Vid, Map 01)	00:22:31	✓	✓	✓ IG	✓ IF	✓	243.5931	798.1307
05 (w/o Vid, Map 01)	00:40:20	✓	✓	✓ IF	✓ IG	✓	331.8071	381.3476
06 (w/o Vid, Map 01)	00:44:07	✓	✓	✗	✓ IF	✓	1,071.3739	1,172.9718
07 (w/o Vid, Map 01)	00:37:55	✓	✓	✓ IG	✓ IF	✓	621.5286	783.2825
08 (w/ Vid, Map 07)	00:59:29	✓	✓	✓ IF	✓ IG	✓	143.7948	273.4398
09 (w/ Vid, Map 07)	00:29:13	✓	✓	✓ IF	✓ IG	✓	287.3204	162.1954
10 (w/ Vid, Map 07)	00:36:07	✓	✓	✓ IG	✓ IF	✓	120.6069	398.9986
11 (w/ Vid, Map 07)	00:43:17	✓	✓	✗	✓ IG	✓	709.6662	1,060.3593
12 (w/ Vid, Map 07)	00:48:19	✓	✓	✓ IF	✓ IG	✓	399.5372	791.9245
13 (w/o Vid, Map 07)	00:26:13	✓	✓	✓ IF	✓ IG	✓	810.7966	858.8580
14 (w/o Vid, Map 07)	00:25:59	✓	✓	✓ IG	✓ IF	✓	315.8362	346.8799
15 (w/o Vid, Map 07)	00:19:38	✓	✓	✓ IF	✓ IG	✓	424.6955	526.3592

Table 4.8: *Synchronised files of the ILMT-s2s corpus*

However, the data files from the biosignal recording device and the SMI Eye Tracking Glasses have not been edited. Therefore, the offset time to synchronise the start are provided in Table 4.8.

4.4.2 Transcription

Two students (one native speaker of English and one native speaker of Portuguese)[40] were recruited to orthographically transcribe the edited audio files using the open source software Wavesurfer (Sjölander & Beskow, 2000). Since the subjects within the dialogues spoke different languages, the dialogue of the subject pair was transcribed separately as two independent dialogues. Furthermore, since the subject's were mainly talking to the ILMT-s2s System, their utterances were short and conclusive so each utterance to the system could be regarded as a turn. Additionally for the Portuguese text, the dedicated annotation tool ELAN (Wittenburg et al., 2006) was used to add accented letters that Wavesurfer could not handle.

4.4.2.1 The Transcription Workflow

The two students transcribed the thirty synchronised audio files (fifteen each in their native language). The steps for the workflow are as follows:

1. Open the stereo audio file that contained both subject utterances in Wavesurfer.

2. Apply the display configuration, as illustrated in Figure 4.23, so that the waveform of both channels and the spectrogram, power plot of the channels to be transcribed is displayed with the text insertion field. The display range was also set to be standardised between the two

[40] Both students that transcribed and annotated the dialogue files were participants of the data collection.

student transcribers at 2 seconds[41] to try to achieve a unified onset gap from the segmentation start line and the onset of the subject utterance.

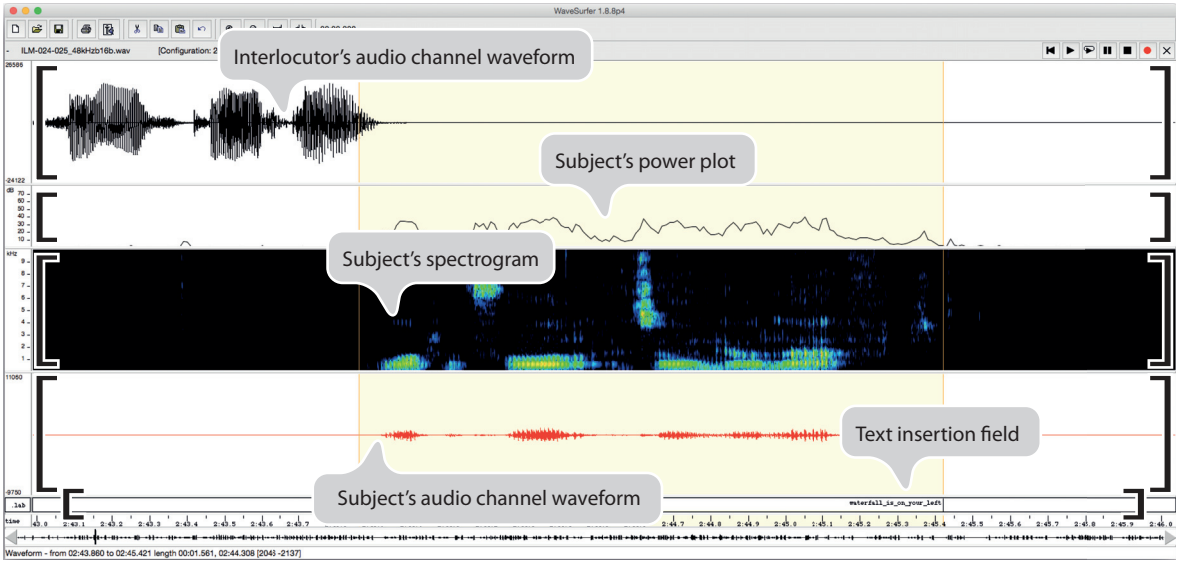


Figure 4.23: *Example of Wavesurfer transcribing configuration for right channel audio*

3. Segment the audio files in chunks that the student transcribers thought the subject stopped talking (turns) and transcribe the utterance of each segment.[42]

4. I verify the segmentation of both English and Portuguese files and the English transcription to correct any obvious errors or highlight for verification with the student transcribers on a later occasion.[43]

5. Convert the transcription text file from the Wavesurfer format of "Start time – Space – End time – Space – Transcription text with words connected by underscores instead of spaces" to a more conventional tab delimited csv format and import into the dedicated annotation tool ELAN.

6. Link the transcriptions to the video that includes both subjects and also the stereo audio file that was used to transcribe the dialogue.

7. For the Portuguese transcription only: add the accentuation of the letters, since I could not setup Wavesurfer to open files containing UTF-8 unicode characters (in other words, Wavesurfer would save accentuated letters, but each time the file was reopened, the accents would be removed).

[41] The range displayed in the window could be adjusted as the student transcribers pleased, but by using a standardised range each time the window was opened, it was thought that it would be easier to obtain a closer onset. This intention was also explained to the transcribers.

[42] The headphones used by the student transcribers was capable of adjusting the volume of either audio channel, so the transcribers were capable to mute the interlocutor's audio channel as they wanted.

[43] Since I had spoken to all the subjects, I had a better understanding of the subjects' interests and background that made it possible to correct utterances that were difficult to hear and understand. One example of this would be "thanks for not responding" being changed to "python not responding". As for the Portuguese transcription, after the process of step 7 was completed, the Portuguese transcription was verified by a different native Portuguese speaker with the request to verify the transcription, spelling and usage of accents.

4.4.2.2 Summary of the Transcription

19h 18 m 54s of audio files (9h 39m 27s each of English and Portuguese dialogues) have been transcribed to create 2h 15m 41s of orthographically transcribed utterances (1h 3m 59s of English utterances and 1h 11m 41s of Portuguese utterances) as summarised in Table 4.9.

	Total	With Eye-Contact (w/ Video)	Without Eye-Contact (w/o Video)
No. of dialogues transcribed	15	8	7
Total duration of dialogues	9h 39m 27s	6h 02m 44s	3h 36m 43s
Duration of transcribed utterances	2h 15m 41s	1h 19m 30s	56m 10s
English speaker	1h 03m 59s	38m 54s	25m 05s
Information Giver	41m 09s	24m 53s	16m 15s
Information Follower	22m 49s	14m 00s	08m 49s
Portuguese speaker	1h 11m 41s	40m 36s	31m 05s
Information Giver	37m 11s	17m 01s	20m 10s
Information Follower	34m 30s	23m 35s	10m 55s
No. of word types (lemmas)			
English speaker	1,019	769	689
Portuguese speaker	1,169	906	599
Portuguese	925 (Pt)+244 (En)	683 (Pt)+223 (En)	549 (Pt)+50 (En)
No. of word tokens	18,678	11,101	7,577
English	9,728	5,854	3,874
Information Giver	6,029	3,562	2,467
Information Follower	3,699	2,292	1,407
Portuguese	8,950	5,247	3,703
Information Giver	4,213	1,906	2,307
Information Follower	4,737	3,341	1,396
Mean word tokens per dialogue	1,245	1,388	1,082

Table 4.9: *ILMT-s2s corpus data transcription summary*

4.4.3 Annotation

The two students who transcribed the data also annotated the data using the dedicated annotation tool ELAN (Wittenburg et al., 2006) for perceived emotions (below), dialogue acts (page 84), and facial expressions and head movements (page 88) annotations. Furthermore, from the XML file created by the ILMT-s2s System (§ 4.2.6: "System Log (XML)") and also the XML file exported from Final Cut Pro X (§ 4.4.1: "Synchronisation"), utterances were labelled as *On-Talk*, *Off-Talk*, *Off-Talk Self*, and *Off-Talk Other* to indicate who the utterance was directed to (page 90) and also imported as tiers to the ELAN data file.

4.4.3.1 Perceived Emotions [3 labels]:

The perceived emotions "*Surprised*", "*Frustrated*", and "*Amused*" were annotated using the dedicated annotation tool ELAN (Wittenburg et al., 2006) by the two students who also transcribed the audio files. The inter-coder agreement for the annotated perceived emotion labels was calculated using the modified kappa feature of ELAN 4.9.0's "Inter-Annotator Reliability..." function on one of the dialogues and the results are well above 60%.[44]

Perceived Emotions	Description
Surprised:	The subject is in a state of surprise.
Amused:	The subject is in a state of amusement.
Frustrated:	The subject is in a state of frustration.

Table 4.10: *Description of perceived emotions annotation labels*

Understanding that there are numerous perceived emotions, proposed by various annotation schemes (McCowan, Carletta, et al., 2005; Popescu-Belis, 2005; Juel Henrichsen & Allwood, 2013), the initial annotation was limited to only "*Surprised*", "*Amused*" and "*Frustrated*". My motivation to only looking into the three perceived emotions of "*Surprised*", "*Frustrated*", and "*Amused*" in the data is related to that of Batliner et al. (2000). However, changing the vantage point from the corporate side of financial cost, to the human aspect of emotional cost of the communication breakdown, both from the manufacturing side and the consumer. Since the data collection is task based, the understanding was that the subjects would be focused on providing and receiving clear instructions. However, due to "*the vocabulary problem*" (Furnas, Landauer, Gomez, & Dumais, 1987), where each individual has their own understanding and preference for describing situations, clear communication is difficult even in Human-to-Human situations. By adding the *filters* (Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech (TTS)) of the ILMT-s2s System, further complication will arise. The assumption was that this would bring out the selected cognitive characteristic in the interaction the subjects had with the ILMT-s2s System. This reasoning is further explained in § 2.2.3: "Perceived Emotion (Cognitive State)".

[44] Due to the limited size of the ILMT-s2s corpus, all perceived emotion labels were verified after each dialogue was completed by me, and ambiguous items were discussed with the student annotators for verification.

■ The Perceived Emotions Workflow.

As part of the preparation, the ELAN files for each dialogue had a tier added for the segmentation of perceived emotions. The three perceived emotions of "*Surprised*", "*Frustrated*", and "*Amused*" were also made selectable through a pull down menu once the segmented section was activated for input. The steps for the workflow are as follows:

1. Open the ELAN file for the dialogue that has been pre-linked to the video that includes both subjects and also the stereo audio file that was used to transcribe the dialogue.

2. Segment the data at locations that the student annotators thought the subject was visibly in one of the three perceived emotions ("*Surprised*", "*Frustrated*", and "*Amused*") and select the perceived emotion from the pull down menu as illustrated in Figure 4.24 for each segment.

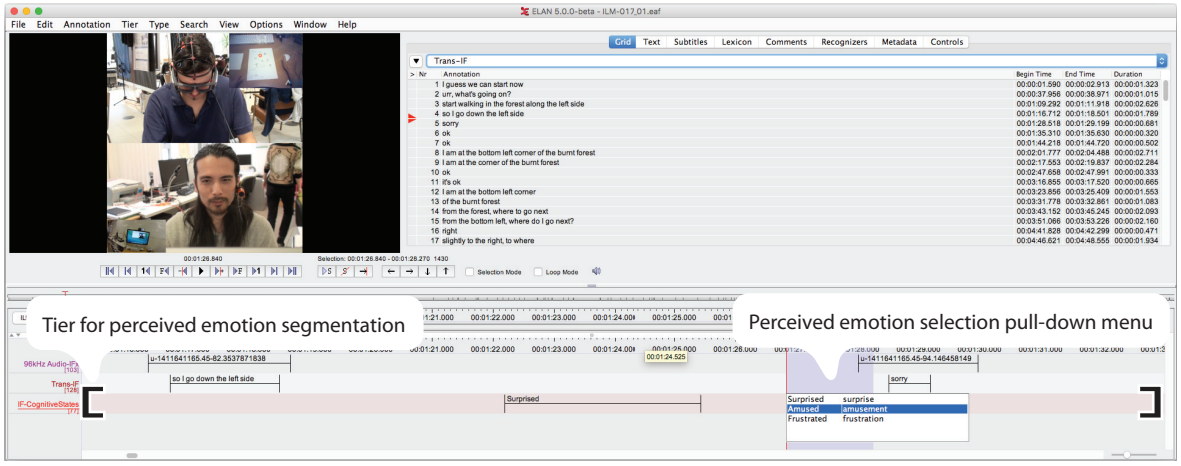


Figure 4.24: *Example of ELAN perceived emotion annotation*

3. I verify all perceived emotion labels and highlighted ambiguous items for verification with the student annotators on a later occasion.

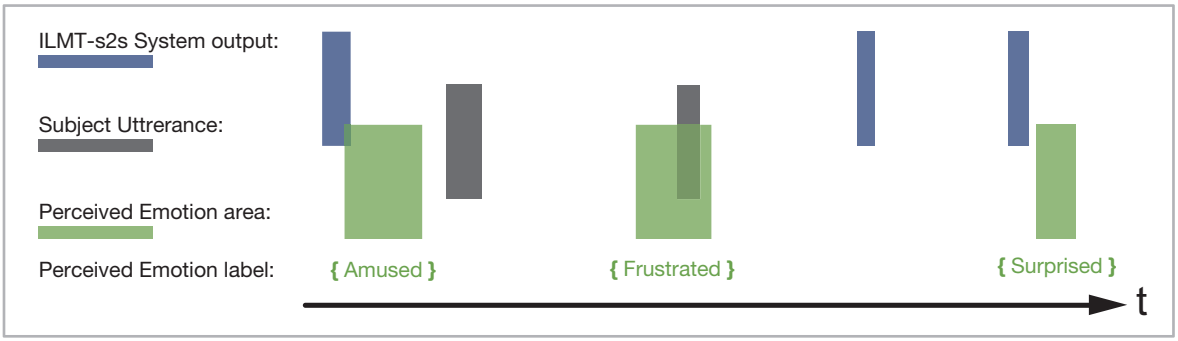
■ Summary of Perceived Emotions.

There are a total of 1,835 labeled perceived emotions summarised as shown in Table 4.11. As previously mentioned, the labels of the perceived emotions are not restricted to the utterance location, but freely applied within the dialogue. In other words, the annotators were allowed to freely segment the

Cognitive states	Count	Total duration (approx.)
Amused	812	1 h 33 m
Frustrated	670	51 m
Surprised	353	17 m
Total	1,835	2 h 41 m

Table 4.11: *Summary of perceived emotions annotated in the ILMT-s2s corpus*

dialogue so that the label covers the area where the subject visually expressed the given state, which, as illustrated in Figure 4.25, does not typically correspond to the location of the subject utterance



ILMT-s2s TTS output:	740.571 – 742.388	*what makes up the bay of goals*
Perceived Emotion:	742.000 – 746.950 *Amused*	
Subject Utterance:	748.458 – 750.747	*there is the crane bay, yes?*
Perceived Emotion:	760.570 – 765.390 *Frustrated*	
Subject Utterance:	763.221 – 764.675	*crane bay*
ILMT-s2s TTS output:	774.668 – 775.786	*excuse could low*
ILMT-s2s TTS output:	784.343 – 785.641	*way toward coqueiro*
Perceived Emotion:	786.100 – 788.650 *Surprised*	

Figure 4.25: *Image of dialogue segmentation — Actual timing from Dialogue 15*

or the ILMT-s2s System output. The distribution of the perceived emotion labels within the major groupings of the ILMT-s2s corpus are illustrated in Figure 4.26. The state of "*Surprised*" being the

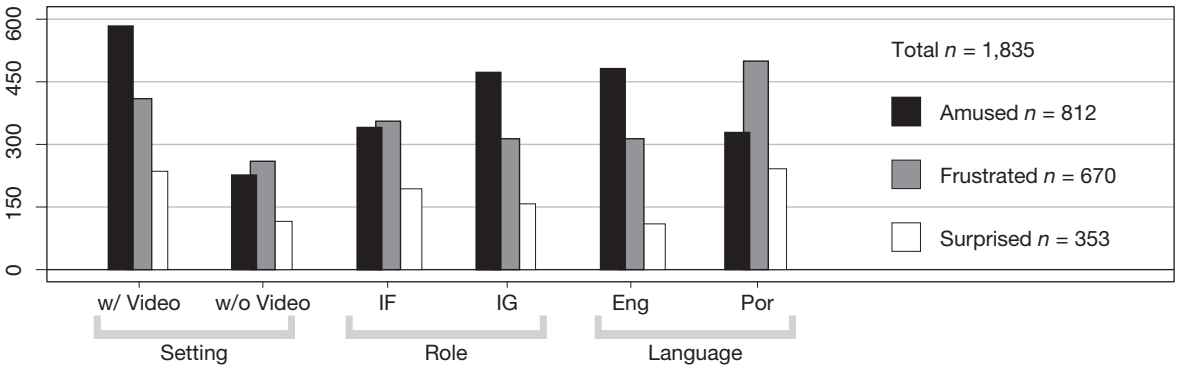


Figure 4.26: *Distribution of perceived emotions within the ILMT-s2s corpus — Categorised by group. The "n" indicates the number of labels within each grouping of "Setting", "Role", and "Language".*

least frequent in all settings, as can be predicted from the overall low total, but interestingly the frequency of the states "*Amused*" and "*Frustrated*" interchange within each grouping.

4.4.3.2 Annotation of Dialogue Acts [25 labels]:

The twenty five dialogue acts of "*Acknowledgement CP/CPU*", "*Align*", "*Check*", "*Clarify*", "*Explain*", "*Instruct*", "*Interjection*", "*Query y-n/w*", "*Reply y/n/w*", with also a "*Solo*" variant of each dialogue act except "*Interjection*" were used as annotation labels to mark the type of utterance being spoken. A description of each dialogue act that was presented to the same student annotators that annotated the perceived emotions and transcribed the data is listed in Table 4.12.

These labels were based on the Dialogue Structure Coding scheme (Carletta et al., 1997), but with modifications to "Acknowledgement" to differentiate the simple acknowledgement (CP) of the

Dialogue Act [25]	Description
Acknowledgement CP:	A minimal verbal response that gives continuation feedback.
Acknowledgement CPU:	A minimal verbal response that gives acceptance/understanding feedback.
Align:	Checks the attention or agreement of the partner, or his/her readiness for the next move.
Check:	Requests the partner to confirm information that the checker has some reason to believe, but is not entirely sure about.
Clarify:	A repetition of information which the speaker has already stated, often in response to a check move.
Explain:	States information which has not been elicited by the partner.
Instruct:	Commands the partner to carry out an action.
Interjection	
Query Y-N:	A yes/no question.
Query W:	Any query which is not covered by the other categories.
Reply Y:	Any reply to any query with a yes-no surface form which means "yes", however that is expressed.
Reply N:	A reply to a query with a yes/no surface form which means "no".
Reply W:	Any reply to any type of query which doesn't simply mean "yes" or "no".
(Solo) variant:	Applies to all the above (e.g., "Acknowledgement CP (Solo)") except "Interjection" and is used when the subject is not talking to the ILMT-s2s System (*Off-Talk*).

Table 4.12: *Description of dialogue acts annotation labels*

utterance from the acknowledgement with the actual understanding of the utterance (CPU) which is defined under the MUMIN coding scheme (Allwood et al., 2007).[45] Also, a "Solo" variant of the dialogue acts, and "Interjection" were added due to the "*Off-Talk*" characteristics (Oppermann et al., 2001) of the collected data.

■ The Dialogue Act Workflow.

One of the dialogue act labels mentioned in Table 4.12 was assigned to each utterance of the dialogue. If the utterance contained more than one dialogue act, the segmentation of the utterance was split on the dialogue act tier at the timing where the dialogue act changed and a dialogue act label was assigned accordingly. As with the perceived emotion labels, the ELAN files for each dialogue had a tier added for the segmentation of dialogue acts. The twenty five dialogues acts were also made selectable through a pull down menu once the segmented section was activated for input to minimise errors. The steps for the workflow are as follows:

1. Open the ELAN file for the dialogue that has been pre-linked to the video that includes both subjects and also the stereo audio file that was used to transcribe the dialogue.

2. Copy the segmentation locations from the transcription tier and select the dialogue act from the pull down menu as illustrated in Figure 4.27 for each segment. If the segmented utterance included more than one dialogue act, split the segment where required, before selecting the dialogue act from the pull down menu.

[45] The decision to use these two coding scheme was thanks to Dr. Loredana Cerrato.

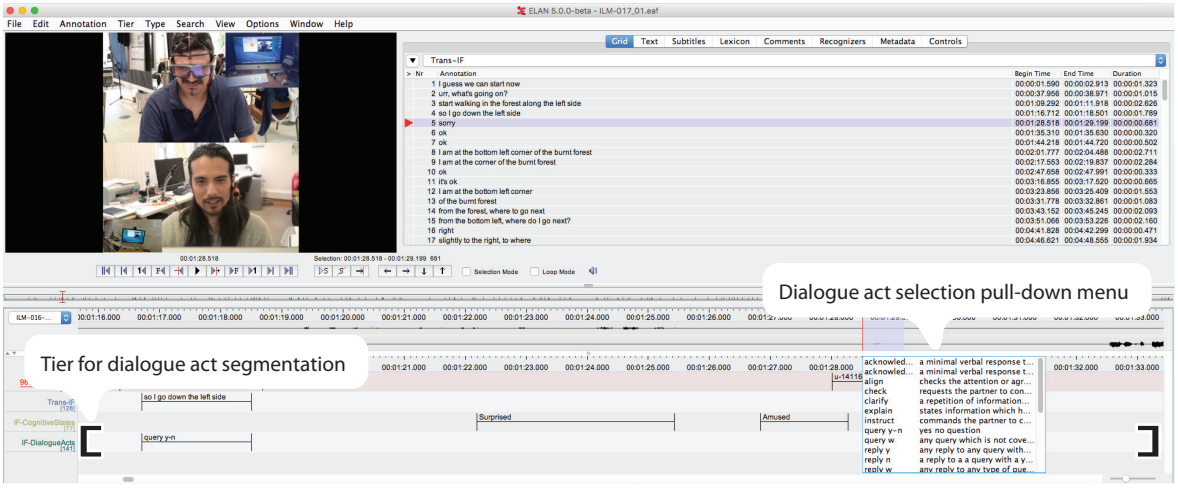


Figure 4.27: *Example of ELAN dialogue act annotation*

3. Proceed to the annotation of facial expressions and head movements (page 88).

■ Summary of Dialogue Acts.

There are in total of 3,850 labeled dialogue acts summarised as shown in Table 4.13 and distributed as illustrated in Figure 4.28. Since the dialogue act scheme used for this ILMT-s2s corpus are fundamentally the same as that used by the HCRC Edinburgh Map Task corpus, this annotation data can be used in a direct comparison with the dialogues of the HCRC Edinburgh Map Task corpus (Chapter 8). It is interesting to note from Table 4.13 and Figure 4.28 that only the dialogue act of

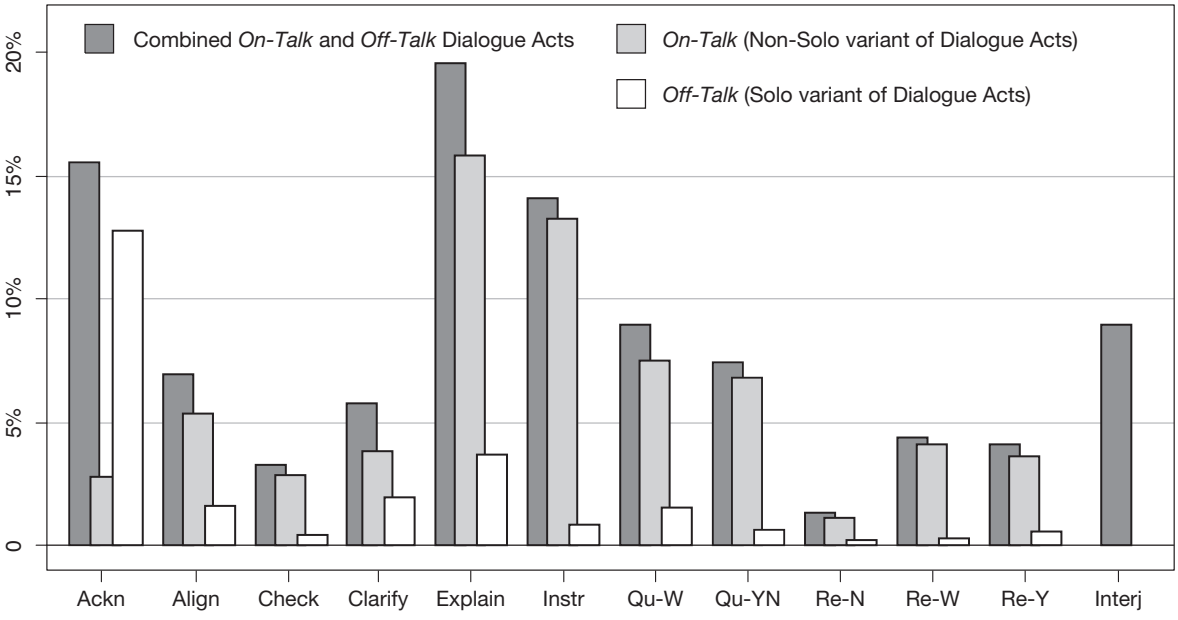


Figure 4.28: *Bar chart of dialogue acts: Total of On-Talk and Off-Talk variant's combined as well as separated — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item*

"*Acknowledgement*" occurs more frequently during *Off-Talk*. Of the 597 "*Acknowledgement*" labels, 490 appear during *Off-Talk* and only 107 appear during *On-Talk*. This indicates that the subjects of the corpus are acknowledging the communication from their interlocutor, but are only transmitting it to their interlocutor ≈18% of the time through the ILMT-s2s System with *On-Talk*, while ≈82%

Dialogue Acts	Total			With Eye-Contact (w/ Video)			Without Eye-Contact (w/o Video)		
	Count	All (%)	Sub (%)	Count	All (%)	Sub (%)	Count	All (%)	Sub (%)
Utterance count	3,628	—	—	2,151	—	—	1,477	—	—
All dialogue acts	3,850	—	—	2,283	—	—	1,567	—	—
Subtotal *On-Talk*	2,574	66.86	—	1,444	63.25	—	1,130	72.11	—
Subtotal "Interjection"	343	8.91	—	240	10.51	—	103	6.57	—
Subtotal *Off-Talk* (Solo)	933	24.23	—	599	26.24	—	334	21.31	—
Acknowledgement CP	10	0.26	0.39	1	0.04	0.07	9	0.57	0.80
Acknowledgement CPU	97	2.52	3.77	42	1.84	2.91	55	3.51	4.87
Align	206	5.35	8.00	91	3.99	6.30	115	7.34	10.18
Check	110	2.86	4.27	66	2.89	4.57	44	2.81	3.89
Clarify	146	3.79	5.67	64	2.80	4.43	82	5.23	7.26
Explain	609	15.82	23.66	428	18.75	29.64	181	11.55	16.02
Instruct	510	13.25	19.81	276	12.09	19.11	234	14.93	20.71
Query Y-N	262	6.81	10.18	142	6.22	9.83	120	7.66	10.62
Query W	287	7.45	11.15	145	6.35	10.04	142	9.06	12.57
Reply Y	137	3.56	5.32	84	3.68	5.82	53	3.38	4.69
Reply N	43	1.12	1.67	29	1.27	2.01	14	0.89	1.24
Reply W	157	4.08	6.10	76	3.33	5.26	81	5.17	7.17
Interjection	343	8.91	—	240	10.51	—	103	6.57	—
Ackn. CP (Solo)	85	2.21	9.11	57	2.50	9.52	28	1.79	8.38
Ackn. CPU (Solo)	405	10.52	43.41	258	11.30	43.07	147	9.38	44.01
Align (Solo)	60	1.56	6.43	32	1.40	5.34	28	1.79	8.38
Check (Solo)	16	0.42	1.71	3	0.13	0.50	13	0.83	3.89
Clarify (Solo)	75	1.95	8.04	33	1.45	5.51	42	2.68	12.57
Explain (Solo)	142	3.69	15.22	128	5.61	21.37	14	0.89	4.19
Instruct (Solo)	31	0.81	3.32	9	0.39	1.50	22	1.40	6.59
Query Y-N (Solo)	23	0.60	2.47	20	0.88	3.34	3	0.19	0.90
Query W (Solo)	57	1.48	6.11	27	1.18	4.51	30	1.91	8.98
Reply Y (Solo)	21	0.55	2.25	18	0.79	3.01	3	0.19	0.90
Reply N (Solo)	8	0.21	0.86	6	0.26	1.00	2	0.13	0.60
Reply W (Solo)	10	0.26	1.07	8	0.35	1.34	2	0.13	0.60

Table 4.13: *Summary of dialogue acts annotated in the ILMT-s2s corpus*

of the time, it is not transmitted verbally to the interlocutor. As illustrated in Figure 4.29, ≈53% of *Off-Talk* dialogue acts have been labeled as an acknowledgement, whereas acknowledgements only represent ≈4% of *On-Talk* dialogue act labels. Even with the labels split into groupings of setting

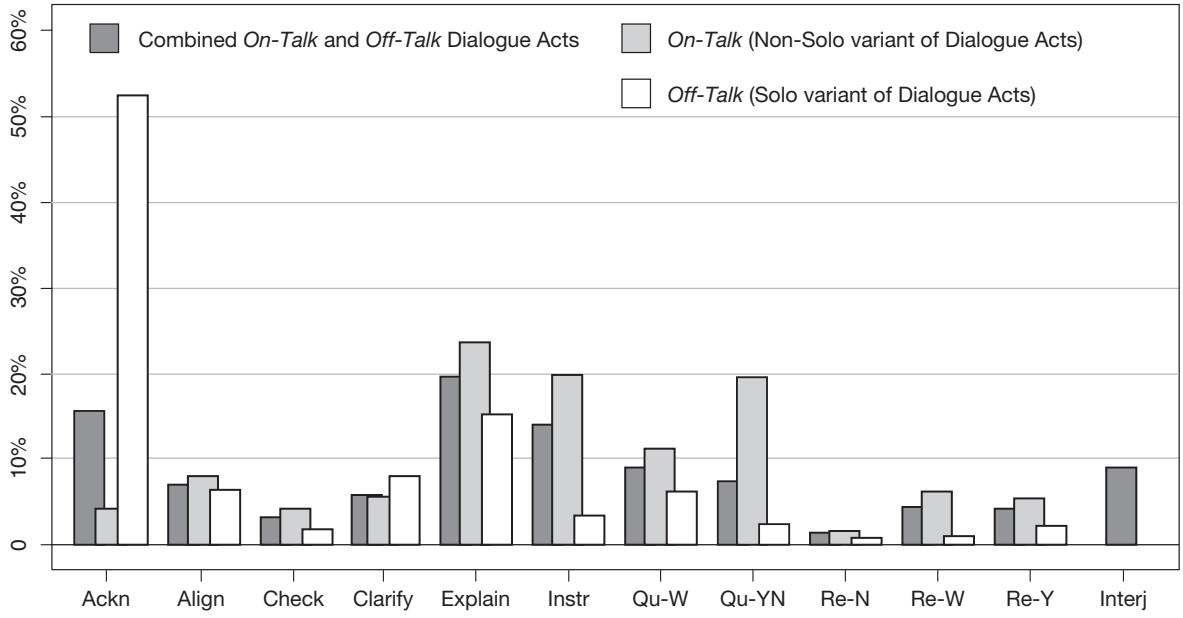


Figure 4.29: *Bar chart of dialogue acts: Proportion of On-Talk and Off-Talk variant's combined as well as separated, with the total of each variant being 100% — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item*

("w/ Video" and "w/o Video"), of the 490 *Off-Talk* "*Acknowledgement*" labels, 315 appear during the w/ Video setting and 175 appear during the w/o Video setting. Though the overall count is different, as a frequency of occurrence during each setting, the difference is only 13.80% of all the *Off-Talk* dialogue acts during the w/ Video setting and 11.17% during the w/o Video setting.

4.4.3.3 Facial Expressions [3 labels] and Head Movements [2 labels]:

Methodological Note

The Facial Expressions and Head Movements have not been used in this study.

Three facial expressions of "*Smile*", "*Laughter*", "*Surprise*", and two head movements of "*Feedback-nod*", "*Feedback-shake*" were annotated using the dedicated annotation tool ELAN (Wittenburg et al., 2006) by the two students that also transcribed the audio files. A description of each facial expressions and head movements that was presented to the student annotators is listed in Table 4.14.

These labels were based upon the "General face" and "Head movement" attributes of the MUMIN coding scheme (Allwood et al., 2007). Of the original "General face" attributes, "Scowl" and "Other" were removed and "*Surprised*" was added, since the ASR and MT results may not provide what the subject was expecting, and result in a surprised rather than a scowling expression. The "Head movement" attributes were summarised into "Feedback-nod" for vertical, and "Feedback-shake" for

Facial Expressions [3]	Description
Smile:	The subject is smiling.
Laughter:	The subject is laughing.
Surprise:	The subject has an expression of surprise.

Head Movements [2]	Description
Feedback-nod:	Head nod or nods related to acknowledgement/feedback.
Feedback-shake:	Shake/s related to rejection/feedback.

Table 4.14: *Description of facial expressions and head movements annotation labels*

horizontal head movements, following the expected ease of visual communication during the "w/ Video" recording setup mentioned in § 4.3.1: "Separation of Subjects".

■ Facial Expressions and Head Movements Workflow.

One of the labels mentioned in Table 4.14 for facial expressions or head movements were assigned to locations where the annotator noticed a facial expression or head movement based on visual cues. Therefore, the duration and location of the segments are not limited to, or directly related to each utterance. As with the perceived emotion and dialogue act labels, the ELAN files for each dialogue had a tier added for the segmentation of facial expressions or head movements. The five labels were also made selectable through a pull down menu once the segmented section was activated for input to minimise errors. The steps for the workflow are as follows:

1. Open the ELAN file for the dialogue that has been pre-linked to the video that includes both subjects and also the stereo audio file that was used to transcribe the dialogue.

2. Segment the data for locations that the student transcribers thought the subject was visibly in one of the five facial expressions or head movements labels ("*Smile*", "*Laughter*", "*Surprise*", "*Feedback-nod*", and "*Feedback-shake*") and select the label from the pull down menu as illustrated in Figure 4.30 for each segment.

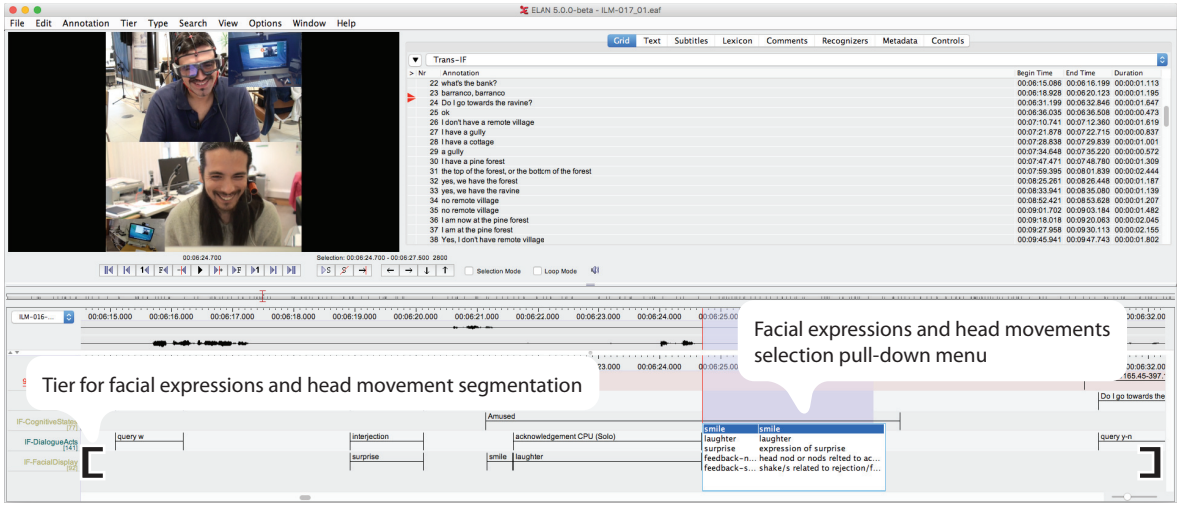


Figure 4.30: *Example of ELAN facial expressions and head movement annotation*

3. Once finished, the student annotators participation with the ILMT-s2s corpus creation was completed.[46]

■ Summary of Facial Expressions and Head Movements.

There are a total of 2,329 labeled facial expressions and head movements summarised as shown in Table 4.15. As previously mentioned, the labels of the facial expressions and head movements are not restricted to the utterance location, but freely applied within the dialogue.

Facial Expressions	Count	Total duration (approx.)
Smile	882	1h 11m 30s
Laughter	443	27m 26s
Surprise	339	15m 32s
Subtotal	1,664	1h 54m 28s
Head Movements	**Count**	**Total duration (approx.)**
Feedback-nod	301	10m 35s
Feedback-shake	364	12m 06s
Subtotal	665	22m 42s
Total	2,329	2h 17m 11s

Table 4.15: *Summary of facial expressions and head movements annotated in the ILMT-s2s corpus*

This thesis does not contain any study with these facial expressions and head movements annotation labels. Future work that looks into the different usage of head movements to indicate backchannel acknowledgements in the settings of "w/ Video" and "w/o Video" could be of interest since it has been reported that the subjects the ILMT-s2s corpus did not utter as many "Acknowledgement" dialogue acts as subjects in the direct Human-to-Human HCRC Edinburgh Map Task corpus (§ 8). This explanation of the annotation has been included in this thesis to provide a more detailed description of the labelling process than what is written in my LREC 2016 paper (Hayakawa, Luz, Cerrato, & Campbell, 2016).

4.4.3.4 Talk Types [4 lables]:

The Talk Type labels were not added by the student annotators, but retrieved from the XML log files of the ILMT-s2s System and the Final Cut Pro X Project exports. The definition of the four talk type labels of "*On-Talk*", "*Off-Talk*", "*Off-Talk Self*", "*Off-Talk Other*" are listed in Table 4.16.

Talk Types	Description
On-Talk:	Utterances spoken to the ILMT-s2s System.
Off-Talk:	Utterances that are not spoken to the ILMT-s2s System.
Off-Talk Self:	*Off-Talk* utterances that are directed to oneself.
Off-Talk Other:	*Off-Talk* utterances that are directed directly to to a fellow human.

Table 4.16: *Description of the talk type labels*

The studies related to the ILMT-s2s corpus are not the first to identify the phenomenon of *Off-Talk*.

[46] I would like to thank Tim Ryan and Cesar Sabato again for their work that has been greatly appreciated by myself.

Batliner et al. (2006) open their paper with an example from Shakespeare's *Hamlet*, where Hamlet seems to change his speaking style when addressing his interlocutor, to that of utterances which are spoken, but not directed towards his interlocutor. This shows that this is not a new phenomenon, but part of human nature that even Shakespeare expressed with his characters. However, the dialogues of the ILMT-s2s corpus are not that of a play like Shakespeare's *Hamlet*. It is a collection of task oriented computer mediated interlingual dialogues. The *Off-Talk* by the subjects of the ILMT-s2s corpus are fundamentally the same as the definition of *Off-Talk* provided by Oppermann et al. (2001, p. 1):

> **"We define Off-Talk to be every utterance that is not directed to the system as a question, a feedback utterance or as an instruction.**
>
> The following is an open list of possible utterances which could be considered as Off-Talk in human-machine-interaction:
>
> - soliloquy / thinking aloud
> - swearing
> - reading from displayed text aloud
> - conversation with other person(s) present
> - telephone conversation (e.g., with cellular phone)
> - extrinsic speech (e.g., video player, TV set, etc.)" (Oppermann et al., 2001, p. 1)

Oppermann et al. (2001) look into the type of *Off-Talk* by differentiating *Off-Talk* into two categories, "ROT" (Read Off-Talk) and "OOT" (Other forms of Off-Talk) which looks at the source of the utterance — Is the *Off-Talk* something that is read, or some other type of communication? For the ILMT-s2s corpus collection, *Off-Talk* is differentiated by the direction of the utterance — Is the subject talking to themselves or to someone else other than themselves and their interlocutor through the ILMT-s2s System?

■ The Talk Type Workflow.

Since the ILMT-s2s System used a "Push-to-talk" activation method, differentiating *On-Talk* and *Off-Talk* locations was a simple task of identifying when the ILMT-s2s System was activated. As for the different types of *Off-Talk*, the *Off-Talk Self* and *Off-Talk Other*, these differences were verified by watching the dialogues and determining the intended direction of the utterance. However, this simple act of identifying the time-stamps of the each *On-Talk* utterance needed the information retrieved from the ILMT-s2s System log and verification with the actual dialogue of the subjects. In this process, the ASR, MT and TTS event time-stamp information and results were also retrieved. During the verification process of this workflow, it became apparent that the time-stamp from the ILMT-s2s System was not uniform, with uneven disparity between logged items within the same dialogue as

illustrated in Figure 4.31. For this reason, the time-stamp information from the Final Cut Pro X Project XML export was used to obtain the main time-stamp while the information from the ILMT-s2s System log file was maintained to calculate the disparity for each item that was not retrievable from the Final Cut Pro X Project XML export file (the ASR, MT and TTS time-stamp information). The steps for the workflow are as follows:

1. Obtain the offset time from the Final Cut Pro Project XML export for each ILMT-s2s System XML log file as listed in Table 4.17.[47]

2. Add the offset time listed in Table 4.17 to each time-stamp within the ILMT-s2s System XML log to obtain the segmentation time in reference to the edited video and audio files that were used to transcribe and annotate the corpus data.

Dialogue	Duration	ILMT-s2s System XML Log Offset (seconds)				
01, IG	00:23:02	−149.52548				
01, IF	00:23:02	−74.64276				
02, IG	00:48:57	−0.11624	2261.30027			
02, IF	00:48:57	−848.57570	2228.82527			
03, IG	01:14:20	5.39359	1968.39613	2551.14246	3118.17747	3557.73962
03, IF	01:14:20	39.49046		2604.05801		3584.22240
04, IG	00:22:31	−145.17648				
04, IF	00:22:31	−131.63327				
05, IG	00:40:20	−62.83798				
05, IF	00:40:20	−84.19018				
06, IG	00:44:07	−68.27064				
06, IF	00:44:07	−52.05637				
07, IG	00:37:55	−1775.88918				
07, IF	00:37:55	−1444.50624				
08, IG	00:59:29	−234.05287		2859.05570		
08, IF	00:59:29	−332.49684	2625.78827	2846.69377		
09, IG	00:29:13	12.84692				
09, IF	00:29:13	−6.12690				
10, IG	00:36:07	−187.72095	917.06778	1578.70855	1578.65421	
10, IF	00:36:07	−139.77255	923.65065		1627.65647	
11, IG	00:43:17	−15.49799				
11, IF	00:43:17	−43.84483				
12, IG	00:48:19	−1976.81268				
12, IF	00:48:19	−1947.33588				
13, IG	00:26:13	28.82652				
13, IF	00:26:13	2.18499				
14, IG	00:25:59	−339.12836				
14, IF	00:25:59	−222.13687				
15, IG	00:19:38	−17.36963				
15, IF	00:19:38	−276.53389				

Table 4.17: *ILMT-s2s System XML log offset time*

3. Extract the segmentation start and end time, with the resulting text from the ILMT-s2s System XML log and save in csv format of; "Start time" – "End time" – "Text".

   The "Text" item output differed with each specific event.

   - For the activation of the ILMT-s2s System (when the subject used the system to talk to the interlocutor): the file name of the recorded utterance.

[47] Dialogues with multiple offset times indicate system crashes.

- For the ASR: the ASR result that were obtained from the Google Speech API.
- For the MT: the MT results that were obtained from the Microsoft Bing translation system.
- For the TTS: the MT result that was sent to the interlocutor's computer or the text received from the interlocutor's computer.

4. Import the csv files created in step 3 to ELAN and merge with the transcription files to verify if the segmentation of the imported csv files are correct or not.

   Unfortunately there was an irregular misalignment of the ILMT-s2s System activation and the transcribed segmentation as illustrated in a mild error sample of Figure 4.31. As mentioned,

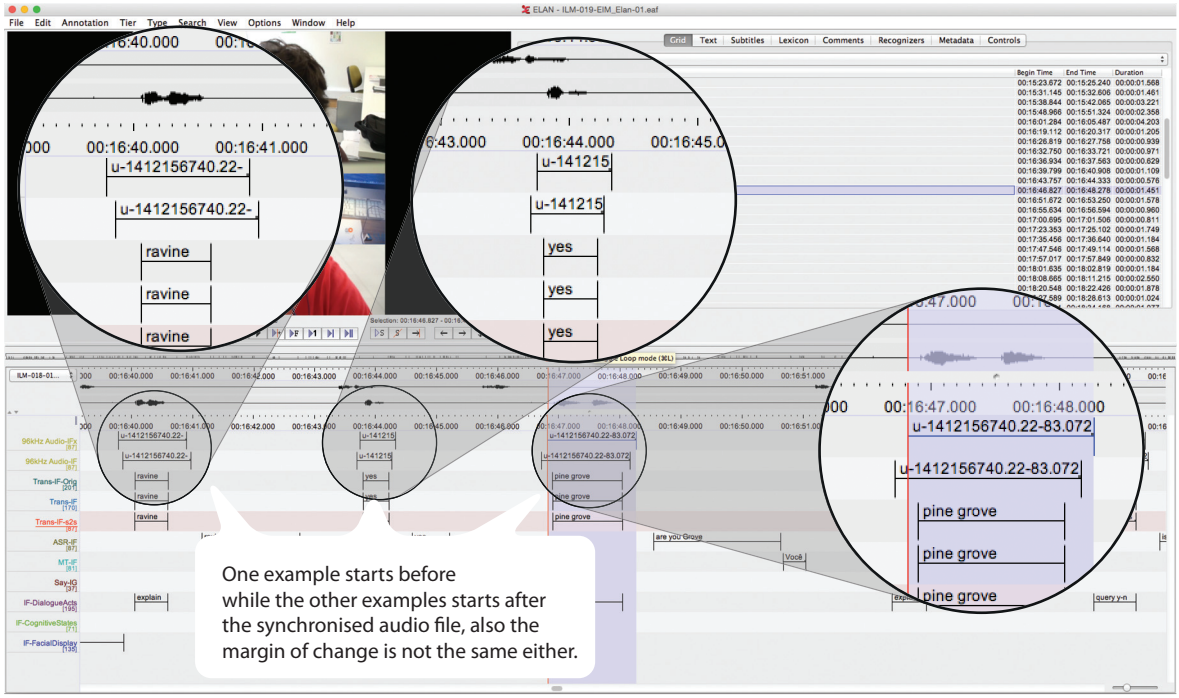


Figure 4.31: *Example of segmentation irregularity*

the sample of Figure 4.31 is a mild error. For extreme items the segmentation would start during the middle of the utterance. This was verified to be a false time stamp since contents of the audio recordings of the utterance differed and covered the whole utterance most of the time. As a countermeasure to this alignment error, the time-stamp information was therefore taken from the Final Cut Pro Project XML export since the utterance recordings of the system (*On-Talk* utterance recordings) were synchronised to the video files from the audio patterns and individually verified during the synchronisation process.

5. Using the segmentation details from the Final Cut Pro Project XML export, label the utterances that were directed to the ILMT-s2s System as *On-Talk*, while all other transcribed utterances were identified as *Off-Talk*.

6. Verify all *Off-Talk* utterances manually using ELAN.

   Using the video and audio, it was possible to label the *Off-Talk* as either *Off-Talk Self* or *Off-*

Talk Other, and also relabel *On-Talk* items that were not processed by the ILMT-s2s System. The relabeling of some *Off-Talk* items to *On-Talk* was possible by listening for the click from the mouse before the utterance and after the utterance thanks to the click-and-hold-to-talk feature of the ILMT-s2s System. Due to this relabeling, the *On-Talk* labels indicate the intention of the subject and not the actual success of the intention.

■ Summary of Talk Types.

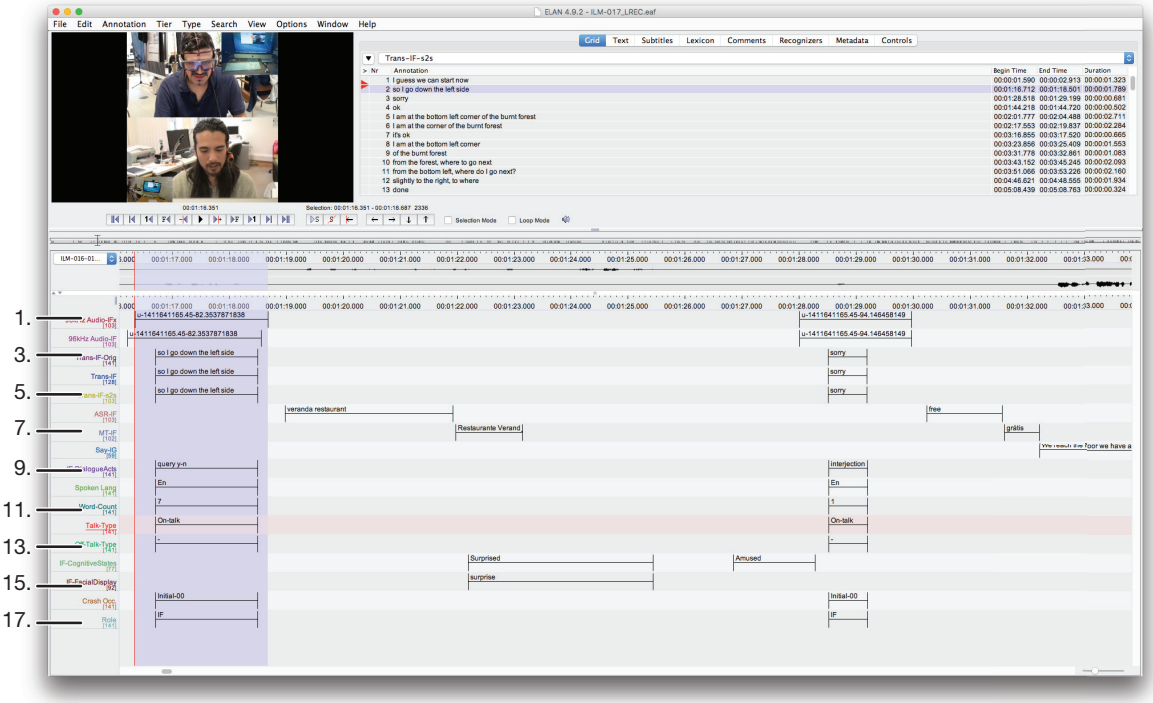
There are a total of 3,628 labeled talk types summarised as shown in Table 4.18. Since the talk type labels are linked one to one with the utterances of the ILMT-s2s corpus, the values in Table 4.18 also refer to the number of utterances in the corpus.

Talk Type	All	English	Portuguese
On-Talk	2,439	1,328	1,111
Off-Talk	1,189	652	537
Off-Talk Self	848	483	365
Off-Talk Other	341	169	172
Total	3,628	1,980	1,648

Table 4.18: *Summary of ILMT-s2s corpus Talk Types*

4.4.3.5 Summary of Annotation

The ILMT-s2s corpus consists of perceived emotion, dialogue act, and facial expressions and head movement annotations that were labeled by two student annotators who also transcribed the audio to text. Additionally as illustrated in Figure 4.32, to the labels that the students annotated, labels of talk type (*On-Talk*, *Off-Talk*, *Off-Talk Self*, and *Off-Talk Other*), segmentation information ASR, MT, TTS delivery and TTS output with the relative text has also been added as meta-data to the corpus.



1. ILMT-s2s System utterance recording file name - Segmentation from FCPX XML output
2. ILMT-s2s System utterance recording file name - Segmentation from ILMT-s2s XML log
3. Original transcriber transcription
4. Transcription of utterances to ILMT-s2s System and self
5. Transcription of utterances to ILMT-s2s System only
6. ASR result - Segmentation from ILMT-s2s XML log
7. MT result - Segmentation from ILMT-s2s XML log
8. TTS output from translation of interlocutor's utterance - Segmentation from ILMT-s2s XML log
9. Annotation of dialogue acts
10. Spoken language of subject
11. Word count of utterance
12. Talk type
13. Off-Talk type
14. Annotation of perceived emotion
15. Annotation of facial display
16. System crash location
17. Role of subject

Figure 4.32: *Representation of ELAN data files*

4.4.4 Calculating the Speech Rate

Methodological Note

Instead of using the "Words Per Minute (wpm)" or "Syllables Per Second (sps)" values, I use the value "Speech Rate Difference" in this study. The reason for this is explained in detail in Chapter 5.

The speech rate used in this study is calculated by creating a Text-to-Speech (TTS) system's output of all the human utterances in the dataset and using the TTS output duration as a reference to compare against the human output. The formula used to indicate speech rate is as follows:
"$Speech\ Rate = 1 - ((Subject\ Utterance\ Duration)/(TTS\ Utterance\ Duration))$"
This results in a value that indicates the difference between the subject's utterance speech rate and the TTS system's speech rate that is set at 180 wpm.

This method is novel, but related to other ideas in the literature. Sztahó, Kiss, & Vicsi (2015) use a similar method of comparing the syllable duration of a specific phrase, with previous recordings of the utterance by the same subject. Sztahó et al. compare like-for-like utterances of the same subject to determine if the subject's speech rate differs between each recording session, whereas I create an artificial like-for-like using the TTS system as a reference for the comparison.

■ Summary of the Speech Rate.

The speech rate difference was calculated on the ≈3,600 utterances of the computer-mediated dialogue dataset. A summary of the speech rate difference is listed in Table 4.19 and illustrated in Figure 4.33.

Speech Rate (%)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
ILMT-s2s	−1539.00	−58.76	−23.39	−37.92	1.86	83.30	74.14	3,628
ILMT-s2s (En)	−878.70	−52.43	−18.80	−28.71	7.56	83.30	58.04	1,980
ILMT-s2s (Pt)	−1539.00	−66.95	−29.12	−48.98	−4.41	58.64	88.50	1,648

Table 4.19: *Summary of subject utterance duration divided by TTS output duration* — "$1 - (S/T)$"

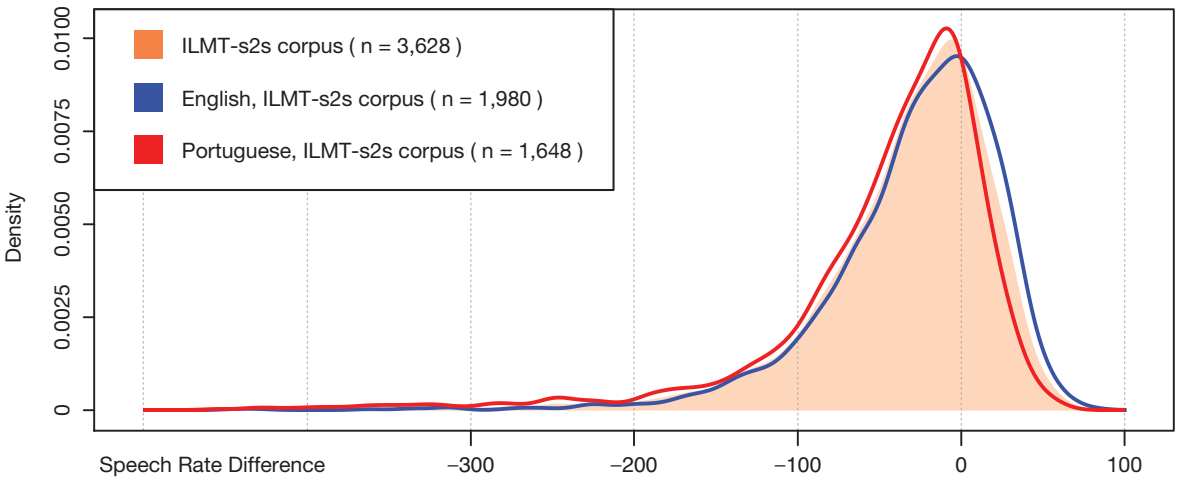


Figure 4.33: *Speech rate difference density plot of English and Portuguese dialogue data*

4.4.5 Calculating the ASR Performance (WER & *F*-Score)

The ILMT-s2s System's ASR performance was assessed by aligning the human transcriptions[48] (§ 4.4.2) and the ASR result transcript to compute the Word Error Rate (WER), Precision, Recall, and *F*-Score of each utterance (McCowan, Moore, et al., 2005).[49]

Word Error Rate (WER): Defined as the ratio of the Levenshtein distance between the aligned ASR result transcript of the utterances to the number of words in the reference human transcription. Figure 4.34 illustrates how it is calculated.

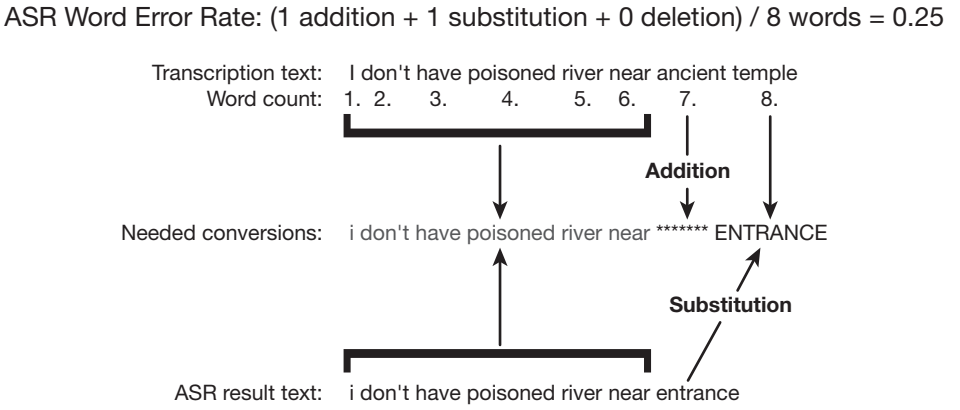


Figure 4.34: *Levenshtein distance between human transcription and ASR result*

Precision (*P*): The ratio of matching words in the alignment text to the number of words in the ASR result text. Figure 4.35 illustrates how it is calculated.

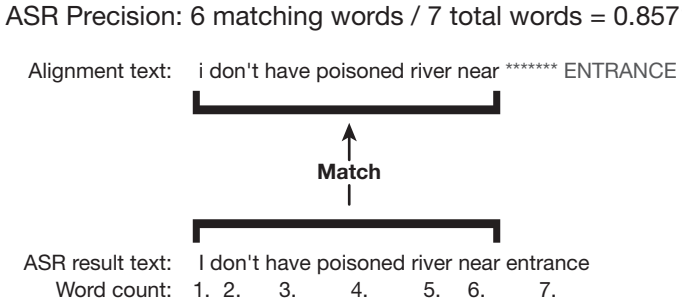


Figure 4.35: *Precision (P): Ratio of matching words in the alignment to the ASR result*

[48] Punctuation is removed from the transcription so as not to interfere with the alignment, and all letters are changed into small cap.

[49] The decision to use the "WER" and "*F*-Score" and the actual R code to calculate it was thanks to Dr. Saturnino Luz, while the act of creating the "Reverse *F*-Score" is the contribution I made.

Recall (R): The ratio of matching words in the alignment text to the number of words in the transcribed text. Figure 4.36 illustrates how it is calculated.

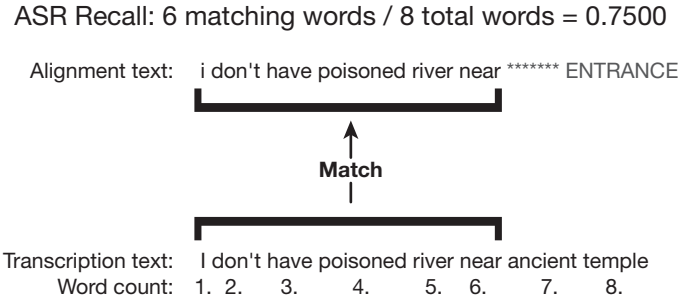


Figure 4.36: *Recall (R): Ratio of matching words in the alignment to the transcribed text*

***F*-Score:** The harmonic mean of Precision (P) and Recall (R) is reported as the F-Score. Figure 4.37 illustrates how it is calculated.

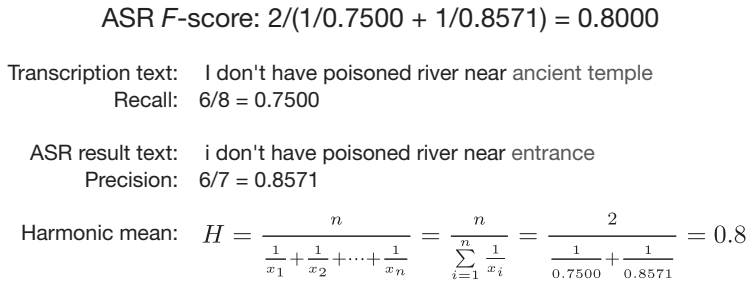


Figure 4.37: *F-Score (F_0): Harmonic mean of P and R*

Additionally, the value of a Reverse F-Score $(1 - F_0)$ was included for easier comparison with the speech rate — the higher the Reverse F-Score the more errors in the ASR result.

■ Summary of the ASR Performance Values (WER & F-Score).

Methodological Note

Instead of using the "F-Score" values straight forward, I use the "*Reverse* F-Score" in this study. This is for two reasons. (i) To keep the direction of the results the same between WER and F-Score values, with a low value indicating good ASR perfornamce and a high value indicating a poor ASR performance., and (ii) As with the Speech Rate Difference where zero (0) indicates the same speech rate as the TTS system output, by using the reverse F-Score, zero is kept as an indicator of the ASR performance being the same as the reference transcription.

The reverse F-Score is calculated by subtracting the F-Score from one (1) as in $(1 - F_0)$.

The ASR performance was calculated on the ≈2,300 *On-Talk*[50] utterances of the computer-mediated dialogue dataset. A summary of the WER and the Reverse F-Score are listed in Table 4.20 and Table 4.21.

[50] "*On-Talk*" are utterances directed to the subject's interlocutor via the ILMT-s2s System as explained in § 4.4.3.4.

ASR Performance	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
All utterances (ILMT-s2s)	0.00	0.00	0.25	0.41	0.67	10.00	0.55	2,370
English utterances	0.00	0.00	0.29	0.43	0.75	10.00	0.60	1,302
Portuguese utterances	0.00	0.00	0.25	0.38	0.54	8.00	0.48	1,068

Table 4.20: *Summary of computer-mediated dialogue WER values*

ASR Performance	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
All utterances (ILMT-s2s)	0.00	0.00	0.22	0.34	0.56	1.00	0.36	2,370
English utterances	0.00	0.00	0.25	0.37	0.67	1.00	0.38	1,302
Portuguese utterances	0.00	0.00	0.20	0.31	0.43	1.00	0.34	1,068

Table 4.21: *Summary of computer-mediated dialogue Reverse F-Score values*

The distribution of the ASR performance values for WER are illustrated as a density plot in Figure 4.38 and the *Reverse F*-Score values are illustrated as a density plot in Figure 4.39.

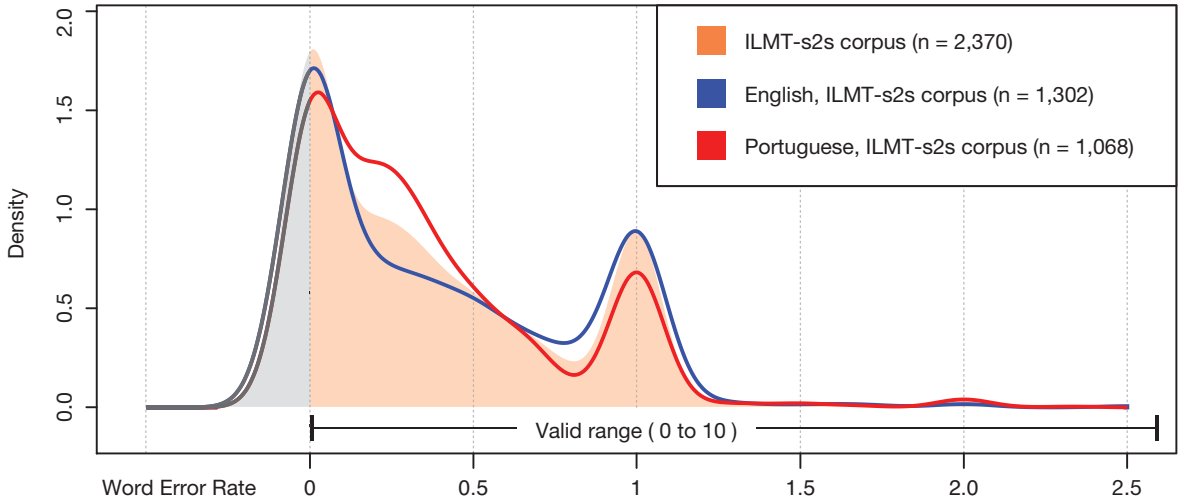


Figure 4.38: *WER: ASR performance density plot of Eng. and Portuguese dialogue data*

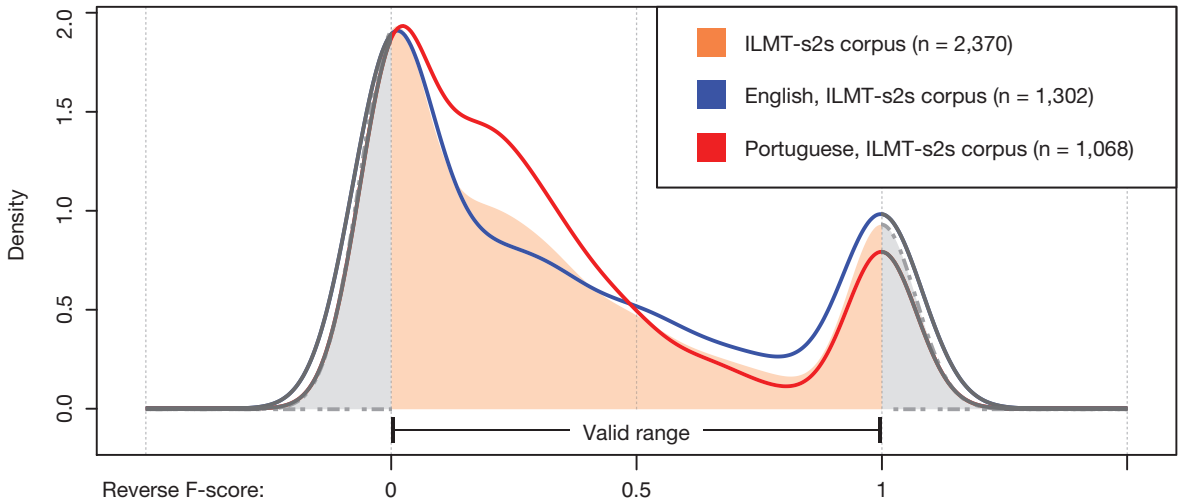


Figure 4.39: *Reverse F-Score: ASR performance density plot of Eng. and Portuguese dialogue data*

4.4.6 20 Frame/Window Normalization

It's good to have a large corpus of data, but it is also necessary to understand the characteristics of the corpus before going into any analysis. The preparations performed in this section provide a value, in the speech rate (§ 4.4.4: "Calculating the Speech Rate") and ASR performance (§ 4.4.5: "Calculating the ASR Performance (WER & *F*-Score)") that changes over the course of the dialogue. For this reason the simplest way to understand the data is to create a graph that displays the flow of change. However there is no point in visualising the data if the visualisation is not standardised and comparable to other dialogues with at least one parameter unified between all dialogues. The ILMT-s2s corpus consists of fifteen dialogues where subject pairs communicated over the internet to complete the map task. The dialogues for the corpus were not scripted and therefore differ in length of between approximately 20 minutes to approximately 1 hour 15 minutes, and an utterance quantity of between 76 *On-Talk* utterances and 262 *On-Talk* utterances as illustrated in the bar chart of Figure 4.40 and detailed in Table A.2 of Appendix A.

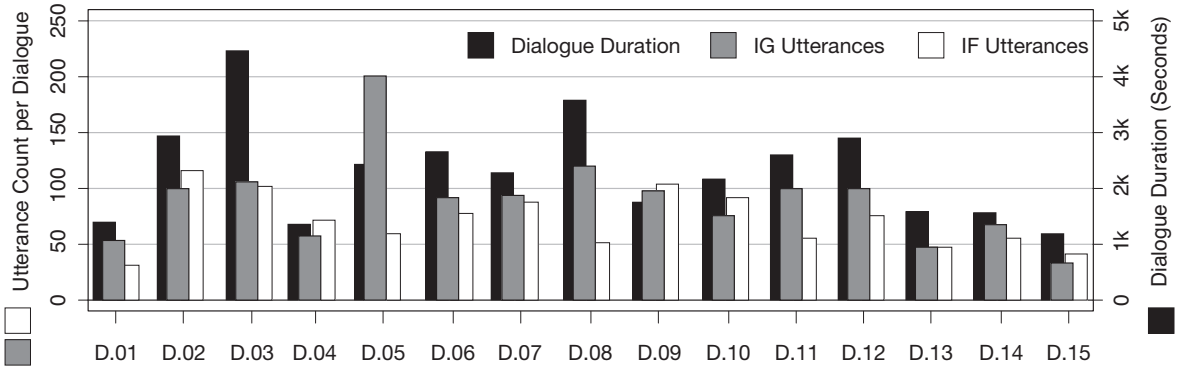


Figure 4.40: *ILMT-s2s corpus utterance count and duration comparison*

Since the emphasis of this study is to look into the speech rate and ASR performance, the difference in the duration of the dialogues is ignored, but the differing quantities of utterances directed towards the interlocutor need to be transformed into a comparable format. There may be better ways to unify the differing dialogue utterance quantities, but for this data each utterance value was duplicated 20 times (or equivalent), and then grouped by the number of utterances there were in the dialogue (or equivalent) as indicated in the seven examples in Table 4.22 so that the total utterance count would be a unified 20.

Dialogue number (En)		D.01	D.02	D.03	D.04	D.05	D.06	D.07
Utterance Count per Dialogue		33	100	106	73	201	93	89
Duplicate each value	×	20	1	10	20	20	20	20
Total number of values		660	100	1,060	1,460	4,020	1,860	1,780
Group values (windows)	÷	33	5	53	73	201	93	89
Number of window samples		20	20	20	20	20	20	20

Table 4.22: Sample window alteration example

Using the new window size (20 windows/samples per dialogue) for each dialogue, the mean value of all the subjects is plotted in Figure 4.41, where in the bottom (blue) curve the deviation in speech rate is observed:

- 0% indicates that the interlocutor's utterance duration is the same as the TTS.
- Positive percentage points indicate that the interlocutor's utterance is faster than the TTS.
- Negative percentage points indicate that the interlocutor's utterance is slower than the TTS.

Thus, the farther away from 0%, the faster (+ figures) or slower (− figure) the interlocutor's utterance.

The top (orange) curve represents a measure of the ASR performance accuracy (Reverse *F*-Score) — 0% if the ASR result is completely correct, and 100% if the ASR result is completely wrong.

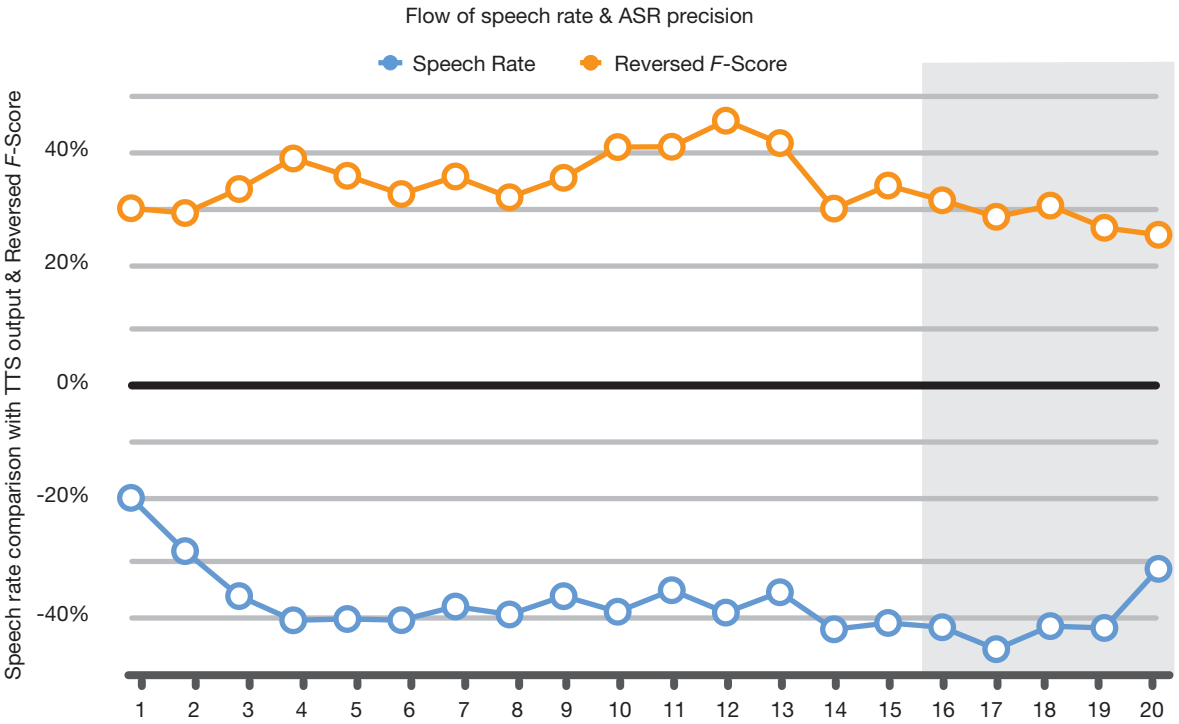


Figure 4.41: *Speech rate related and ASR accuracy for all subjects combined — Dialogue duration standardised into 20 frame windows*

	Speech Rate (%)					Reverse *F*-Score (%)				
	Min.	*Mdn*	*M*	Max.	*SD*	Min.	*Mdn*	*M*	Max.	*SD*
All dialogues	−185.80	−31.31	−37.61	28.59	33.64	0	29.77	34.16	100	23.85

Table 4.23: *Summary of 20 frame window dialogue data*

Plots for individual subjects can be found between Figure A.14 to Figure A.17 in § A.3 of Appendix A.

4.5 Summary

The ILMT-s2s corpus is a corpus collected with the intention of observing and exploring the adaptation of subjects to computer mediated interlingual communication. It is not a corpus that has been created in finely controlled environments that wishes to replicate specific characteristics of human communication.

The ILMT-s2s System that was created for the data collection of this corpus used a combination of three off-the-shelf components: Automatic Speech Recognition (ASR) from the Google Speech API, Machine Translation (MT) from Microsoft Bing translation services and the Text-to-Speech (TTS) from the Apple TTS system voices of Mac OS X operating system. The user interface of the ILMT-s2s System was designed following the simple structural interface of widely used messaging software to provide familiarity to the subjects but with the addition of an integrated push-to-talk Speech-to-Speech Machine Translation (S2S-MT) system. The ILMT-s2s System was installed on two computers in two different rooms and communicated via the general student wireless network provided by Trinity College Dublin, The University of Dublin. A microphone on each subject was used to record the subject's utterance to the ILMT-s2s System, while four camcorders recorded the video and audio of the whole session. Additionally eye gaze and biosignals (EEG, SC and BVP) were recorded from two separate devices on one of the participating subjects of the dialogue pair. The recorded files have been edited to remove recordings before and after the session and all video and audio files have been synchronised to a total of 9h 39m 27s of dialogue data. The data has also been transcribed and annotated for dialogue acts, perceived emotions, facial expressions and head movements.

This data, that I have called the ILMT-s2s corpus,[51] has been used, along with the data from the HCRC Edinburgh Map Task corpus (Anderson et al., 1991) to explore the research questions that were described in Chapter 1. The analysis of the data is presented in the next chapter.

Of the data explained in this chapter, the subject utterance duration, the ASR results, the perceived emotion annotation, the dialogue act annotation, and the user survey results have been used for the research presented in this thesis and the papers that I have published. The biosignal data collected in this corpus has been used to study the perceived emotions, and the three talk types of the subjects, but since the analysis of the biosignals was preformed by my colleague at the University of Dublin, Trinity College, Fasih Haider, they are not reported in this thesis.

[51] International Standard Language Resource Number (ISLRN): 100-610-774-625-0

Chapter 5

Speech Rate Difference

This chapter explains the details of why the speech rate difference was used in this study instead of the more orthodox speech rate measurements such as "*Word-Error-Rate*" and "*Syllables-Per-Second*".

5.1 Introduction

Literature has indicated that subjects adapt their speech to their interlocutor, even when the interlocutor is an artificial agent (L. Bell et al., 2003; Oviatt, Darves, & Coulston, 2004; Suzuki & Katagiri, 2007). Oviatt et al. (1998, p. 473), reported in her study that when users were handling dialogue system errors the users would try using what they thought of as more comprehensive speech. The subjects hyperarticulated, with slower speech. A predictable hypothesis from the literature would be that the subjects would align their speech rate with the ILMT-s2s System, once an error occurs from the ASR, the subject would reduce their speech rate and probably not return to their original speech rate, based on the report from Branigan et al. (2010, p. 2366);

> **"**We hypothesize that people may not feel confident in their ability to make accurate judgements about the capability of a computer, and might therefore tend to adopt a conservative initial model of the computer. Having adopted such a model, they might be reluctant to update it dynamically, or they might require substantial evidence of the computer's abilities before updating it.**"** (Branigan et al., 2010, p. 2366)

For this reason, the ASR performance (WER) of the ILMT-s2s System, along with the subject speech rate were thought to be an important aspect to study. However, the orthodox methods of measuring the speech rate have concerns that I have listed below.

5.1.1 Methodological Reason to Using Speech Rate Difference

The speech rate difference method (§ 5.5) of verifying if the subject was speaking quickly or slowly was chosen over word per minute (wpm) or syllables per second (sps) for the following three reasons:

1. To obtain a reliable reference of speech rate by comparing the subject utterance with a stable output reference.

2. To verify if the subjects align their speech rate to the communication tool's (the ILMT-s2s System's) output speech rate.

3. To compare the speech rate of non-mediated dialogue and computer-mediated dialogue.

This method is novel, but related to other ideas in the literature. Sztahó et al. (2015) use a similar method of comparing the syllable duration of a specific phrase, with previous recordings of the utterance by the same subject. Sztahó et al. compare like-for-like utterances of the same subject to determine if the subject's speech rate differs between each recording session, whereas I create an artificial like-for-like using the TTS system as a reference for the comparison.

5.1.2 Practical Reason to Using Speech Rate Difference

The utterances of the computer-mediated dialogues (the ILMT-s2s corpus) were expected to be low in word count and short in duration from the pre-data collection test runs, and hence, make it difficult to get a meaningfull speech rate value. This expectation materialised in the computer-mediated dialogue dataset, and is also present in the non-mediated dialogue data (the HCRC Edinburgh Map Task corpus).

Non-mediated dialogues: *Mdn* word count: 3 words, *Mdn* duration: 0.8 sec

Computer-mediated dialogues: *Mdn* word count: 4 words, *Mdn* duration: 1.5 sec

For details refer to Table 5.1.

Word Count	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	1	1	3	5.572	7	56	6.52	3,790
ILMT-s2s	1	2	4	5.168	7	179	6.01	3,628
ILMT-s2s (En)	1	1	4	4.919	6	179	6.76	1,980
ILMT-s2s (Pt)	1	2	4	5.466	8	42	4.97	1,648
Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	0.048	0.349	0.797	1.559	1.830	21.380	2.06	3,790
ILMT-s2s	0.097	0.748	1.493	2.244	2.907	77.720	2.75	3,628
ILMT-s2s (En)	0.097	0.651	1.285	1.939	2.404	77.720	2.96	1,980
ILMT-s2s (Pt)	0.136	0.893	1.874	2.610	3.509	26.360	2.49	1,648
Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	21.61	171.10	225.70	238.50	289.60	1263.00	95.12	3,790
ILMT-s2s	15.24	107.80	141.00	153.40	186.40	618.60	65.28	3,628
ILMT-s2s (En)	16.61	119.20	155.30	167.70	205.00	618.60	68.75	1,980
ILMT-s2s (Pt)	15.24	97.52	127.90	136.20	166.00	459.60	56.26	1,648

Table 5.1: *Summary of word count, duration and wpm*

With such a low word count and short duration for each utterance, it is difficult to illustrate and represent the speed in which each utterance was uttered using the orthodox method of wpm as explained below (subsection 5.1.3).

5.1.3 The Dataset from the Perspective of *Words per Minute*

A summary of the speed that each utterance in the computer-mediated dialogue dataset was uttered is reported in Table 5.2 using the measure of wpm.

Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	21.61	171.10	225.70	238.50	289.60	1263.00	95.12	3,790
ILMT-s2s All	15.24	107.80	141.00	153.40	186.40	618.60	65.28	3,628
ILMT-s2s En	16.61	119.20	155.30	167.70	205.00	618.60	68.75	1,980
ILMT-s2s Pt	15.24	97.52	127.90	136.20	166.00	459.60	56.26	1,648

Table 5.2: *Summary of wpm per utterance*

Non-mediated dialogues: The utterances are indicated to have been spoken as slowly as 15.24 wpm and as fast as 618.60 wpm with an arithmetic mean speed of 153.4 wpm and a *SD* of 65.28. Of this variance, 60% of the utterances were reported to have been uttered within a 100 wpm range between 100.03 wpm (20^{th} percentile) and 200.00 wpm (80^{th} percentile).

Computer-mediated dialogues: The utterances are indicated to have been spoken as slowly as 21.61 wpm and as fast as 1,263.00 wpm with a mean speech rate of 238.50 wpm and a *SD* of 95.12. Of this variance, 60% of the utterances were reported to have been uttered within a range of ≈150 wpm between 161.21 wpm (20^{th} percentile) and 308.26 wpm (80^{th} percentile).

These wpm values indicate that the subject uttered their speech relatively slowly in the computer-mediated dialogue data while the subjects speech rate is relatively fast in the non-mediated dialogue dataset. However, it is not clear yet if the sample utterances was intentionally spoken slowly or fast, or just producing deceptive values.

5.1.4 The Point of Concern with *Words per Minute*

When the orthodox way of calculating the Words Per Minute (wpm) as $(W/T \times 60)$, where W is the word count per utterance and T is the duration of the utterance in seconds, is used, the imbalanced nature of the spoken words would create a variance that is difficult to interpret.

One example of this is the different wpm value of the three single word utterances, “Perfect”, “Yes” and “Okay” that are indicated in Table 5.3. The maximum wpm value for the word “Perfect” is 130.40 wpm, but did the subject utter this word at the same intentional speed as the median wpm value for “Yes” at 138.90 wpm or “Okay” at 139.90 wpm?

Unlike measurements for traveling distance speeds of kilometres per hour (kph) or miles per hour (mph) where 2 items are fixed, the length of distance that is measured as kilometres or miles, and the duration of time that are measured as hours, the measurement of wpm only has 1 fixed item, the durations that is measured in minutes. The measurement of a word is not a fixed length — one might say that this variable length in words is a characteristic of the “stress-timed” English and Portuguese languages, but the variability in duration has also been found in a theoretically isochronous “syllable-timed” language such as Japanese too (Arai & Greenberg, 1997). Unlike the wpm measurement for

Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
Perfect	0.460	0.485	0.543	0.535	0.577	0.611	0.06	5
Yes	0.327	0.392	0.432	0.460	0.509	0.821	0.10	49
Okay	0.097	0.345	0.429	0.456	0.535	1.081	0.18	83
Pebbled shore	0.994	1.087	1.180	1.132	1.201	1.222	0.12	3
Go down	0.882	0.902	0.926	0.976	0.988	1.219	0.13	6
Then where?	0.536	0.542	0.547	0.554	0.564	0.5800	0.02	3
Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
Perfect	98.20	104.00	110.50	113.40	123.70	130.40	13.46	5
Yes	73.08	117.90	138.90	135.80	153.10	183.50	25.23	49
Okay	55.50	112.20	139.90	154.00	173.90	618.60	75.16	83
Pebbled shore	98.20	99.95	101.70	106.90	111.20	120.70	12.12	3
Go down	98.44	121.60	129.60	124.50	133.10	136.10	14.04	6
Then where?	206.90	213.10	219.40	216.70	221.60	223.90	8.80	3

Table 5.3: *Summary of single and two word subject utterance duration and wpm sample*

typing speed which often classifies one word as five letters, the amount of time it takes to utter the single word is different and varies.

It is not possible to know if the 130.40 wpm value for the single word utterance of "Perfect" is fast or slow when compared with the given example of "Yes" or "Okay", as it is possible with kph/mph where the numerator as well as the denominator is of a fixed measurement. This is not just a problem of single word utterances, but also in multiple word utterances as indicated in Table 5.3 with the different wpm values of the three multi word utterances.

The median wpm values for "Pebbled shore" (an item on the map), "Go down" and "Then where?" are 101.70 wpm, 129.60 wpm and 219.40 wpm respectively. These wpm values would indicate that on average, the first is spoken extremely slowly, the second quite slowly and finally, the last, quite quickly. However, there is no way of knowing if the difference in the wpm values of these three multi word utterances and even the single word utterances indicated in Table 5.3 are the result of a speech rate difference (intentionally or unintentionally changing the speed of speaking) or the physical difference in the duration to pronounce the given word. The bar chart in Figure 5.1 provides no more information

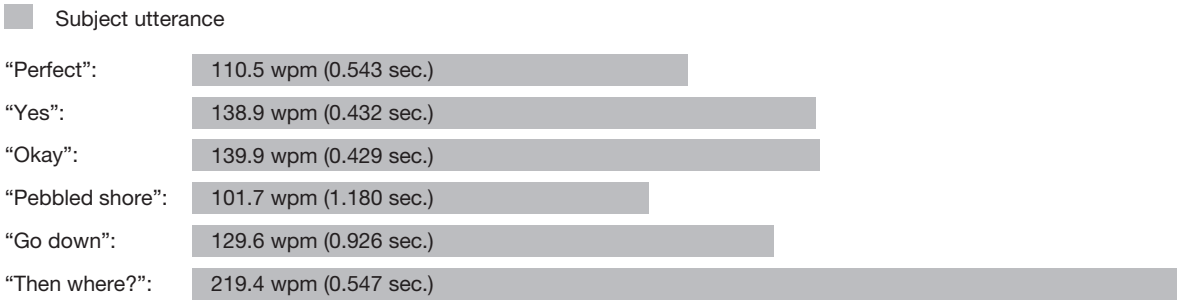


Figure 5.1: *Mdn wpm for the six provided samples — Subject utterance*

than five words were spoken slowly, between 101.7 wpm and 139.9 wpm, and one was spoken quite quickly at 219.4 wpm.

Wpm as a reference of speech rate is already an average of word combinations within the utterance of a minute. The random mixture of words with long and short duration are mixed into an utterance that is then calculated into a quantitive measure. Though the frequency of word duration has been

previously investigated in written language (Zipf, 1945) and also in spoken language (Greenberg, 1999; Batliner et al., 2001; A. Bell et al., 2002; A. Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), given the varying short sentences, in the unscripted speech of the dataset used in this study, the irregular duration of a word used in the short phrases would not provide an accurate wpm value to provide a reliable quantitive measure, due to the different word-frequency distributions. If the window were a longer period of time, such as the first quarter of the dialogue, there might be enough words to balance out the variance, but not in this situation where the window is a single utterance.

5.1.5 The Point of Concern with *Syllables per Second*

The method of using *syllables per second* as a measure may be more reliable than wpm, however this would require highly trained annotators to phonologically segment the data. With this comes a new issue of whether the utterance is annotated as indicated in the dictionary or following the actual phonetic sound that was uttered.[1] A fully unsupervised method of calculating the speech rate estimation based on syllable nuclei detection was presented by de Jong & Wempe (2009) as a Praat (Boersma & van Heuven, 2001) script, however I have yet to test this method on the computer-mediated dialogue data to verify that a similar reliability can be obtained. However, analysis of the *syllables per second* may not provide any better results since an analysis of the Switchboard corpus data (Godfrey, Holliman, & McDaniel, 1992) which is a corpus of a collection of natural and spontaneous telephone conversations, by Greenberg (1999, p. 167) indicated the following:

> "Although only 22% of the Switchboard lexicon is composed of monosyllabic forms, approximately 80% of the corpus tokens are just one syllable in length [...]. The portion of the lexicon consisting of three or more syllables (38%) is rarely exhibited in spontaneous language, accounting for less than 5% of the spoken instances [...]." (Greenberg, 1999, p. 167)

This would indicate that even using the measure of syllables per second, the length of the dialogues in the computer-mediated dialogue dataset may be too short and the problem identified with the wpm measure may also continue to persist.

5.1.6 Other Concerns

Another concern about the usage of wpm as a value to indicate the speech rate in my studies can be noticed in the paper by Yuan, Liberman, & Cieri (2006, p. 542). In their study of the Switchboard corpus (Godfrey et al., 1992) and the Chinese CallHome and CallFriend corpus (Yuan & Jurafsky, 2005) the utterances speech rates between one and seven words change drastically:

[1] Greenberg (1999, p. 163) reported eighty different pronunciations of the word "and" in a study of the Switchboard Transcription Corpus (Greenberg, 1997).

> "We can see that in both English and Chinese, there is an abrupt rise of speaking rate for the segments containing from one to seven words. For the segments having eight to about 30 words, however, the speaking rate stays level. And then, especially in English, the speaking rate rises again, but with a more gradual slope." (Yuan et al., 2006, p. 542)

This would indicate that utterances with a word count of one to eight words could not be compared reliably using the wpm value and that only utterances with the same word count can be compared within that range.[2]

5.2 Dataset

In this chapter, the transcription files (including segmentation time-stamps) of the non-mediated dialogue dataset described in § 3.2.5 and the transcription files (including segmentation time-stamps) of the computer-mediated dialogue dataset described in § 3.3.3 of Chapter 3 are used for the calculation.

5.3 Preparations for Calculating the Speech Rate Difference

Since the main objective of my studies was to investigate the alignment or adaptation of the subject to the ILMT-s2s System, it was perceived that a comparison with the ILMT-s2s System's TTS speech rate would be adequate as it would provide a one-to-one reference point with the utterances of the computer-mediated dialogue dataset.

The rationale behind this idea was to recreate every utterance that was uttered by the subjects of the computer-mediated dialogue dataset, using the same TTS output system used in the ILMT-s2s System and compare like-for-like utterance durations between the human subject and the ILMT-s2s System's TTS voice. This method would theoretically remove the variance in the resulting value that is created by the different duration in pronouncing words of different lengths, since the reference will also be pronouncing the same word — therefore, creating a more stable speech rate value.

To compare the speech rate, the transcription text of all speakers was output to a plain text file[3] that was then read out by the TTS system at a speech rate setting of 180 wpm and saved as an audio file. Using Praat, the audio file was segmented manually, using the transcription from the plain text file to provide the boundaries of the TTS output of each utterance, as illustrated in Figure 5.2. Once completed, the duration of each TTS output reference was calculated from the start and end times of the segmented audio files.

Non-mediated dialogues: *Mdn* duration: 0.8 sec (*Mdn* word count: 3 words)

Computer-mediated dialogues: *Mdn* duration: 1.3 sec (*Mdn* word count: 4 words)

For details of the duration and wpm of each TTS output refer to Table 5.4.

[2] This concern is verified in § 7.6 of Chapter 7

[3] All utterances within the transcription file were used with each utterance separated with a comma since the TTS

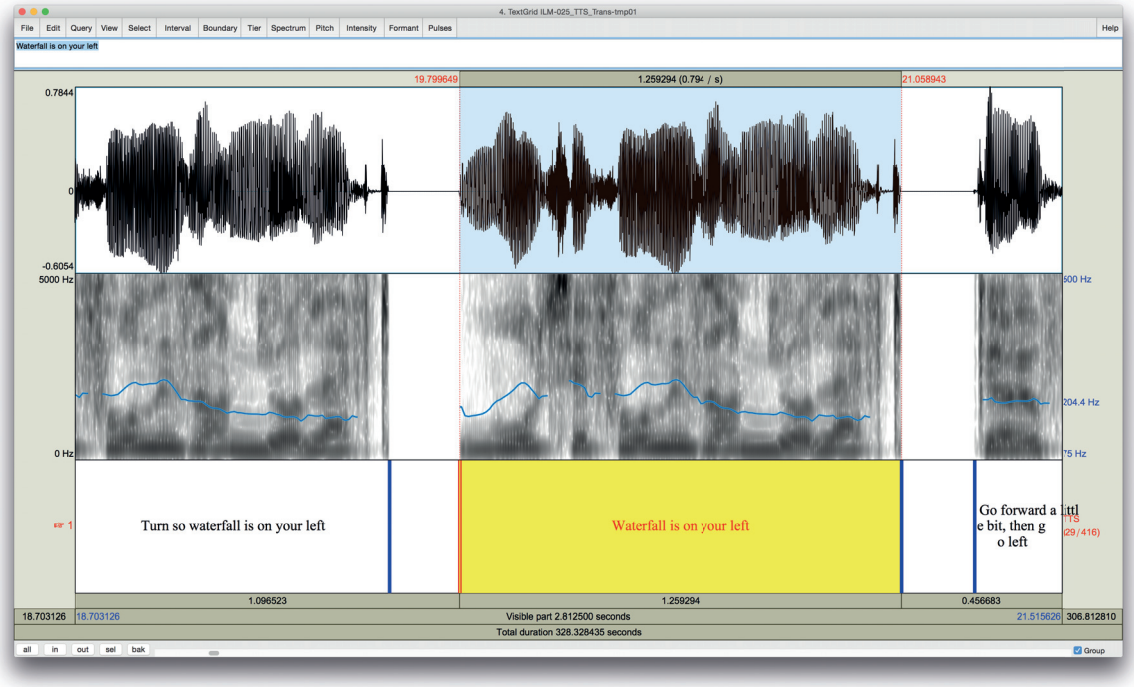


Figure 5.2: *Transcription image from Praat*

Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	0.121	0.432	0.905	1.481	1.924	15.040	1.53	3,790
ILMT-s2s All	0.112	0.597	1.250	1.694	2.250	56.500	1.90	3,628
ILMT-s2s En	0.166	0.578	1.124	1.508	1.888	56.500	2.03	1,980
ILMT-s2s Pt	0.112	0.653	1.491	1.917	2.622	14.240	1.69	1,648
Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	71.68	143.70	204.00	199.90	245.30	495.90	60.29	3,790
ILMT-s2s All	45.09	144.40	179.40	185.50	220.70	536.10	56.79	3,628
ILMT-s2s En	45.82	147.80	190.90	192.90	232.10	390.20	58.50	1,980
ILMT-s2s Pt	45.09	142.10	170.10	176.60	199.10	536.10	53.33	1,648

Table 5.4: *Summary of duration and wpm of the subject utterances output by the TTS system*

5.3.1 The Dataset from the Perspective of the TTS Output

As with the subject utterances wpm values in Table 5.2, the TTS output data show a big variance between the slow output measures and the fast output measures, but with a smaller *SD* when compared to the subject utterance measures.

Non-mediated dialogues: The utterances are indicated to have been output as slowly as 71.68 wpm and as fast as 495.90 wpm with an arithmetic mean speed of 199.9 wpm and a *SD* of 60.29. Of this variance, 60% of the utterances were reported to have been uttered within a 100 wpm range between 139.86 wpm (20^{th} percentile) and 254.28 wpm (80^{th} percentile).

Computer-mediated dialogues: The utterances are indicated to have been spoken as slowly as 45.09 wpm and as fast as 536.10 wpm with a mean speech rate of 185.50 wpm and a *SD* of 56.79. Of this variance, 60% of the utterances were reported to have been uttered within a range of ≈100 wpm between 137.57 wpm (20^{th} percentile) and 231.69 wpm (80^{th} percentile).

A linear regression model of the two wpm values (human utterances & TTS output) highlights and

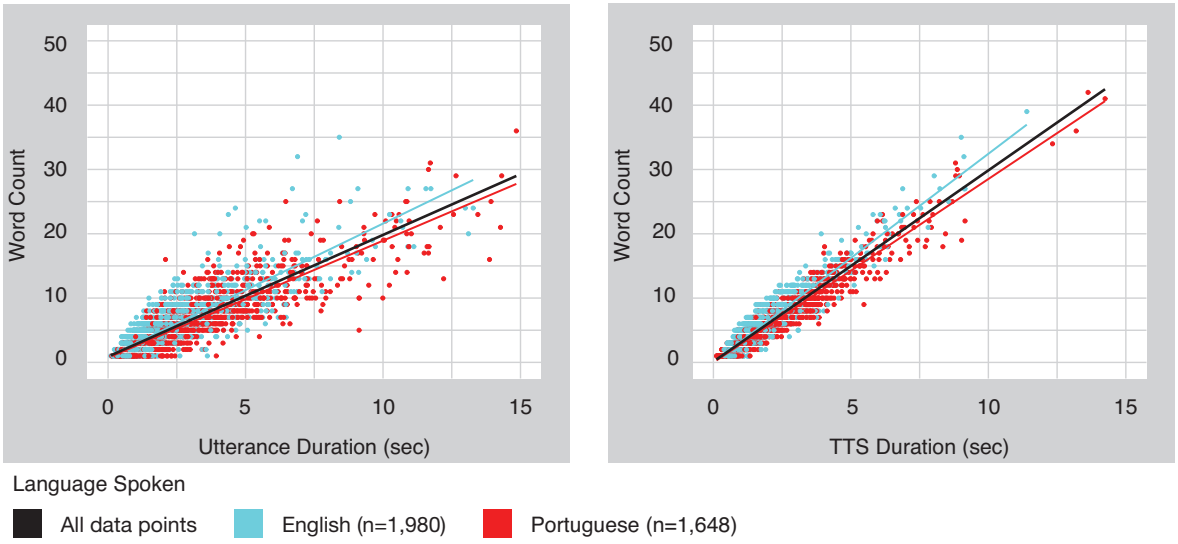


Figure 5.3: *Scatter plot of ILMT-s2s corpus utterance durations by word count*

confirms the differences between the two speech rates within the computer-mediated dialogue data (Figure 5.3). The regression equation of the word count and the duration for the subject utterance is "$(Word\ Count) = 0.61648 + 2.02693 * (Utterance\ duration)$", with an *Adjusted* R^2 value of 0.858 ($p < 2.2e-16$) and the regression equation for the TTS output is "$(Word\ Count) = -0.06870 + 3.08959 * (TTS\ output\ duration)$", with an *Adjusted* R^2 value of 0.948 ($p < 2.2e-16$).

There are two interesting details from these results that illustrate the data: (i) the slope difference of 1.062 which pridicts that the speech rate of the TTS output will be ≈ 30% faster than the subject utterance, and (ii) the 0.948 *Adj* R^2 value that indicates there is a variance in the speech rate of the TTS output since it is not the maximum value of 1.0, but is still smaller than that of the subject

system did not recognise a full stops as a pause in the utterance.

utterances. These two points are retrievable from the summary of the data in Table 5.2, and Table 5.4, however the linear regression equation and Figure 5.3 represents this more clearly. Although this identifies the speech rate difference and variation difference of the two measurements of subject utterance speed and TTS output speed, it doesn't help clarify the variable speech rate differences of the six single and multiple word examples provided in Table 5.3. The TTS output summary for the six utterance examples are displayed in Table 5.5, and Figure 5.4.

Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
Perfect	0.478	0.479	0.480	0.482	0.485	0.488	0.00	5
Yes	0.522	0.527	0.527	0.528	0.529	0.534	0.00	49
Okay	0.410	0.466	0.481	0.505	0.582	0.590	0.05	83
Pebbled shore	0.789	0.790	0.790	0.790	0.790	0.790	0.00	3
Go down	0.562	0.562	0.563	0.563	0.564	0.565	0.00	6
Then where?	0.471	0.472	0.473	0.472	0.473	0.473	0.00	3
Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
Perfect	122.80	123.70	124.90	124.50	125.40	125.50	1.16	5
Yes	112.40	113.40	113.90	113.70	113.90	114.90	0.61	49
Okay	101.70	103.10	124.60	120.20	128.70	146.20	12.31	83
Pebbled shore	151.8	151.8	151.8	151.9	151.9	152.0	0.12	3
Go down	212.3	213.0	213.4	213.2	213.6	213.6	0.51	6
Then where?	253.5	253.5	253.5	254.0	254.2	254.8	0.73	3

Table 5.5: *Summary of single and two word TTS output duration and wpm sample*

The main difference that can be observed when comparing the one and two word example phrases of the subject utterances and TTS output in Table 5.5, is again the reduction in the variance, the *SD* values. This reduction indicates the regularity in which the phrase is repeatedly output by the TTS system. This is an effect that is completely expected, since it is the exact reason that the TTS output was used: to provide a standardised, regular speech rate. However the question of which phrase is spoken fast or slow is not identifiable from this data either. The wpm values or the bar chart in Figure 5.4 do not provide any better understanding of the speech rate of short utterances on its own.

Figure 5.4: *Mdn wpm for the six provided samples — TTS output*

5.4 Calculating the Speech Rate Difference

5.4.1 Creating the Speech Rate Difference

Taking advantage of the regular speech rate of the TTS output, the TTS output duration can be used as a reference that the subject utterance is compared against. By dividing the subject utterance duration by the TTS output duration, a value that identifies the difference from the regular speech rate obtained by the TTS output can be retreived with an equation of (S/T), where S is the duration of the speaker's utterance and T is the duration of the TTS output.

"*Speech Rate* = (*Subject Utterance Duration*)/(*TTS Utterance Duration*)"

The resulting values of this measurement method are reported in Table 5.6.

Speech Rate Difference	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	0.178	0.673	0.844	0.939	1.089	6.472	0.44	3,790
ILMT-s2s All	0.167	0.982	1.233	1.379	1.588	16.400	0.74	3,628
ILMT-s2s En	0.167	0.925	1.188	1.287	1.525	9.759	0.58	1,980
ILMT-s2s Pt	0.414	1.044	1.291	1.490	1.670	16.400	0.88	1,648

Table 5.6: *Summary of subject utterance duration divided by TTS output duration — "S/T"*

The range between zero (0) and one (1) are values from utterances that were spoken faster than the TTS output and the range of one (1) and above are utterances that were spoken slower than the TTS output. The *Mdn* value being slower than the TTS output can also be confirmed from the slope of the regression equation of "*Subject utterance duration* = 0.239752+0.648037 *TTS output duration*" with an *Adjusted* R^2 value of 0.883 ($p < 2.2e{-}16$) as indicated in the graph on the right in Figure 5.5. With the slope being smaller than one (1), the regression equation is indicating that the subject utterance duration is not shorter than the TTS output duration, and therefore the majority of subject utterance are slower than the TTS output. This is evident from the lack of values within the pink coloured out region of Figure 5.5, which covers the area of a slope greater than one (1).

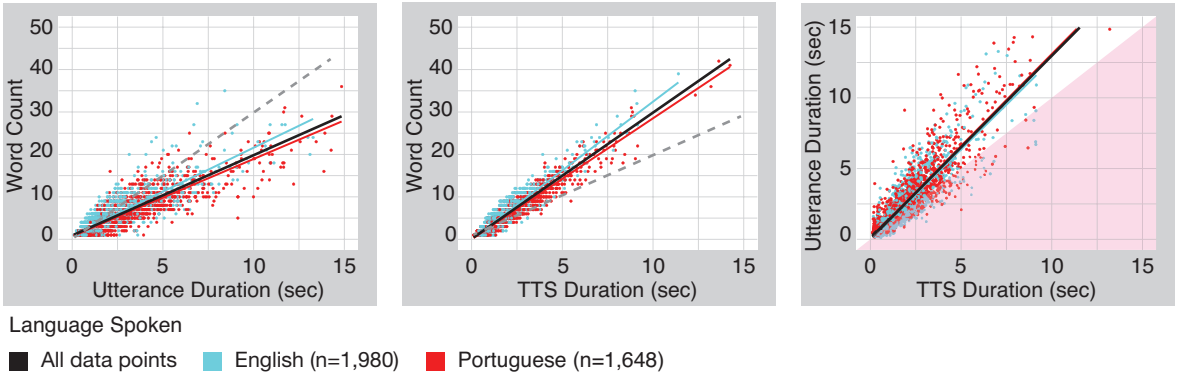


Figure 5.5: *Scatter plot of word count by subject utterance duration, word count by TTS output duration, and subject utterance duration by TTS output duration*

5.4.2 Simplifying the Speech Rate Difference

Unfortunately understanding the figures of Table 5.6 requires a moment of mental arithmetic and it is not immediately identifiable from just a glance at the results if the utterance was spoken quickly or not. To make the measurement values immediately evident, one (1) was subtracted from the resulting value to differentiate utterances that were faster or slower than the TTS output by using zero (0) as a reference division line, and the value was also multiplied by minus one (−1) to display utterances spoken slower than the TTS output as a negative value and faster utterances as a positive value. The resulting values of this conversion are summarised in Table 5.7.

Speech Rate Difference (%)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC	−547.20	−8.94	15.65	6.06	32.69	82.25	43.73	3,790
ILMT-s2s All	−1539.00	−58.76	−23.39	−37.92	1.86	83.30	74.14	3,628
ILMT-s2s En	−878.70	−52.43	−18.80	−28.71	7.56	83.30	58.04	1,980
ILMT-s2s Pt	−1539.00	−66.95	−29.12	−48.98	−4.41	58.64	88.50	1,648

Table 5.7: *Summary of subject utterance duration divided by TTS output duration — "*$1-(S/T)$*"*

5.5 Definition

Placing the priority on the simplicity of understanding the resulting value, the formula used to indicate speech rate that is used in this thesis is as follows:

"Speech Rate = 1 − ((Subject Utterance Duration) / (TTS Utterance Duration))"

This formula results in a value that indicates the difference between (i) the subject's utterance speech rate and (ii) the Text-to-Speech (TTS) system's speech rate that is set at 180 wpm. Obvious from the formula above, the frame of the value is one utterance and a positive value indicates that the utterance is faster than the TTS system and a negative value indicates that it is slower.

5.6 Conclusion

Using the equation "$1-(S/T)$" the results becoming more intuitive, and immediately understandable for the author as illustrated in Figure 5.6.

Using this method of comparing the subject utterance duration and the TTS output duration, the meaning that was lacking in the wpm values in Table 5.3, and Table 5.5, can now be presented with a more comprehensive value — the durational difference between the subject utterance and a regular speech rate of the TTS output, as indicated at the bottom of Table 5.8.

With this comparison, where the duration required to pronounce the word is taken into account, it is possible to say with a higher degree of certainty that the "4.22%" maximum value of the single word utterance of "Perfect" represents a speech rate that is a similar rate to the "3.45%" first quarter value of "Yes" and the "9.00%" mean value of "Okay". Also, the previous assumption that the three multi-word examples were spoken extremely slowly, quite slowly and quite quickly in the order of

ILMT-s2s: Utt. Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Perfect	0.460	0.485	0.543	0.535	0.577	0.611	0.06
Yes	0.327	0.392	0.432	0.460	0.509	0.821	0.10
Okay	0.097	0.345	0.429	0.456	0.535	1.081	0.18
Pebbled shore	0.994	1.087	1.180	1.132	1.201	1.222	0.12
Go down	0.882	0.902	0.926	0.976	0.988	1.219	0.13
Then where?	0.536	0.542	0.547	0.554	0.564	0.5800	0.02
HCRC: Utt. Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Yes	0.255	0.333	0.395	0.402	0.447	0.647	0.10
No	0.092	0.209	0.308	0.301	0.364	0.748	0.12
Okay	0.143	0.317	0.382	0.386	0.447	0.771	0.11
Got you	0.247	0.304	0.323	0.321	0.342	0.389	0.05
And then	0.253	0.469	0.500	0.493	0.550	0.758	0.15
Right okay	0.249	0.492	0.603	0.633	0.726	1.366	0.23
ILMT-s2s: Utt. Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Perfect	98.2	104.0	110.5	113.4	123.7	130.4	13.46
Yes	73.1	117.9	138.9	135.8	153.1	183.5	25.23
Okay	55.5	112.2	139.9	154.0	173.9	618.6	75.16
Pebbled shore	98.2	99.9	101.7	106.9	111.2	120.7	12.12
Go down	98.4	121.6	129.6	124.5	133.1	136.1	14.04
Then where?	206.9	213.1	219.4	216.7	221.6	223.9	8.80
HCRC: Utt. Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Yes	92.7	134.2	151.8	157.3	180.7	235.6	37.22
No	80.3	164.8	194.6	238.8	287.2	653.6	113.11
Okay	77.8	134.2	156.9	169.5	189.3	421.1	55.92
Got you	308.7	351.2	372.7	381.1	395.1	485.8	60.41
And then	158.3	218.1	239.9	271.2	255.7	475.1	103.72
Right okay	87.8	165.4	199.1	214.1	243.9	481.3	79.07
ILMT-s2s: TTS Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Perfect	0.478	0.479	0.480	0.482	0.485	0.488	0.00
Yes	0.522	0.527	0.527	0.528	0.529	0.534	0.00
Okay	0.410	0.466	0.481	0.505	0.582	0.590	0.05
Pebbled shore	0.789	0.790	0.790	0.790	0.790	0.790	0.00
Go down	0.562	0.562	0.563	0.563	0.564	0.565	0.00
Then where?	0.471	0.472	0.473	0.472	0.473	0.473	0.00
HCRC: TTS Duration (sec.)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Yes	0.487	0.524	0.527	0.522	0.529	0.532	0.01
No	0.410	0.446	0.448	0.444	0.450	0.453	0.01
Okay	0.532	0.582	0.584	0.578	0.585	0.593	0.02
Got you	0.554	0.559	0.560	0.572	0.563	0.633	0.03
And then	0.359	0.389	0.394	0.401	0.426	0.437	0.03
Right okay	0.764	0.829	0.832	0.824	0.833	0.842	0.02
ILMT-s2s: TTS Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Perfect	122.8	123.7	124.9	124.5	125.4	125.5	1.16
Yes	112.4	113.4	113.9	113.7	113.9	114.9	0.61
Okay	101.7	103.1	124.6	120.2	128.7	146.2	12.31
Pebbled shore	151.8	151.8	151.8	151.9	151.9	152.0	0.12
Go down	212.3	213.0	213.4	213.2	213.6	213.6	0.51
Then where?	253.5	253.5	253.5	254.0	254.2	254.8	0.73
HCRC: TTS Speech Rate (wpm)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
Yes	112.8	113.4	114.0	114.9	114.6	123.2	3.06
No	132.4	133.3	133.9	135.1	134.5	146.3	3.64
Okay	101.2	102.6	102.7	103.9	103.1	112.8	3.20
Got you	189.6	213.1	214.3	210.4	214.7	216.6	10.26
And then	274.6	281.7	304.6	300.3	308.5	334.3	19.94
Right okay	142.5	144.1	144.2	145.8	144.8	157.1	4.28
ILMT-s2s: Difference from TTS (%)	**Min.**	**1st Qu.**	***Mdn***	***M***	**3rd Qu.**	**Max.**	***SD***
Perfect	**−27.65**	**−18.98**	**−13.61**	**−11.06**	**0.71**	**4.22**	**13.39**
Yes	**−55.82**	**3.45**	**18.01**	**12.91**	**25.93**	**38.21**	**19.16**
Okay	**−130.70**	**−8.14**	**13.54**	**9.00**	**32.94**	**83.30**	**36.51**
Pebbled shore	**−54.61**	**−52.06**	**−49.50**	**−43.29**	**−37.64**	**−25.77**	**15.39**
Go down	**−117.00**	**−74.98**	**−64.39**	**−73.36**	**−60.32**	**−56.97**	**22.60**
Then where?	**−22.55**	**−19.07**	**−15.58**	**−17.31**	**−14.70**	**−13.81**	**4.62**
HCRC: Difference from TTS (%)	**Min.**	**1st Qu.**	***Mdn***	***M***	**3rd Qu.**	**Max.**	***SD***
Yes	**−23.00**	**15.58**	**23.56**	**23.01**	**37.02**	**51.30**	**18.51**
No	**−66.90**	**18.43**	**31.64**	**32.30**	**52.78**	**79.60**	**27.52**
Okay	**−31.78**	**22.46**	**33.29**	**33.27**	**45.52**	**75.93**	**18.32**
Got you	**38.59**	**38.93**	**42.60**	**43.96**	**45.78**	**55.42**	**6.54**
And then	**−98.43**	**−47.24**	**−26.98**	**−23.83**	**−12.98**	**39.67**	**39.89**
Right okay	**−64.22**	**12.93**	**26.54**	**23.14**	**40.30**	**67.41**	**27.72**

Table 5.8: *Comparison with TTS duration sample*

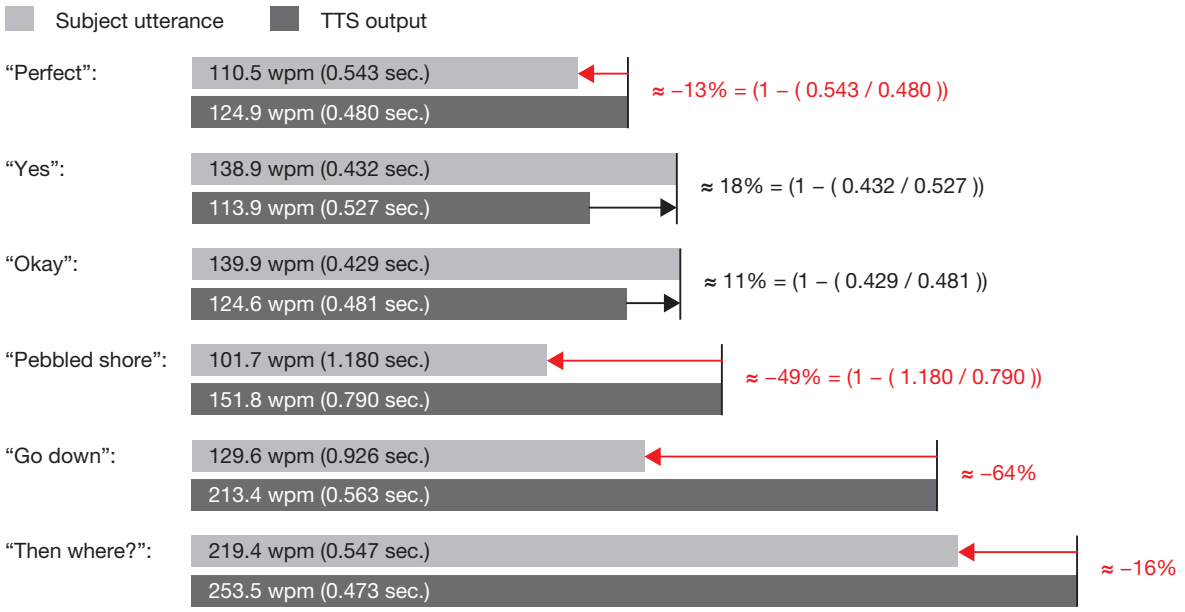


Figure 5.6: *Example speech rate comparison — Subject utterance & TTS output*

“Pebbled shore (101.7 wpm)” – “Go down (129.6 wpm)” – “Then where? (219.4 wpm)”, can now be reassessed to all examples being slower than the reference TTS output and reordered accordingly using a percentage value that indicate the difference to the duration of the TTS output. “Go down (−73.36%)” – “Pebbled shore (−43.29%)” – “Then where? (−15.58%)”.

5.7 Summary

By comparing the speech rate of the subject utterance with a reference TTS output of the same utterance (Figure 5.7), I have indicated that the resulting value is easier to interpret and more robust than the wpm measurement method, by comparing the variable subject utterance duration with a standard measurement of the TTS output duration.

Using this method, it becomes possible to visualise the speech rate of short utterances with a higher degree of certainty. It is already possible to say with confidence that there is a difference between the speech rates of computer mediated dialogues, and the non-mediated dialogues. From Figure 5.7, I can understand that the Information Giver is talking slower than the Information Follower in both corpora, while the female subjects of the computer mediated dialogue dataset are talking much slower than male subjects to a degree grater than female subjects within the non-mediated dialogue dataset.

Furthermore, Figure 5.8 illustrates how the values from this method can illustrate how the speech rate changes from utterance to utterance, indicating how speech rates differ when talking directly to ones interlocutor (the HCRC Edinburgh Map Task corpus) and when one’s communication is mediated by a Speech-to-Speech Machine Translation system (the ILMT-s2s corpus). This is studied in more detail in Chapter 7: “Speech Rate Differences Between Datasets”, and diagrams of all dialogues and various groupings can be found in § A.3 of Appendix A, and § B.3.2 of Appendix B.

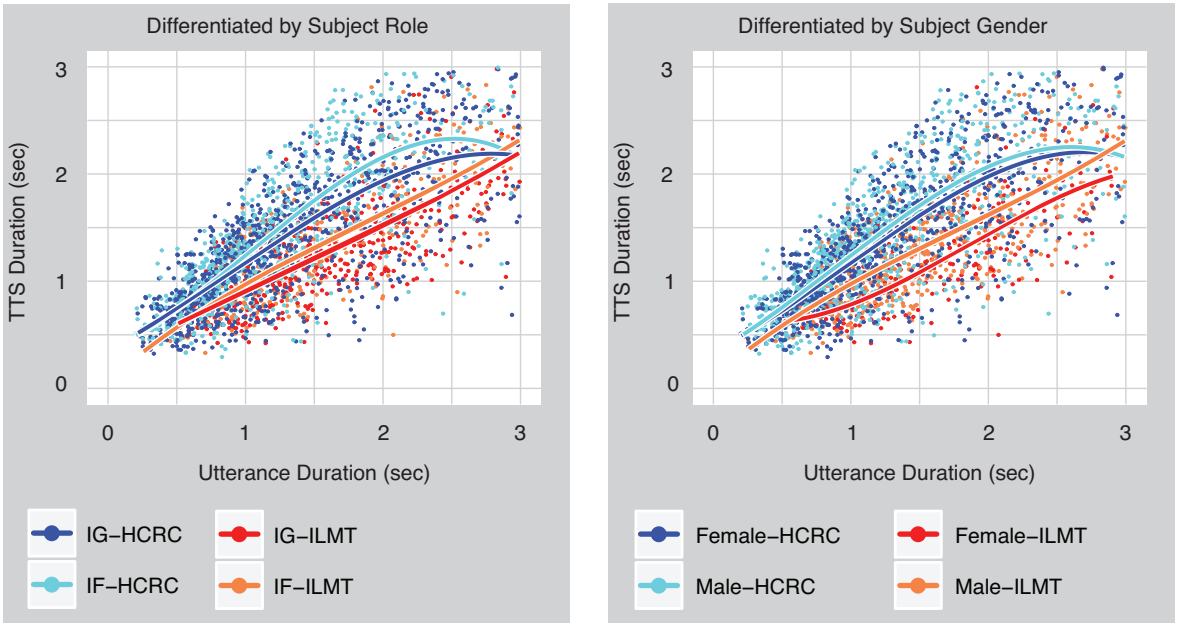


Figure 5.7: *Example of speech rate duration difference by role and gender*

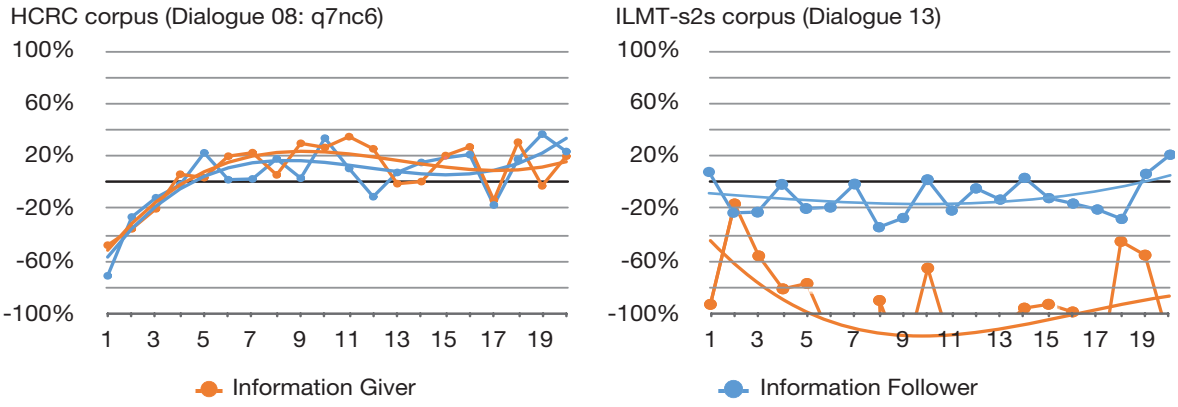


Figure 5.8: *Example of speech rate comparison of speakers — Y-axis is the speech rate difference, X-axis is the frame when all dialogues are divided into 20 frames, using the method discribed in § 4.4.6*

Chapter 6

The Three Talk Types

Using a different style of speech towards oneself and others is not a new phenomena. Batliner et al. (2009, p. 171) open their paper with an example from Shakespeare's *Hamlet*, where Hamlet seems to change his speaking style when addressing his interlocutor to that of utterances that are spoken, but not directed towards his interlocutor.

> "In this passage from Shakespeare's *Hamlet*, we find two '*Asides*', one for speaking aside to a third person, the other one for speaking to oneself. Implicitly we learn that such asides are produced with a lower voice: when Hamlet addresses Guildenstern and Rosencrantz again, the stage direction reads *Aloud*. And we can imagine that while speaking aside, Rosencrantz is turning towards Guildenstern, and Hamlet away from both — maybe towards the audience. Thus both speech characteristics and head/body orientation can play a role." (Batliner et al., 2009, p. 171)

6.1 Introduction

During the recording of the dialogues within the computer-mediated dialogue dataset, the subjects were able to direct their speech in three different directions.

- *On-Talk* — Subject utterances to the interlocutor via the ILMT-s2s System (§ 4.2).
- *Off-Talk Self* — Subject utterances to themselves.
- *Off-Talk Other* — Subject utterances to someone else; the other person could be the technician of the experiment who entered the room on the few occasions when the system crashed, or a university member using the office for other purposes.

Off-Talk is a phenomenon described by Oppermann et al. (2001) that is said to be a characteristic of Human-Computer-Interaction where the subject is not directing the interaction towards the system (computer). Oppermann et al. (2001) differentiated *Off-Talk* into two categories, "ROT" (Read Off-Talk) and "OOT" (Other forms of Off-Talk) which looks at the type of utterance, but in this report

Off-Talk is differentiated by the direction of the utterance. Is the subject talking to themselves or to someone else?

As mentioned in § 4.4.3: "Annotation", to annotate the data, the dedicated annotation tool ELAN was used to label the transcription with *On-Talk*, *Off-Talk* (*Self* and *Other*). *On-Talk* locations were retrieved from the ILMT-s2s System's log and all other utterances were annotated manually as *Off-Talk Self* and *Off-Talk Other*. This resulted in 2,439 *On-Talk*, 1,189 *Off-Talk* (848 *Self* and 341 *Other*) utterances in the computer-mediated dialogue dataset.

6.2 Dataset

In this chapter, the talk type labels of the computer-mediated dialogue dataset described in § 3.3.3.3 and the speech rate difference labels of the computer-mediated dialogue dataset described in § 3.3.3.4 of Chapter 3 are used for the analysis.

6.3 Methods

The talk type annotation labels (§ 4.4.3.4) are used to identify which utterances were directed to whom; the ILMT-s2s System (*On-Talk*), themselves (*Off-Talk Self*) or to another person Face-to-Face (*Off-Talk Other*).

6.3.1 Talk Type Analysis of Each Subject

Using these labels, the speech rate of *On-Talk* and *Off-Talk* are compared, and furthermore, the speech rates of *On-Talk*, *Off-Talk Self* and *Off-Talk Other* are compared for each individual and also for each grouping (role, gender, language spoken, setting of the recording).

- The speech rate of *On-Talk* and *Off-Talk* of each of the subjects are tested to verify if there is a significant difference (*t-test*).
- The speech rate of *On-Talk*, *Off-Talk Self* and *Off-Talk Other* of each of the subjects are tested to verify if there is a significant difference (*ANOVA*).

6.3.2 Talk Type Analysis by Utterance Word Count

Due to the greater variability (standard deviation) in the speech rate of single word utterances (e.g., "umm", "ok", "yes", "what?", "ah", etc.), single and multiple word utterances are differentiated and the above verifications are re-conducted as the following four variants.

- The speech rate of *[Refer to grouping list below]* utterances of *On-Talk* and *Off-Talk* of each of the subjects are tested to verify if there is a significant difference (*t-test*).

- The speech rate of *[Refer to grouping list below]* utterances of *On-Talk*, *Off-Talk Self* and *Off-Talk Other* of each of the subjects are tested to verify if there is a significant difference (*ANOVA*).

Grouping list:

1. Single word
2. Multiple word

6.3.3 Talk Type Analysis by Grouping

Finally, the multiple word data of the subjects is divided into ten groups to verify if any difference is common to all subject's or if it depends upon a specific circumstance.

The speech rate of *multiple word* utterances of *On-Talk*, *Off-Talk Self* and *Off-Talk Other* of *[Refer to grouping list below]* is tested to verify if there is a significant difference (*ANOVA*).

Grouping list:

1. Subject's role is Information Giver
2. Subject's role is Information Follower
3. Subject is female
4. Subject is male
5. English speaking subject
6. Portuguese speaking subject
7. Can see the other (w/ Video)
8. Cannot see the other (w/o Video)
9. Wearing biosignal recording devices (w/ Bio)
10. Not wearing biosignal recording devices (w/o Bio)

Methodological Note

Biosignal are not used in this study, but they were recorded during the corpus collection. Therefore they have been added to the groupings to verify that the devices do not apply a significant influence to the results.

6.4 Results

Aside from literature, the indication that the speech rate of *On-Talk* and *Off-Talk* are different within the computer-mediated dialogue data was present in a data sheet that was created to store the data. In Figure 6.1, the colour red mostly occurs on yellow rows: The red text indicates speech rates slower than the reference TTS, and the yellow rows indicate *On-Talk*. The actual values in values of Figure 6.1 do not individually matter (summary follows) — the purpose of this illustration is to highlight the obvious overlap of red text and yellow rows.

A summary of the *On-Talk* and *Off-Talk* speech rates of all thirty subject's utterances combined, are listed in Table 6.1 and it is evident from the "*Mdn*" and "3rd Qu." values that *On-Talk* utterances are spoken slower than *Off-Talk* utterances.

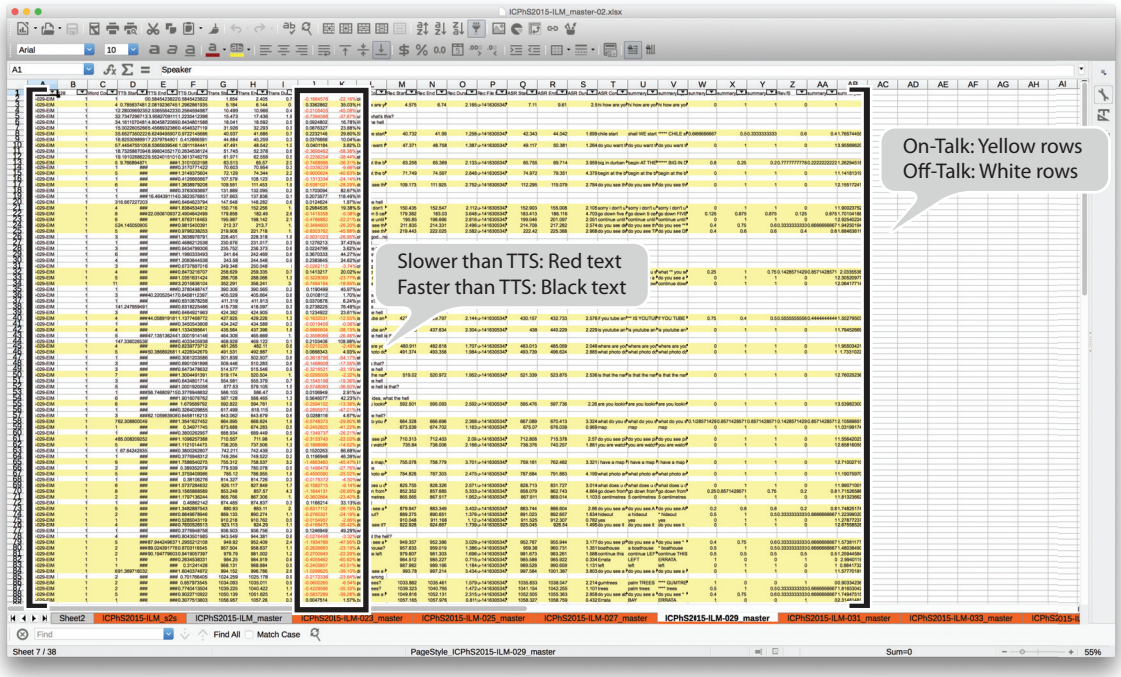


Figure 6.1: *Data sheet with On-Talk highlighted in yellow and utterances slower than the TTS reference output in red. The actual values in the data sheet rows do not matter for this illustration.*

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Count
On-Talk — All utterances	−383.90	−58.94	−28.31	−36.79	−4.82	58.28	47.15	2,439
Off-Talk — All utterances	−1,539.00	−57.66	−8.14	−40.24	15.73	83.30	110.51	1,189

Table 6.1: *Summary and standard deviation of the 2 talk types (All subjects combined)*

6.4.1 Talk Type Analysis of Each Subject

■ *On-Talk* & *Off-Talk* Significance Test of Each Subject.

Of the thirty subjects, the t-test results of thirteen subjects have a significant difference between the speech rate of *On-Talk* and *Off-Talk*.[1] Of the thirteen subjects with a significant difference, ten subjects obtained an effect size (Cohen's d) categorised as large. Refer to Table 6.2 and Table 6.2 for the test results.

No.	Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen's d	ES
1	D.07	IF - En (♂)	89 : 7 (7.3%)	6 (6.3%)	1 (1.0%)	$p < 0.0000$	***	$d = 1.3307$	large
2	D.03	IG - En (♀)	106 : 81 (43.3%)	9 (4.8%)	72 (38.5%)	$p < 0.0000$	***	$d = 1.3018$	large
3	D.08	IG - En (♀)	121 : 41 (25.3%)	11 (6.8%)	30 (18.5%)	$p < 0.0000$	***	$d = 1.0733$	large
4	D.11	IF - En (♀)	56 : 73 (56.6%)	59 (45.7%)	14 (10.9%)	$p < 0.0000$	***	$d = 1.0413$	large
5	D.14	IG - En (♂)	69 : 59 (46.1%)	59 (46.1%)	– –	$p < 0.0000$	***	$d = 0.9758$	large
6	D.13	IG - Pt (♀)	49 : 106 (68.4%)	106 (68.4%)	– –	$p < 0.0000$	***	$d = 0.7862$	med
7	D.01	IF - En (♂)	33 : 25 (43.1%)	25 (43.1%)	– –	$p < 0.0001$	***	$d = 1.4665$	large
8	D.06	IF - Pt (♀)	78 : 44 (36.1%)	44 (36.1%)	– –	$p < 0.0041$	**	$d = 0.7591$	med
9	D.01	IG - Pt (♀)	55 : 6 (9.8%)	6 (9.8%)	– –	$p < 0.0106$	*	$d = 0.8577$	large
10	D.06	IG - En (♀)	93 : 100 (51.8%)	91 (47.2%)	9 (4.7%)	$p < 0.0136$	*	$d = 0.3522$	small
11	D.02	IG - En (♂)	100 : 59 (37.1%)	55 (34.6%)	4 (2.5%)	$p < 0.0209$	*	$d = 0.4566$	small
12	D.05	IG - En (♂)	201 : 2 (1.0%)	2 (1.0%)	– –	$p < 0.0343$	*	$d = 2.9422$	large
13	D.13	IF - En (♂)	49 : 2 (3.9%)	2 (3.9%)	– –	$p < 0.0414$	*	$d = 0.8022$	large

Table 6.2: *The 13 subjects with significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (all utterances)*

No.	Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen's d	ES
1	D.02	IF - Pt (♂)	117 : 45 (27.8%)	31 (19.1%)	14 (8.6%)	$p < 0.0548$	.	$d = 0.4975$	small
2	D.12	IF - Pt (♀)	76 : 12 (13.6%)	10 (11.4%)	2 (2.3%)	$p < 0.0605$	.	$d = 1.5212$	large
3	D.03	IF - Pt (♀)	103 : 116 (53.0%)	59 (26.9%)	57 (26.0%)	$p < 0.0974$	.	$d = 0.2194$	small
4	D.07	IG - Pt (♀)	94 : 4 (4.1%)	– –	4 (4.1%)	$p < 0.1242$	–	$d = 0.6444$	med
5	D.14	IF - Pt (♀)	57 : 10 (14.9%)	10 (14.9%)	– –	$p < 0.1606$	–	$d = 1.2068$	small
6	D.09	IG - Pt (♂)	98 : 5 (4.9%)	5 (4.9%)	– –	$p < 0.1685$	–	$d = 2.3886$	large
7	D.04	IF - En (♂)	73 : 46 (38.7%)	40 (33.6%)	6 (5.0%)	$p < 0.1711$	–	$d = 0.3058$	small
8	D.09	IF - En (♂)	104 : 24 (18.8%)	24 (18.8%)	– –	$p < 0.1914$	–	$d = 0.3951$	small
9	D.11	IG - Pt (♀)	101 : 2 (1.9%)	2 (1.9%)	– –	$p < 0.2964$	–	$d = 4.8509$	large
10	D.08	IF - Pt (♀)	53 : 87 (62.1%)	30 (21.4%)	57 (40.7%)	$p < 0.2999$	–	$d = 0.1540$	neg
11	D.12	IG - En (♂)	100 : 54 (35.1%)	54 (35.1%)	– –	$p < 0.3042$	–	$d = 0.2263$	small
12	D.10	IG - Pt (♂)	76 : 52 (40.6%)	21 (16.4%)	31 (24.2%)	$p < 0.3571$	–	$d = 0.1932$	neg
13	D.10	IF - En (♂)	92 : 78 (45.9%)	46 (27.1%)	32 (18.8%)	$p < 0.6776$	–	$d = 0.0677$	neg
14	D.15	IG - Pt (♂)	34 : 15 (30.6%)	13 (26.5%)	2 (4.1%)	$p < 0.9102$	–	$d = 0.0495$	neg
15	D.04	IG - Pt (♀)	59 : 33 (35.9%)	28 (30.4%)	5 (5.4%)	$p < 0.9658$	–	$d = 0.0113$	neg
16	D.15	IF - En (♀)	42 : 1 (2.3%)	– –	1 (2.3%)	–	–	–	–
17	D.05	IF - Pt (♀)	61 : 0 –	– –	– –	–	–	–	

Table 6.3: *The 17 subjects without significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (all utterances)*

A boxplot of the *On-Talk* and *Off-Talk* speech rates of all the subjects combined can be found on left side of Figure 6.2.

[1] Of the thirty dialogues, nine of the dialogues had less than 10% of *Off-Talk* utterances. If we remove the nine dialogues where the *Off-Talk* utterances accounted for less than 10% of the total utterances count, then the ratio becomes nine out of the twenty one dialogues.

■ *On-Talk*, *Off-Talk Self* & *Off-Talk Other* Significance Test of Each Subject.

Following the positive results from the two categories of *On-Talk* and *Off-Talk* speech rates using the *t*-test, an ANOVA test was performed on the speech rates of the three categories *On-Talk*, *Off-Talk Self*, and *Off-Talk Other* of each subject.

By differentiating the direction of the *Off-Talk* to talking to oneself (*Off-Talk Self*) and talking directly to a fellow human (*Off-Talk Other*) twenty one subjects out of the thirty[2] showing a significant difference in the three talk type combination of *On-Talk*, *Off-Talk Self* and *Off-Talk Other*. Refer to Table 6.4 and Table 6.5 for the ANOVA test results.

No.	Dialog	Role - Lang	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA	sig.
1	D.01	IF - En (♂)	33	25 (43.1%)	– –	$p < 0.0000$	***
2	D.02	IG - En (♂)	100	55 (34.6%)	4 (2.5%)	$p < 0.0000$	***
3	D.03	IG - En (♀)	106	9 (4.8%)	72 (38.5%)	$p < 0.0000$	***
4	D.08	IG - En (♀)	121	11 (6.8%)	30 (18.5%)	$p < 0.0000$	***
5	D.09	IG - Pt (♂)	98	5 (4.9%)	– –	$p < 0.0000$	***
6	D.10	IG - Pt (♂)	76	21 (16.4%)	31 (24.2%)	$p < 0.0000$	***
7	D.11	IG - Pt (♀)	101	2 (1.9%)	– –	$p < 0.0000$	***
8	D.11	IF - En (♀)	56	59 (45.7%)	14 (10.9%)	$p < 0.0000$	***
9	D.12	IF - Pt (♀)	76	10 (11.4%)	2 (2.3%)	$p < 0.0000$	***
10	D.13	IG - Pt (♀)	49	106 (68.4%)	– –	$p < 0.0000$	***
11	D.14	IG - En (♂)	69	59 (46.1%)	– –	$p < 0.0000$	***
12	D.03	IF - Pt (♀)	103	59 (26.9%)	57 (26.0%)	$p < 0.0001$	***
13	D.06	IF - Pt (♀)	78	44 (36.1%)	– –	$p < 0.0001$	***
14	D.05	IG - En (♂)	201	2 (1.0%)	– –	$p < 0.0001$	***
15	D.08	IF - Pt (♀)	53	30 (21.4%)	57 (40.7%)	$p < 0.0003$	***
16	D.14	IF - Pt (♀)	57	10 (14.9%)	– –	$p < 0.0008$	***
17	D.04	IG - Pt (♀)	59	28 (30.4%)	5 (5.4%)	$p < 0.0013$	***
18	D.07	IF - En (♂)	89	6 (6.3%)	1 (1.0%)	$p < 0.0044$	**
19	D.02	IF - Pt (♂)	117	31 (19.1%)	14 (8.6%)	$p < 0.0166$	*
20	D.06	IG - En (♀)	93	91 (47.2%)	9 (4.7%)	$p < 0.0168$	*
21	D.10	IF - En (♂)	92	46 (27.1%)	32 (18.8%)	$p < 0.0392$	*

Table 6.4: *The 21 subjects with significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)*

No.	Dialog	Role - Lang	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA	sig.
1	D.01	IG - Pt (♀)	55	6 (9.8%)	– –	$p < 0.0507$	.
2	D.04	IF - En (♂)	73	40 (33.6%)	6 (5.0%)	$p < 0.0633$	.
3	D.09	IF - En (♂)	104	24 (18.8%)	– –	$p < 0.0835$	.
4	D.15	IG - Pt (♂)	34	13 (26.5%)	2 (4.1%)	$p < 0.1810$	–
5	D.12	IG - En (♂)	100	54 (35.1%)	– –	$p < 0.1820$	–
6	D.07	IG - Pt (♀)	94	– –	4 (4.1%)	$p < 0.2100$	–
7	D.13	IF - En (♂)	49	2 (3.9%)	– –	$p < 0.2720$	–
8	D.15	IF - En (♀)	42	– –	1 (2.3%)	$p < 0.7910$	–
9	D.05	IF - Pt (♀)	61	– –	– –	–	

Table 6.5: *The 9 subjects without significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)*

Post-hoc comparisons (Tukey HSD test) revealed significant differences in most combinations of *On-Talk* and *Off-Talk Self* (17 out of 21) and in more than half the combinations of *On-Talk* and

[2]Of the thirty dialogues, nine of the dialogues had less than 10% of *Off-Talk* utterances. If we remove the nine dialogues where the *Off-Talk* utterances accounted for less than 10% of the total utterances count, then the ratio becomes seventeen out of the twenty one dialogues.

Off-Talk Other (7 out of 13) and *Off-Talk Self* and *Off-Talk Other* (8 out of 13). Refer to Table 6.6 and Table 6.7 for the Post-hoc comparisons (Tukey HSD test) results.

No.	Dial. - Role	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
		On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
1	D.01 - IF	0.0000 ***	–	–	$d = 1.4665$, large	–	–
2	D.02 - IG	0.0011 **	0.0204 *	0.0005 ***	$d = 0.6061$, med	$d = 2.0031$, large	$d = 1.4491$, large
3	D.03 - IG	0.0023 **	0.0000 ***	0.9279	$d = 1.1309$, large	$d = 1.3156$, large	$d = 0.1394$, neg
4	D.08 - IG	0.7356	0.0000 ***	0.0026 *	$d = 0.2217$, small	$d = 1.6068$, large	$d = 1.0144$, large
5	D.09 - IG	0.0000 ***	–	–	$d = 2.3886$, large	–	–
6	D.10 - IG	0.0000 ***	0.0247 *	0.0000 ***	$d = 1.2639$, large	$d = 0.9367$, large	$d = 1.4095$, large
7	D.11 - IG	0.0000 ***	–	–	$d = 4.8509$, large	–	–
8	D.11 - IF	0.0000 ***	0.0000 ***	0.3604	$d = 0.9345$, large	$d = 1.2344$, large	$d = 0.5009$, med
9	D.12 - IF	0.0000 ***	0.9051	0.0120 *	$d = 1.9544$, large	$d = 0.6287$, med	$d = 0.8757$, large
10	D.13 - IG	0.0000 ***	–	–	$d = 0.7862$, med	–	–
11	D.14 - IG	0.0000 ***	–	–	$d = 0.9758$, large	–	–
12	D.03 - IF	0.0010 **	0.6279	0.0002 ***	$d = 0.5781$, med	$d = 0.1864$, neg	$d = 0.6314$, med
13	D.06 - IF	0.0001 ***	–	–	$d = 0.7591$, med	–	–
14	D.05 - IG	0.0001 ***	–	–	$d = 2.9422$, large	–	–
15	D.08 - IF	0.0029 **	0.6949	0.0002 ***	$d = 0.7055$, med	$d = 0.2117$, small	$d = 0.7752$, med
16	D.14 - IF	0.0008 ***	–	–	$d = 1.2068$, large	–	–
17	D.04 - IG	0.4161	0.0036 **	0.0008 ***	$d = 0.2849$, small	$d = 2.2399$, large	$d = 1.3315$, large
18	D.07 - IF	0.0091 **	0.2427	0.9424	$d = 1.2737$, large	–	–
19	D.02 - IF	0.0172 *	0.4174	0.8057	$d = 0.5919$, med	$d = 0.4938$, small	$d = 0.1208$, neg
10	D.06 - IG	0.0985 .	0.0453 *	0.2813	$d = 0.3004$, small	$d = 1.3270$, large	$d = 0.4282$, small
21	D.10 - IF	0.6139	0.1132	0.0328 *	$d = 1.2639$, large	$d = 0.7130$, med	$d = 1.4095$, large

Table 6.6: *The 21 subjects with significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)*

No.	Dial. - Role	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
		On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
1	D.01 - IG	0.0507 .	–	–	$d = 0.8577$, large	–	–
2	D.04 - IF	0.0633 .	0.1011	0.6921	$d = 0.4049$, small	$d = 0.5454$, med	$d = 0.5316$, med
3	D.09 - IF	0.0835 .	–	–	$d = 0.3951$, small	–	–
4	D.15 - IG	0.7420	0.2445	0.1571	$d = 0.2384$, small	$d = 2.1540$, large	$d = 0.8586$, large
5	D.12 - IG	0.1822	–	–	$d = 0.2263$, small	–	–
6	D.07 - IG	–	0.2100	–	–	$d = 0.6444$, med	–
7	D.13 - IF	0.2716	–	–	$d = 0.8022$, large	–	–
8	D.15 - IF	–	0.7912	–	–	–	–
9	D.05 - IF	–	–	–	–	–	–

Table 6.7: *The 9 subjects without significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)*

This increase shows that the differentiation of *Off-Talk Self* and *Off-Talk Other* is meaningful. A boxplot of the *On-Talk*, *Off-Talk Self* and *Off-Talk Other* speech rates of all the subjects combined can be found on the right side of Figure 6.2.

6.4.2 Talk Type Analysis by Utterance Word Count

A greater variability (standard deviation) in the speech rate of one word utterances (e.g., "umm", "ok", "yes", "what?", "ah", etc.) was identified when compared to multiple word utterances.

Table 6.8 and Table 6.9 shows a summary of the speech rate values of single word utterances, and

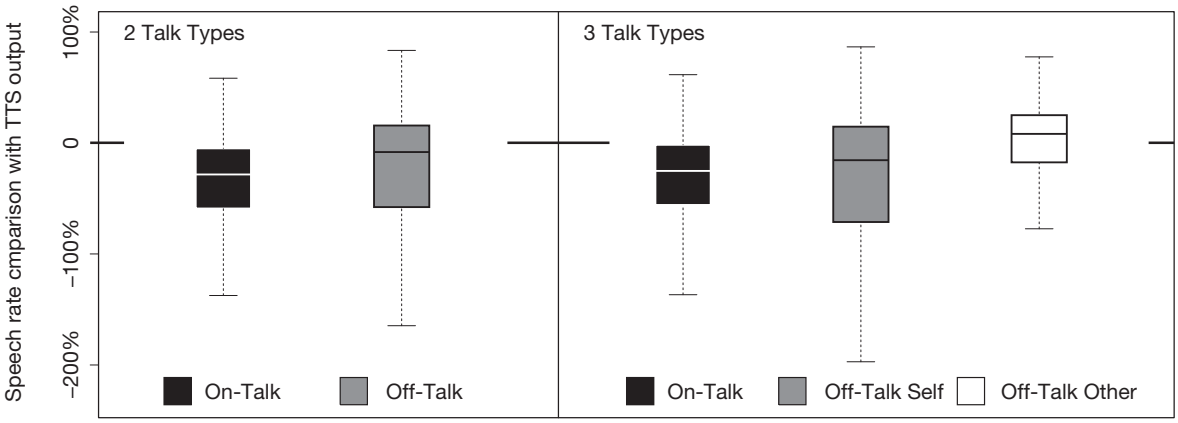


Figure 6.2: *Boxplot of all word, single word, and multi word utterance speech rates of all thirty subjects*

two or more word utterances with a clear indication that the single word utterances have a wider variability (standard deviation).

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Count
On-Talk — 1 word utterances	−328.00	−43.48	−13.92	−28.44	7.98	51.90	59.30	344
On-Talk — 2 ≤ word utterances	−383.90	−60.50	−31.10	−38.17	−7.32	58.28	44.69	2,095
Off-Talk — 1 word utterances	−1,539.00	−114.60	−30.07	−73.29	11.59	83.30	145.65	562
Off-Talk — 2 ≤ word utterances	−443.50	−27.79	0.65	−10.81	18.43	71.48	48.75	627

Table 6.8: *Summary and standard deviation of 2 talk types of single word, and multi word utterance speech rates of all thirty subjects*

Furthermore, from Table 6.8 it is evident that the characteristics of the single word utterances affect the variability of *Off-Talk*, with not only a wider standard deviation (*SD*), but greatly differing means (*M*) and median (*Mdn*) values.

Looking closer at *Off-Talk*, Table 6.9 indicates that, in particular, *Off-Talk Self* is affected by the single word utterances with a standard deviation (*SD*) of 155.17 in single word utterances as compared to the 55.59 of two or more word utterances.[3]

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Count
Off-Talk Self — 1 word utterances	−1539.00	−126.90	−39.68	−83.80	6.46	83.30	155.17	455
Off-Talk Self — 2 ≤ word utterances	−443.50	−38.42	−6.33	−18.17	13.70	65.48	55.59	393
Off-Talk Other — 1 word utterances	−433.40	−68.77	1.42	−28.82	21.62	74.29	82.11	107
Off-Talk Other — 2 ≤ word utterances	−139.40	−13.85	7.35	1.62	21.95	71.48	30.52	234

Table 6.9: *Summary and standard deviation of 3 talk types of single word, and multi word utterance speech rates of all thirty subjects*

■ Removal of the Single Word Utterances.

The removal of one word utterances resulted in 2,095 *On-Talk*, 632 *Off-Talk* (397 *Self* and 235 *Other*) utterances being used for the analysis.

A boxplot of all three situations (all utterances, only single word utterances, and only multi word utterances) is illustrated in Figure 6.3 and shows how the removal of single word utterances clarifies the pattern of the speech rates of the three talk types (*On-Talk*, *Off-Talk Self*, and *Off-Talk Other*).

[3]T-test results of the comparison of *On-Talk* and *Off-Talk* speech rates of only 1 word utterances show that nine of the thirty subjects show a significant difference with medium to large effect sizes (Table A.9 of Appendix A).

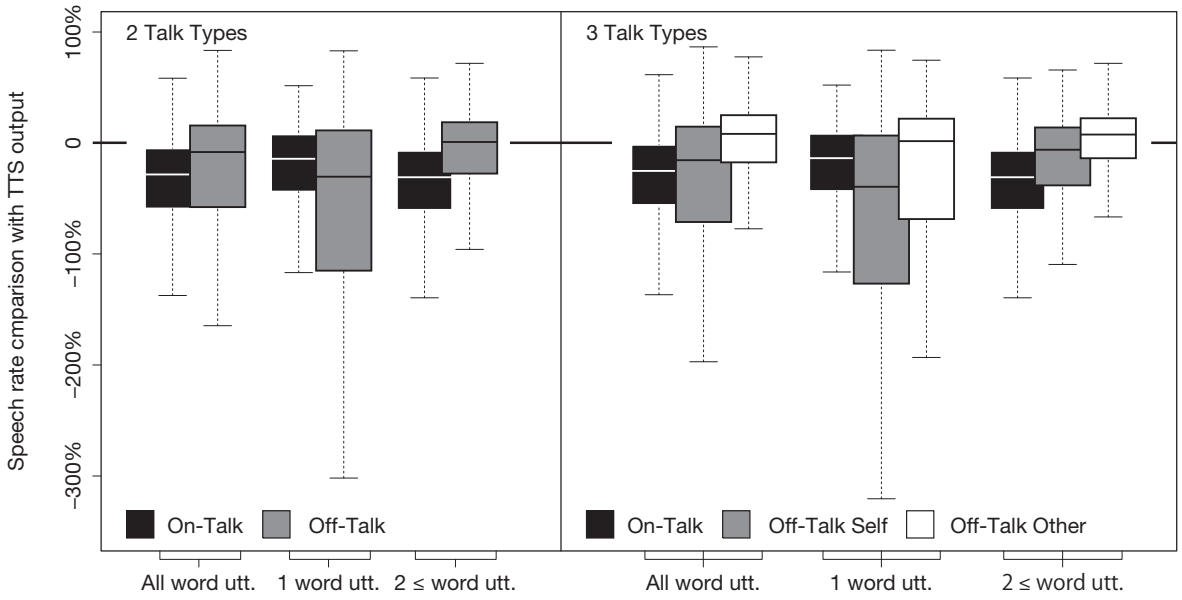


Figure 6.3: *Boxplot of all word, single word, and multi word utterance speech rates of all thirty subjects*

■ *On-Talk* & *Off-Talk* Multi Word Utterance Significance Test of Each Subject.

As performed when all utterances were analysed, the *t*-test was used to compare the *On-Talk* and *Off-Talk* speech rates. The *t*-test results show that fourteen of the thirty subjects show a significant difference and consist of mainly medium to large effect sizes. Refer to Table 6.10 and Table 6.11 for the test results.

No.	Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen's d	ES
1	D.03	IG - En (♀)	86 : 48 (35.8%)	4 (3.0%)	44 (32.8%)	$p < 0.0000$	***	$d = 1.8069$	large
2	D.08	IG - En (♀)	104 : 23 (18.1%)	1 (0.8%)	22 (17.3%)	$p < 0.0000$	***	$d = 1.6702$	large
3	D.07	IF - En (♂)	80 : 7 (8.1%)	6 (6.9%)	1 (1.1%)	$p < 0.0000$	***	$d = 1.5463$	large
4	D.11	IF - En (♀)	48 : 51 (51.5%)	42 (42.4%)	9 (9.1%)	$p < 0.0000$	***	$d = 1.2100$	large
5	D.13	IG - Pt (♀)	43 : 70 (61.9%)	70 (61.9%)	– –	$p < 0.0000$	***	$d = 1.1616$	large
6	D.06	IG - En (♀)	89 : 67 (43.0%)	59 (37.8%)	8 (5.1%)	$p < 0.0000$	***	$d = 0.9390$	large
7	D.14	IG - En (♂)	59 : 29 (33.0%)	29 (33.0%)	– –	$p < 0.0001$	***	$d = 0.8651$	large
8	D.04	IG - Pt (♀)	48 : 13 (21.3%)	8 (13.1%)	5 (8.2%)	$p < 0.0002$	***	$d = 1.3152$	large
9	D.12	IG - En (♂)	86 : 12 (12.2%)	12 (12.2%)	– –	$p < 0.0004$	***	$d = 1.4567$	large
10	D.10	IF - En (♂)	69 : 50 (42.0%)	29 (24.4%)	21 (17.6%)	$p < 0.0021$	**	$d = 0.6007$	med
11	D.15	IG - Pt (♂)	33 : 4 (10.8%)	3 (8.1%)	1 (2.7%)	$p < 0.0022$	**	$d = 1.8327$	large
12	D.08	IF - Pt (♀)	51 : 52 (50.5%)	14 (13.6%)	38 (36.9%)	$p < 0.0094$	**	$d = 0.5226$	med
13	D.03	IF - Pt (♀)	85 : 63 (42.6%)	21 (14.2%)	42 (28.4%)	$p < 0.0169$	*	$d = 0.3998$	small
14	D.07	IG - Pt (♀)	93 : 3 (3.1%)	– –	3 (3.1%)	$p < 0.0178$	*	$d = 0.9134$	large

Table 6.10: *The 14 subjects with significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (w/o 1 word utterances)*

No.	Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen's d	ES
1	D.04	IF - En (♂)	50 : 23 (31.5%)	19 (26.0%)	4 (5.5%)	$p < 0.0542$	.	$d = 0.5551$	med
2	D.13	IF - En (♂)	46 : 2 (4.2%)	2 (4.2%)	– –	$p < 0.0571$	.	$d = 1.0330$	large
3	D.06	IF - Pt (♀)	64 : 26 (28.9%)	26 (28.9%)	– –	$p < 0.0852$	.	$d = 0.5054$	med
4	D.01	IG - Pt (♀)	47 : 3 (6.0%)	3 (6.0%)	– –	$p < 0.0856$	.	$d = 0.7786$	med
5	D.10	IG - Pt (♂)	56 : 27 (32.5%)	4 (4.8%)	23 (27.7%)	$p < 0.2158$	–	$d = 0.3752$	small
6	D.02	IF - Pt (♂)	108 : 23 (17.6%)	14 (10.7%)	9 (6.9%)	$p < 0.2682$	–	$d = 0.3221$	small
7	D.01	IF - En (♂)	32 : 2 (5.9%)	2 (5.9%)	– –	$p < 0.4909$	–	$d = 0.6810$	med
8	D.12	IF - Pt (♀)	68 : 3 (4.2%)	2 (2.8%)	1 (1.4%)	$p < 0.6478$	–	$d = 0.4365$	small
9	D.02	IG - En (♂)	79 : 11 (12.2%)	9 (10.0%)	2 (2.2%)	$p < 0.6601$	–	$d = 0.2149$	small
10	D.14	IF - Pt (♀)	37 : 2 (5.1%)	2 (5.1%)	– –	$p < 0.7728$	–	$d = 0.2997$	small
11	D.09	IF - En (♂)	95 : 13 (12.0%)	12 (12.0%)	– –	$p < 0.9090$	–	$d = 0.0528$	neg
12	D.15	IF - En (♀)	35 : 1 (2.8%)	– –	1 (2.8%)	–	–	–	–
13	D.09	IG - Pt (♂)	94 : 1 (1.1%)	1 (1.1%)	– –	–	–	–	–
14	D.05	IG - En (♂)	163 : 0 –	– –	– –	–	–	–	
15	D.11	IG - Pt (♀)	87 : 0 –	– –	– –	–	–	–	–
16	D.05	IF - Pt (♀)	58 : 0 –	– –	– –	–	–	–	

Table 6.11: *The 16 subjects without significant difference: Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (w/o 1 word utterances)*

■ *On-Talk*, *Off-Talk Self* & *Off-Talk Other* Multi Word Utterance Significance Test of Each Subject.

As performed when all utterances were analysed, the ANOVA test to compare the *On-Talk*, *Off-Talk Self*, and *Off-Talk Other* speech rates. The ANOVA test results show that seventeen[4] of the thirty subjects show a significant difference and consist of mainly medium to large effect sizes. Refer to Table 6.12 and Table 6.13 for the ANOVA test results.

No.	Dialog	Role - Lang	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA	sig.
1	D.03	IG - En (♀)	86	4 (3.0%)	44 (32.8%)	$p < 0.0000$	***
2	D.04	IG - Pt (♀)	48	8 (13.1%)	5 (8.2%)	$p < 0.0000$	***
3	D.06	IG - En (♀)	89	59 (37.8%)	8 (5.1%)	$p < 0.0000$	***
4	D.08	IG - En (♀)	104	1 (0.8%)	22 (17.3%)	$p < 0.0000$	***
5	D.10	IG - Pt (♂)	56	4 (4.8%)	23 (27.7%)	$p < 0.0000$	***
6	D.11	IF - En (♀)	48	42 (42.4%)	9 (9.1%)	$p < 0.0000$	***
7	D.12	IG - En (♂)	86	12 (12.2%)	– –	$p < 0.0000$	***
8	D.13	IG - Pt (♀)	43	70 (61.9%)	– –	$p < 0.0000$	***
9	D.14	IG - En (♂)	59	59 (33.0%)	– –	$p < 0.0003$	***
10	D.10	IF - En (♂)	69	29 (24.4%)	21 (17.6%)	$p < 0.0006$	***
11	D.07	IF - En (♂)	80	6 (6.9%)	1 (1.1%)	$p < 0.0009$	***
12	D.15	IG - Pt (♂)	33	3 (8.1%)	1 (2.7%)	$p < 0.0065$	**
13	D.03	IF - Pt (♀)	85	21 (14.2%)	42 (28.4%)	$p < 0.0109$	*
14	D.08	IF - Pt (♀)	51	14 (13.6%)	38 (36.9%)	$p < 0.0318$	*
15	D.02	IG - En (♂)	79	9 (10.0%)	2 (2.2%)	$p < 0.0430$	*
16	D.06	IF - Pt (♀)	64	26 (28.9%)	– –	$p < 0.0325$	*
17	D.04	IF - En (♂)	50	19 (26.0%)	4 (5.5%)	$p < 0.0356$	*

Table 6.12: *The 17 subjects with significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)*

[4]Thirteen subjects not having a significant difference may seem to be a large portion of the corpus. However, of the thirteen subjects, only two subject had more than three *Off-Talk* utterances.

No.	Dialog	Role - Lang	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA sig.
1	D.07	IG - Pt (♀)	93	– –	3 (3.1%)	$p < 0.1230$ –
2	D.13	IF - En (♂)	46	2 (4.2%)	– –	$p < 0.1590$ –
3	D.01	IG - Pt (♀)	47	3 (6.0%)	– –	$p < 0.1970$ –
4	D.02	IF - Pt (♂)	108	14 (10.7%)	9 (6.9%)	$p < 0.3280$ –
5	D.01	IF - En (♂)	32	2 (5.9%)	– –	$p < 0.3570$ –
6	D.15	IF - En (♀)	35	– –	1 (2.8%)	$p < 0.6590$ –
7	D.14	IF - Pt (♀)	37	2 (5.1%)	– –	$p < 0.6820$ –
8	D.12	IF - Pt (♀)	68	2 (2.8%)	1 (1.4%)	$p < 0.7350$ –
9	D.09	IG - Pt (♂)	94	1 (1.1%)	– –	$p < 0.7710$ –
10	D.09	IF - En (♂)	95	12 (12.0%)	– –	$p < 0.8590$ –
11	D.05	IG - En (♂)	163	– –	– –	– –
12	D.05	IF - Pt (♀)	58	– –	– –	– –
13	D.11	IG - Pt (♀)	87	– –	– –	– –

Table 6.13: *The 13 subjects without significant difference: ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)*

Post-hoc comparisons (Tukey HSD test) revealed significant differences in most combinations of *On-Talk* and *Off-Talk Other* (10 out of 13) and in more than half the combinations of *On-Talk* and *Off-Talk Self* (10 out of 17) and the least combinations of *Off-Talk Self* and *Off-Talk Other* (3 out of 13). Refer to Table 6.14 and Table 6.15 for the Post-hoc comparisons (Tukey HSD test) results.

No.	Dial. - Role	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
		On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
1	D.03 - IG	0.0113 *	0.0000 ***	0.7976	$d = 1.3210$, large	$d = 1.8159$, large	$d = 0.5202$, med
2	D.04 - IG	0.0920 .	0.0000 ***	0.0346 *	$d = 0.7883$, med	$d = 2.2029$, large	$d = 2.4135$, large
3	D.06 - IG	0.0000 ***	0.0138 *	0.9362	$d = 0.9069$, large	$d = 1.2415$, large	$d = 0.1124$, neg
4	D.08 - IG	0.6534	0.0000 ***	0.7035	–	$d = 1.7034$, large	–
5	D.10 - IG	0.0001 ***	0.0010 ***	0.0000 ***	$d = 2.0147$, large	$d = 1.3099$, large	$d = 2.4285$, large
6	D.11 - IF	0.0000 ***	0.0006 ***	0.8000	$d = 1.1427$, large	$d = 1.2100$, large	$d = 0.3152$, small
7	D.12 - IG	0.0000 ***	–	–	$d = 1.4567$, large	–	–
8	D.13 - IG	0.0000 ***	–	–	$d = 1.1616$, large	–	–
9	D.14 - IG	0.0003 ***	–	–	$d = 0.8651$, large	–	–
10	D.10 - IF	0.2680	0.0004 ***	0.0740 .	$d = 0.3211$, small	$d = 1.1240$, large	$d = 0.5924$, med
11	D.07 - IF	0.0021 **	0.1612	0.9368	$d = 1.4851$, large	–	–
12	D.15 - IG	0.0182 *	0.1309	0.9686	$d = 1.7384$, large	–	–
13	D.03 - IF	0.9525	0.0085 **	0.1553	$d = 0.0651$, neg	$d = 0.6225$, med	$d = 0.5135$, med
14	D.08 - IF	0.3362	0.0296 *	0.9129	$d = 0.4187$, small	$d = 0.5420$, med	$d = 0.1368$, neg
15	D.02 - IG	0.9289	0.0370 *	0.0415 *	$d = 0.1283$, neg	$d = 1.9285$, large	$d = 1.3227$, large
16	D.06 - IF	0.0325 *	–	–	$d = 0.5054$, med	–	–
17	D.04 - IF	0.2679	0.0575 .	0.3266	$d = 0.4371$, small	$d = 1.2586$, large	$d = 0.6697$, med

Table 6.14: *The 17 subjects with significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)*

No.	Dial. - Role	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
		On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
1	D.07 - IG	–	0.1228	–	–	$d = 0.9134$, large	–
2	D.13 - IF	0.1595	–	–	$d = 1.0330$, large	–	–
3	D.01 - IG	0.1973	–	–	$d = 0.7786$, med	–	–
4	D.02 - IF	0.3195	0.8608	0.8523	$d = 0.4300$, small	$d = 0.1854$, neg	$d = 0.1745$, neg
5	D.01 - IF	0.3571	–	–	$d = 0.6810$, med	–	–
6	D.15 - IF	–	0.6587	–	–	–	–
7	D.14 - IF	0.6821	–	–	$d = 0.2997$, small	–	–
8	D.12 - IF	0.7264	–	–	$d = 0.5479$, med	–	–
9	D.09 - IG	0.7709	–	–	–	–	–
10	D.09 - IF	0.8585	–	–	$d = 0.0528$, neg	–	–
11	D.05 - IG	–	–	–	–	–	–
12	D.05 - IF	–	–	–	–	–	–
13	D.11 - IG	–	–	–	–	–	–

Table 6.15: *The 13 subjects without significant difference: Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)*

6.4.3 Talk Type Analysis by Grouping

To verify if this speech rate difference was not a trait of a specific group that was affecting the whole data, ANOVA tests were performed on the groupings of the subjects (role, gender, spoken language, etc.) which resulted in all groupings showing a significant difference between the speech rates. Refer to Table 6.16 for the ANOVA test results.

No.	Category	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA	sig.
1	All	2,095	397 (14.6%)	235 (8.6%)	$p < 0.0000$	***
2	IG	1,169	205 (13.8%)	109 (7.3%)	$p < 0.0000$	***
3	IF	926	192 (15.4%)	126 (10.1%)	$p < 0.0000$	***
4	♀	1,044	254 (17.3%)	174 (11.8%)	$p < 0.0000$	***
5	♂	1,051	143 (11.4%)	61 (4.9%)	$p < 0.0000$	***
6	Eng	1,122	228 (15.6%)	113 (7.7%)	$p < 0.0000$	***
7	Por	973	169 (13.4%)	122 (9.7%)	$p < 0.0000$	***
8	w/ Video	1,197	171 (10.8%)	212 (13.4%)	$p < 0.0000$	***
9	w/o Video	898	226 (19.7%)	23 (2.0%)	$p < 0.0000$	***
10	w/ Bio	1,108	202 (14.4%)	93 (6.6%)	$p < 0.0000$	***
11	w/o Bio	987	195 (14.7%)	142 (10.7%)	$p < 0.0000$	***

Table 6.16: *ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance by groupings)*

Post-hoc comparisons (Tukey HSD test) revealed significant differences in all but one combination of Portuguese *On-Talk* and *Off-Talk Self*. Refer to Table 6.17 for the Post-hoc comparisons (Tukey HSD test) results.

The boxplots of Figure 6.4 shows the distribution of the ten groupings that were analysed.

No.	Category	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
		On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
1	All	0.0000 ***	0.0000 ***	0.0000 ***	$d = 0.4128$, small	$d = 0.9022$, large	$d = 0.4764$, small
2	IG	0.0000 ***	0.0000 ***	0.0000 ***	$d = 0.5250$, med	$d = 1.2098$, large	$d = 0.6519$, med
3	IF	0.0022 **	0.0000 ***	0.0226 *	$d = 0.2625$, small	$d = 0.5871$, med	$d = 0.3189$, small
4	♀	0.0000 ***	0.0000 ***	0.0000 ***	$d = 0.4985$, small	$d = 0.9729$, large	$d = 0.4564$, small
5	♂	0.0000 ***	0.0000 ***	0.0001 ***	$d = 0.4249$, small	$d = 1.1158$, large	$d = 0.6781$, med
6	Eng	0.0000 ***	0.0000 ***	0.0046 **	$d = 0.6427$, med	$d = 1.0539$, large	$d = 0.4229$, small
7	Por	0.2398	0.0000 ***	0.0000 ***	$d = 0.1307$, neg	$d = 0.7748$, med	$d = 0.6242$, med
8	w/ Video	0.0000 ***	0.0000 ***	0.0004 ***	$d = 0.4607$, small	$d = 0.8994$, large	$d = 0.3779$, small
9	w/o Video	0.0000 ***	0.0000 ***	0.0187 *	$d = 0.3700$, small	$d = 1.4483$, large	$d = 0.5051$, med
10	w/ Bio	0.0000 ***	0.0000 ***	0.0000 ***	$d = 0.3607$, small	$d = 1.0101$, large	$d = 0.6950$, med
11	w/o Bio	0.0000 ***	0.0000 ***	0.0189 *	$d = 0.4705$, small	$d = 0.8301$, large	$d = 0.3061$, small

Table 6.17: *Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance by groupings)*

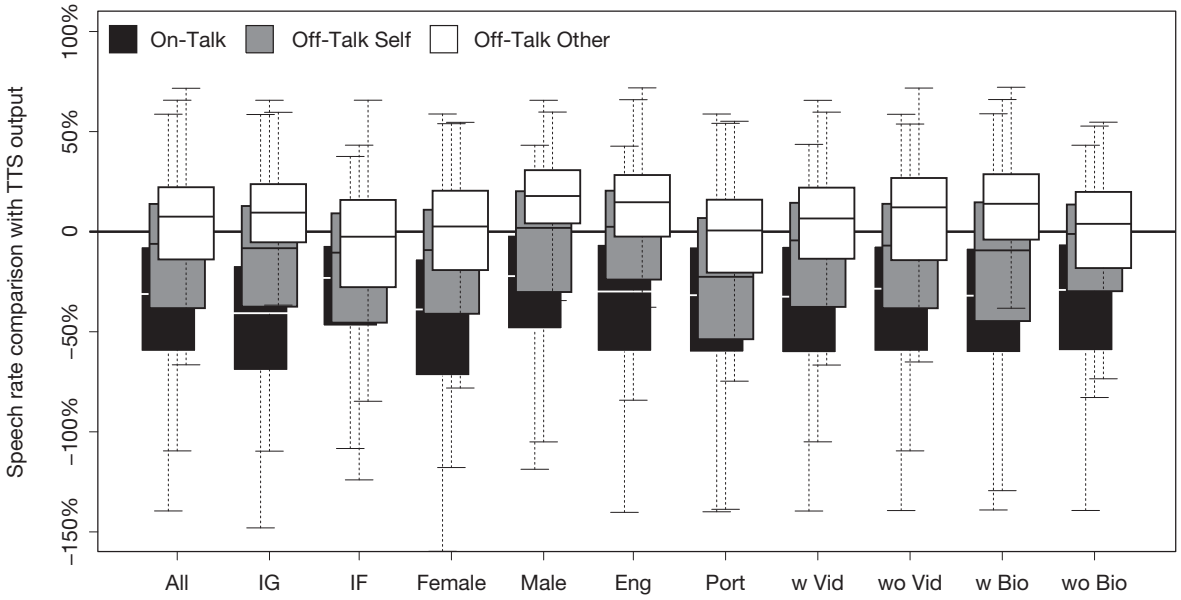


Figure 6.4: *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.*

6.5 Discussion

The statistical analysis of the three talk types (*On-Talk*, *Off-Talk Self*, and *Off-Talk Other*) have provided the following contributions to the understanding of interlingual computer mediated communication.

■ Key Findings.

- Though the speech rates of each talk type overlap, there is a clear indication that the subjects of the computer-mediated dialogue dataset use a different speech rate depending on to whom they are directing their utterance. Utterances directed towards the ILMT-s2s System (*On-Talk*) where spoken with a mean speech rate slower then the TTS output of 180 Word Per Minute (wpm) of the ILMT-s2s System. Furthermore, this *On-Talk* mean speech rate that is slower than the TTS output does not seem to be the subject's natural speech rate. When the subjects are directing their utterances to themselves (*Off-Talk Self*) or to another human in direct Human-to-Human communication (*Off-Talk Other*) the subjects mean speech rate is faster than when directed to the computer system.
- With utterances of two or more words, *On-Talk* is the slowest, followed by *Off-Talk Self* and then *Off-Talk Other* being the utterance type with the fastest mean value.
- When looking at single word utterances, the slowest utterances changes to *Off-Talk Self*, then *On-Talk*, with *Off-Talk Other* remaining as the utterance type with the fastest mean value.

The obvious question following these results is what is causing the hyperarticulation in *On-Talk*. The following is an extract from the draft submitted for review of my Speech Prosody paper (Hayakawa, Haider, et al., 2016) that I deleted for the camera ready draft after the reviewers asked for clarification that all subjects were also analysed seperately, and not only all the utterances as a whole. In the initial draft submitted for the review, I reported the observation of the data where the onset of slower speech was not obviously caused by the negative ASR performance.

> "[*I*] investigate whether there is a difference in the speech rate of *On-Talk* and *Off-Talk*. Initial analysis of the dialogue (Figure 6.5) have shown that within the first thirty seconds of the dialogue, there is no significant difference between the speech rate difference of *On-Talk* and *Off-Talk* ($F_{1,44} = 0.031; p = 0.862$). Even when the data is expanded to the first one hundred seconds, the significance is still small ($p = 0.048$).
>
> Furthermore, as previously studied in Stent, Huffman, & Brennan (2008) and Hayakawa, Cerrato, et al. (2015), a correlation between Word Error Rate (WER) and hyperarticulation has been identified. However, it was observed that of the fourteen subjects that start with one hundred percent accurate ASR results, the onset of hyperarticulation precedes the ASR result error (Figure 6.6). If it was a reaction of WER, then hyperarticulation

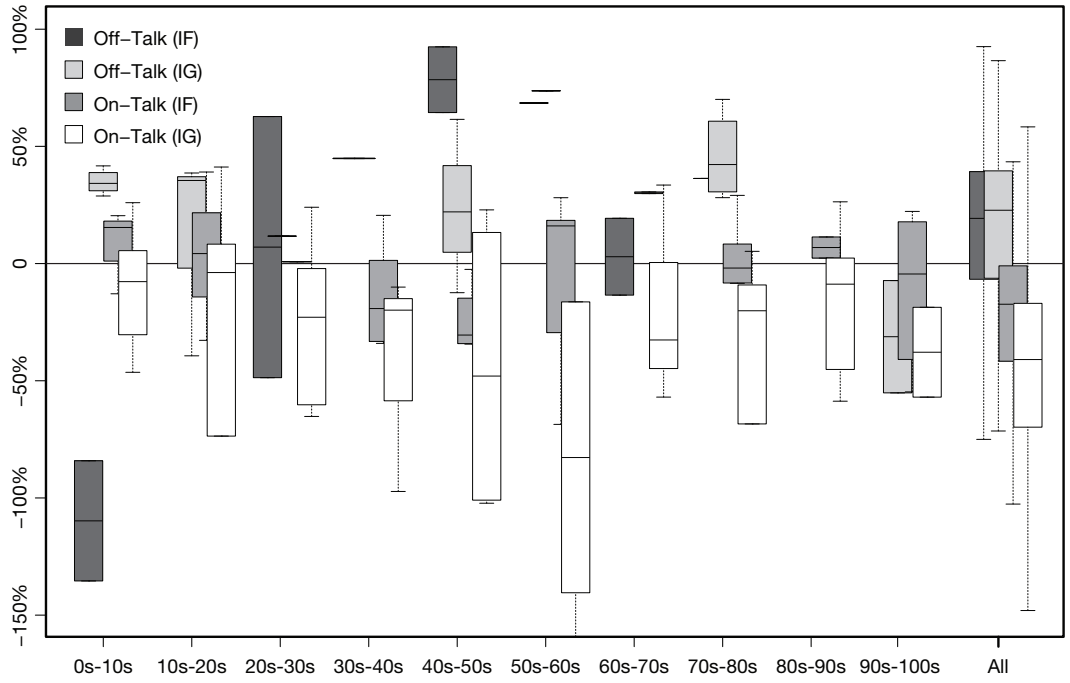


Figure 6.5: *10 sec. interval speech rate difference comparison for first 100 sec. of all dialogues*

should start at $(n+2)$ utterances. This is an indication that communication through the ILMT-s2s System was not the only cause of hyperarticulation for the subjects.

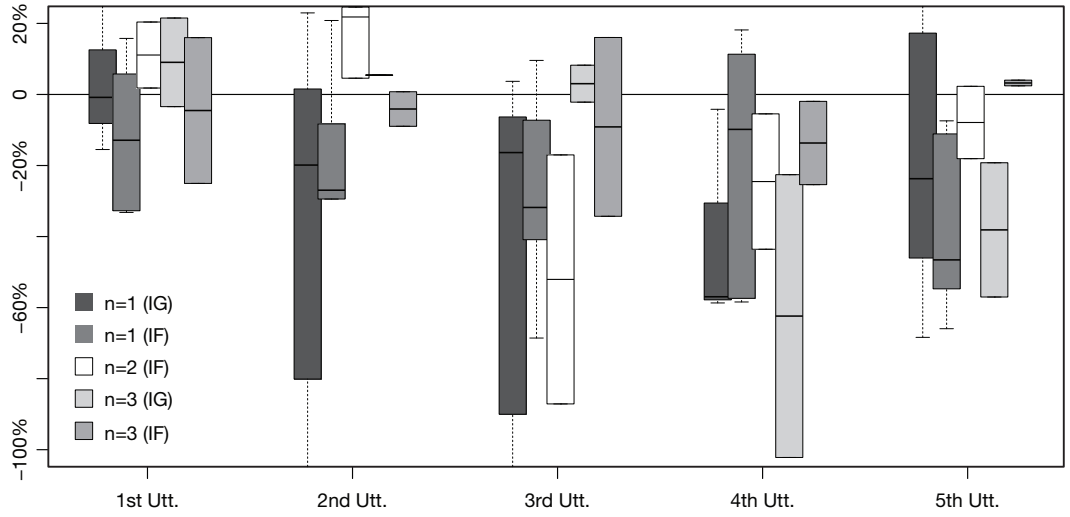


Figure 6.6: *First* (n) *utterances with 100% accurate ASR results and following* $(5-n)$ *utterances with ASR errors*

[...]

The main motivation of this study, apart from its novelty, was to verify if there was a distinguishable difference between *On-Talk*, *Off-Talk Self* and *Off-Talk Other* for a interactive system to provide better performance and a better understanding of the interlocutor. This was achieved with the clear significant difference, moderate Cohen's d estimate and good prosodic prediction results. However the sub-finding that hyperarticulation was not initiated by the ASR WER is of interest and also the significant difference between the

On-Talk of IG ($M = -47.05$, $SD = 44.43$) and IF ($M = -26.99$, $SD = 42.48$) in Figure 6.7 needs further investigation. It is easily imaginable that the perception of simplicity of the map task with the actual complexity of providing understandable instruction caused the initial hyperarticulation. Combine this with the difficulty of using the computer interaction systems may be the cause of this difference, and it will be interesting to see if the speakers of the original HCRC map task also displayed similar hyperarticulation differences."

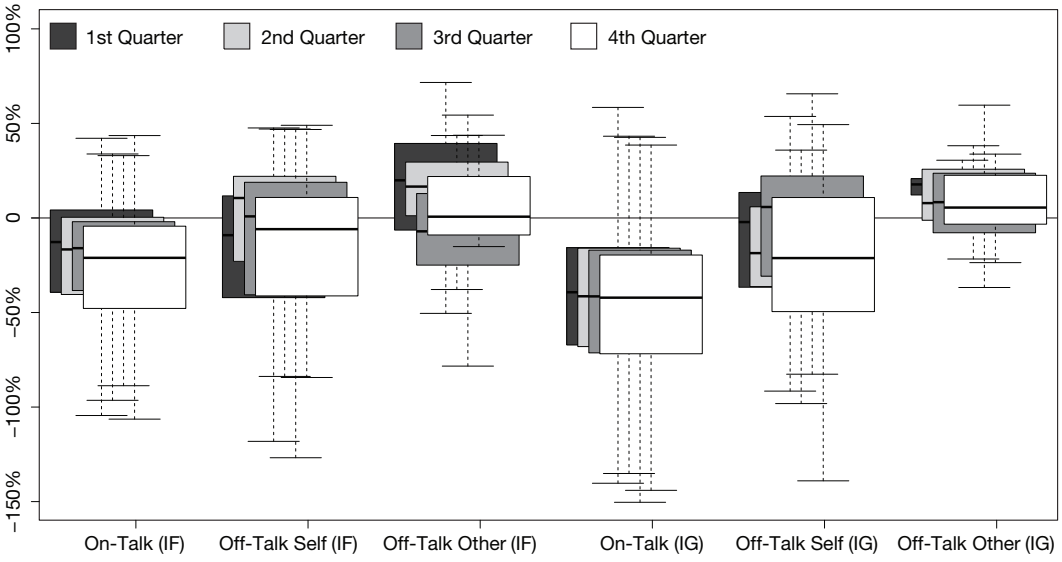


Figure 6.7: *Talk type speech rate difference comparison separated by IG, IF and dialogue quarter*

Figure 6.4 (and Figure 6.7) indicates that of the three talk types, the *On-Talk* speech rate for the Information Giver is slower than that of the Information Follower. Although further research is required in perhaps a Wizard-of-Oz situation where the ASR performance of the interlingual computer mediated communication system can be controlled to provide a 100% accurate reply for a fixed amount of time, I think it is highly likely that the onset of the slower speech rate (hyperarticulation) is a result of the difficulty of the task on hand as well as the adaptation to the communication style of the ILMT-s2s System. Branigan, Lickley, & McKelvie (1999, p. 389), in a study of disfluency in the HCRC Edinburgh Map Task corpus, indicated that the role of the subject had a strong effect on the disfluency rate of the subjects, with the role of Information Giver being more strongly affected:

> "Overall, these results suggest that conversational role has a considerable influence upon disfluency rate and type of disfluency. The different in rate cannot be wholly attributed to differences in utterance length. This is an important issue for future investigation." (Branigan et al., 1999, p. 389)

Though a slower speech rate is not the same as a higher disfluency rate, the data of the computer-mediated dialogue dataset is also showing that the Information Giver planning demands of the the role of Information Giver may be increasing the cognitive load of the subject (Oviatt, 1995) and causing the speech rate to drop even before the ASR performance causes any effect on the subject.

Chapter 7

Speech Rate Differences Between Datasets

7.1 Introduction

The computer-mediated dialogue dataset was created with the intention of comparing interlingual language-technology mediated communication with the direct Human-to-Human (H2H) communication of the non-mediated dialogue dataset. In this section, I compare the speech rate of the utterances in the computer-mediated dialogue dataset with the utterances of the non-mediated dialogue dataset to see if there is a common pattern between them. In the following Chapter 8: "Dialogue Acts Differences Between Datasets", the dialogue acts of the two datasets are compared.

As mentioned in Chapter 6: "The Three Talk Types", the analysis of the three talk types of the subject utterances in the computer-mediated dialogue dataset indicated that the subject talked more slowly when directing there utterances towards the ILMT-s2s System (*On-Talk*), while the subject's speech rate was faster when they were talking to others, in a direct Human-to-Human situation (*Off-Talk Other*). Furthermore, as mentioned in Chapter 9: "ASR Performance & Speech Rate", the speech rate was affected by the Automatic Speech Recognition (ASR) performance, as identified by Dr. Saturnino Luz, but there was also a clear affect on the speech rate of the subject by role, with the Information Giver being slower than the Information Follower, and also gender with the female subjects being slower than the male subjects while the ASR performance did not differ much for either.

From the comparison in this section with the non-mediated dialogue dataset, the speech rate of *Off-Talk Other* shows similarities with the overall speech rate of the non-mediated dialogue data, and similar differences were identified for the speech rate difference in role. However, though a significant difference was observed for the speech rate difference between genders, the difference in the non-mediated dialogue dataset were not as obvious as the difference in the computer-mediated dialogue data.

7.2 Dataset

In this chapter, the speech rate difference labels of the non-mediated dialogue dataset described in § 3.2.5.2 and the speech rate difference labels of the computer-mediated dialogue dataset described in § 3.3.3.4 of Chapter 3 are used for the analysis.

7.3 Methods

The data from the two datasets are compared to establish the following:

1. Compare the speech rate of *Off-Talk Other* with the dialogues of the non-mediated dataset.
2. Compare the speech rate of *On-Talk* with the dialogues of the non-mediated dataset.

7.3.1 Visualisation of the Dialogues' Speech Rate

Using the values of the speech rate (§ 5.5: "Definition") a graph was created to displays the flow of the speech rate for each dialogue.

As with the visualisation of the computer-mediated dialogues (the ILMT-s2s corpus), the speech rate values calculated for each dialogue act in the non-mediated dialogue data (the HCRC Edinburgh Map Task corpus) are normalised so that each recording session (dialogue) contained the same number of windows/frames (20 frames). An example of this normalisation is described in Table 7.1.

Dialogue number		D.01 (IG)	D.01 (IF)	D.02 (IG)	D.02 (IF)
Dialogue Act Count per Dialogue		88	66	44	36
Duplicate each value	×	5	10	5	5
Total number of values		440	660	220	180
Group values (windows)	÷	22	33	11	9
Number of window samples		20	20	20	20

Table 7.1: *Sample window alteration example*

7.3.2 Comparison Between *Off-Talk Other* and the Non-Mediated Dialogues

The purpose of this comparison is to verify if the speech rate of the subjects in the computer-mediated dialogues are similar to the non-mediated dialogues when they are talking directly with another human and not the communication system. This comparison is made to verify that any difference in the three talk type's speech rates are not a result of slow or fast speaking subjects in the computer-mediated dataset.

- The *word count* of each utterance in the both datasets are tested to verify if there is a significant difference (*Mann-Whitney U test*).[1]

[1]This verification is to reduced the possibility of the utterance length affecting the speech rate (Yuan et al., 2006).

- The *speech rate* of each utterance in the both datasets are tested to verify if there is a significant difference (*Mann-Whitney U test*).

7.3.3 Identifying Speech Rate patterns in the Computer-Mediated Dialogues

To identify the patterns between datasets, the English, *On-Talk*, multi word utterances of the computer-mediated dialogue dataset (the ILMT-s2s corpus) data was used to verify any significant difference between the groupings of the subjects.

The speech rate of utterances in the computer-mediated dialogues of *[Refer to grouping list below]* is tested to verify if there is a significant difference (*Mann-Whitney U test*).
Grouping list:

1. Role (Information Giver or Information Follower)
2. Gender (female or male)
3. Setting (w/ Video or w/o Video)

7.3.4 Identifying Speech Rate patterns in the Non-Mediated Dialogues

As verified for the computer-mediated dialogues in § 6.3, the speech rate is analysed to verified the patterns of the various groupings of the subjects.

The speech rate of utterances in the non-mediated dialogues of *[Refer to grouping list below]* is tested to verify if there is a significant difference (*Mann-Whitney U test*).
Grouping list:

1. Role (Information Giver or Information Follower)
2. Gender (female or male)
3. Setting (w/ Eye-Contact or w/o Eye-Contact)
4. Familiarity (Familiar or Unfamiliar)

Methodological Note
Familiarity of the subjects is not a grouping criteria in the computer-mediated dialogues. All subject pairs of the computer-mediated dialogues are unfamiliar with each other. Therefore they have been added to the groupings to verify that familiarity does not apply a significant influence to the results.

7.3.5 Comparison Between *On-Talk* and the Non-Mediated Dialogues

The speech rate used in the computer-mediated dialogues and the non-mediated dialogues was expected to be different (Oviatt et al., 1998; L. Bell et al., 2003; Oviatt, Darves, & Coulston, 2004; Suzuki & Katagiri, 2007). Therefore, once the mean speech rate of both datasets are adjusted to be the same, the data of both datasets are compared to identify any differences within the groupings.

The speech rate of *[Refer to grouping list below]* utterances in the computer-mediated dialogues and the non-mediated dialogues are tested to verify if there is a significant difference (*Mann-Whitney U test*).

Grouping list:

1. Role: Information Giver
2. Role: Information Follower
3. Gender: Female
4. Gender: Male
5. Setting: w/ Eye-Contact (w/ Video)
6. Setting: w/o Eye-Contact (w/o Video)
7. All

Methodological Note

The settings of "*with* or *without eye-contact*" and "*with* or *without video*" are considered as equal. The method used during the two different data collections are different, but the effect of "*having* or *not having visual contact*" with the interlocutor is the same.

7.4 Results

The purpose of this comparison is to see where the speech rate patterns of the computer-mediated dialogue dataset differ from the non-mediated dialogue dataset. In Chapter 6 and Chapter 9, the data of the computer-mediated dialogue dataset was compared to itself, and to reported theories. As indicated in Figure 7.1, the speech rate for *On-Talk* in the computer-mediated dialogue dataset is slower than that of the utterances in the non-mediated dialogue dataset. Up to this point of this thesis, I have yet to report if the subjects of the computer-mediated dialogue data collection are a fluke selection of slow speakers or not.

7.4.1 Visualisation of the Dialogues' Speech Rate

Using the speech rate values that are calculated in § 5.5, the values of each dialogue are standardised into 20 frame windows and visualised as Figure 7.1.

■ The Individual Dialogue Charts.

For the individual dialogue charts, refer to the following figures in § B.3 of Appendix B.

The non-mediated dialogue data (the HCRC Edinburgh Map Task corpus):

Figure B.14: Individual dialogue figures — Dialogues 1 to 4.

Figure B.15: Individual dialogue figures — Dialogues 5 to 8.

Figure B.16: Individual dialogue figures — Dialogues 9 to 12.

Figure B.17: Individual dialogue figures — Dialogues 13 to 16.

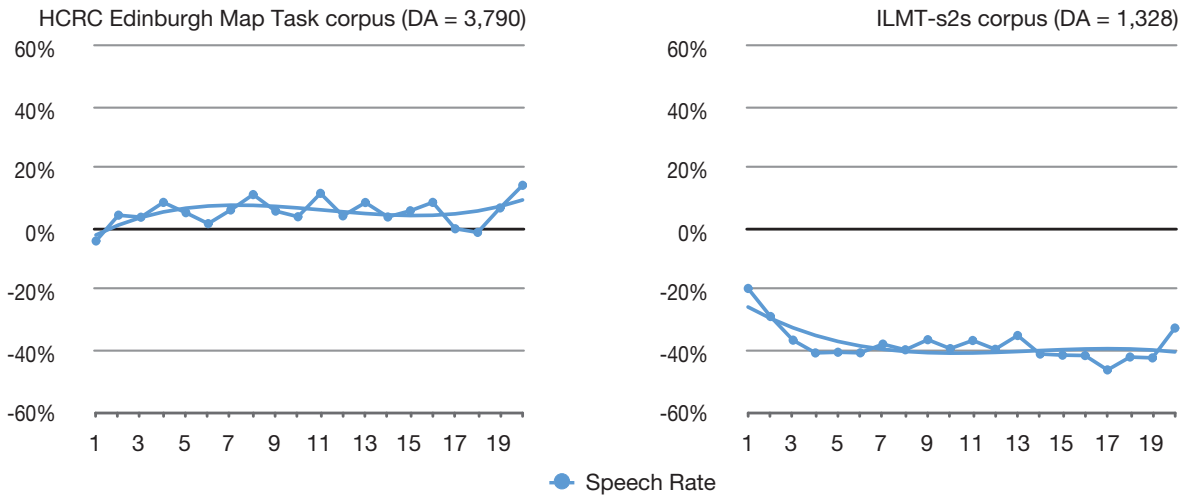


Figure 7.1: *Mean speech rate of the HCRC Edinburgh Map Task corpus (n = 24 subjects and 16 dialogues) and the ILMT-s2s corpus (n = 15 subjects and 15 dialogues)*

The computer-mediated dialogue data (the ILMT-s2s corpus):

Figure A.14: Individual dialogue figures — Dialogues 1 to 4.

Figure A.15: Individual dialogue figures — Dialogues 5 to 8.

Figure A.16: Individual dialogue figures — Dialogues 9 to 12.

Figure A.17: Individual dialogue figures — Dialogues 13 to 15.

■ The Dialogue Charts by Grouping.

For the charts of the grouping within each dataset, refer to the following figures in § B.3 of Appendix B.

The non-mediated dialogue data (the HCRC Edinburgh Map Task corpus):

Figure B.2: Subject speech rate means by role.

Figure B.3: Subject speech rate means by gender, eye contact setting, map, and familiarity.

Figure B.4: Subject speech rate means of roles divided by gender and eye contact setting.

Figure B.5: Subject speech rate means of roles divided by map and familiarity.

Figure B.6: Subject speech rate means of gender divided by role and eye contact setting.

Figure B.7: Subject speech rate means of gender divided by map used and familiarity.

Figure B.8: Subject speech rate means of eye contact divided by role and gender.

Figure B.9: Subject speech rate means of eye contact divided by map used and familiarity.

Figure B.10: Subject speech rate means of map divided by role and gender.

Figure B.11: Subject speech rate means of map divided by eye contact and familiarity.

Figure B.12: Subject speech rate means of familiarity divided by role and gender.

Figure B.13: Subject speech rate means of familiarity divided by eye contact and map used.

The computer-mediated dialogue data (the ILMT-s2s corpus):

Figure A.3: Grouping by role, spoken language, video setting and crashes during video usage.

Figure A.4: Grouping of gender divided by role and spoken language.

Figure A.5: Grouping of gender divided by video setting and crashes during video usage.

Figure A.6: Grouping of roles divided by gender and spoken language.

Figure A.7: Grouping of roles divided by video setting and crashes during video usage.

Figure A.8: Grouping of spoken languages divided by role and gender.

Figure A.9: Grouping of spoken languages divided by video setting and crashes during video usage.

Figure A.10: Grouping of video setting divided by role and spoken language.

Figure A.11: Grouping of video setting divided by gender and crashes during video usage.

Figure A.12: Grouping of dialogues with systems crashes divided by role and spoken language.

Figure A.13: Grouping of dialogues with systems crashes divided by gender.

7.4.2 Comparison Between *Off-Talk Other* and the Non-Mediated Dialogues

The speech rates of the three talk types of the computer-mediated dialogues and that of the non-mediated dialogues are plotted in Figure 7.2. The box plots of *Off-Talk Other* and the dialogues of

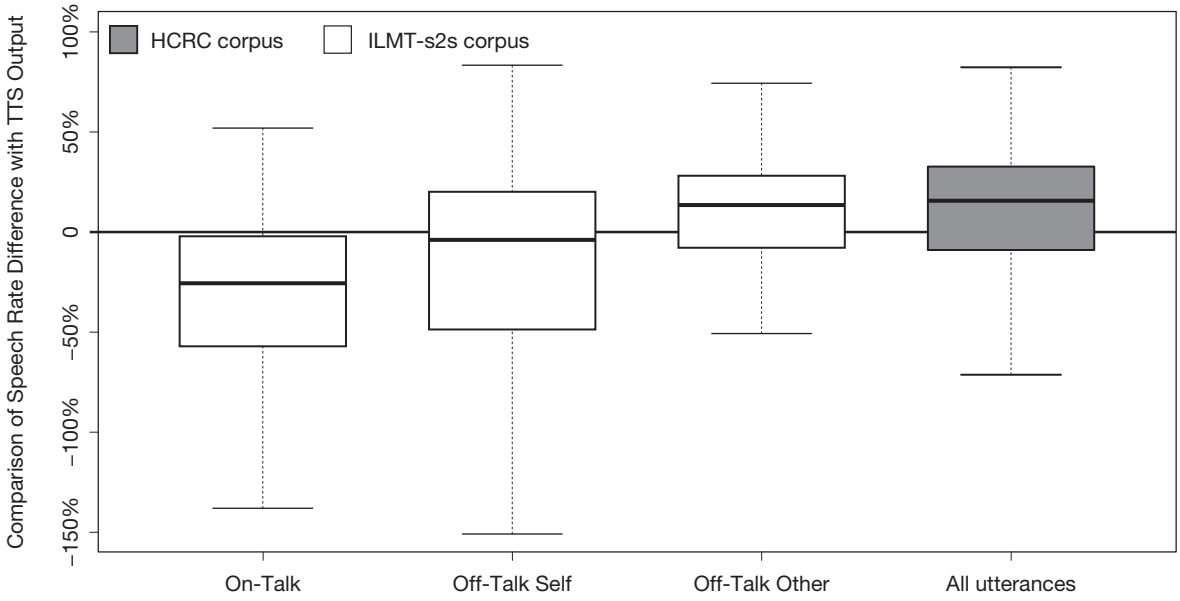


Figure 7.2: *Speech rates boxplots of both datasets*

the non-mediated dialogues show a similar "1st Qu.", "*Mdn*", "*M*" and "3rd Qu." values, as indicated in Table 7.2.

Speech Rate — English	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC corpus	−547.20	−8.94	15.65	6.06	32.69	82.25	43.73	3,790
ILMT-s2s *Off-Talk Other*	−146.90	−7.82	13.48	5.39	28.12	74.29	37.51	169
ILMT-s2s *Off-Talk Self*	−875.90	−48.58	−3.89	−27.94	20.11	83.30	85.04	483
ILMT-s2s *On-Talk*	−314.80	−57.01	−25.54	−33.33	−2.10	51.88	45.24	1,328

Table 7.2: *Summary of the various speech rates*

Off-Talk Other is defined as direct communication with a fellow human. The speech rate of *Off-Talk Other* was therefore compared with the speech rate values of the non-mediated dialogues, which is a collection of direct communication with a fellow human.

- The word count of the utterances was verified and returned to indicate that no significant difference was observed, and the effect size was negligible (Mann-Whitney U test: $U = 309,010,\ p = 0.4274$, Cliff's δ estimate: 0.035).
- Analysis of the two speech rates resulted with no significant difference being observed (Mann-Whitney U test: $U = 336,750,\ p = 0.2566$). Also the effect size resulted in a negligible estimate (Cliff's δ estimate: 0.051).

7.4.3 Identifying Speech Rate patterns in the Computer-Mediated Dialogues

To identify the patters between datasets, the English, *On-Talk*, multi word utterances of the computer-mediated dialogue data (the ILMT-s2s corpus) was used to verify any significant difference between the groupings of the subjects. From this verification a Mann-Whitney U test of the groupings of role, gender, and setting indicate that there is a significant difference within the following groupings (refer to Table 7.3 for the summary of each grouping);

- The gender of the subjects as female or male ($U = 115,720,\ p < 2.2e-16$),
- The role of subject as Information Giver or Information Follower ($U = 253,670,\ p = 1.985e{-}09$),
- The setting of the subjects as w/ Video or w/o Video ($U = 240,930,\ p = 0.001908$),

Speech Rate — English	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
Female	−314.80	−78.70	−44.45	−53.80	−21.14	34.79	49.18	418
Male	−238.70	−44.52	−17.81	−23.93	4.08	51.88	39.97	910
Information Giver	−257.50	−62.03	−31.86	−38.02	−6.94	51.88	43.42	790
Information Follower	−314.80	−46.68	−16.93	−26.46	3.82	41.96	47.00	538
w/ Video (w EC)	−314.80	−58.59	−29.67	−36.06	−4.65	51.88	45.22	712
w/o Video (wo EC)	−257.50	−53.04	−20.71	−30.18	1.63	41.96	45.11	616

Table 7.3: *Summary of the speech rates with a significant difference.*

7.4.4 Identifying Speech Rate patterns in the Non-Mediated Dialogues

For the non-mediated dialogue dataset (the HCRC Edinburgh Map Task corpus) data, a Mann-Whitney U test of the groupings of role, gender, setting, and familiarity indicate that there is a

significant difference within the following groupings (refer to Table 7.4 for the summary of each grouping);

- The role of subject as Information Giver or Information Follower ($U = 857,890,\ p < 2.2e-16$),
- The gender of the subjects as female or male ($U = 671,600,\ p = 0.0093$),
- The familiarity of the subjects as familiar with each other or not ($U = 733,090,\ p = 0.00549$).

Speech Rate — English	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
Information Giver	−438.70	−30.53	0.66	−10.48	20.71	69.33	47.31	1,540
Information Follower	−379.90	−1.84	17.17	9.16	30.02	72.00	37.61	857
Female	−438.70	−18.92	6.07	−4.22	23.15	72.00	43.96	1,271
Male	−379.90	−19.94	11.38	−2.60	27.10	69.63	46.30	1,126
Familiar	−438.70	−15.66	9.79	−1.04	25.53	72.00	42.45	1,448
UnFamiliar	−379.90	−27.13	5.23	−7.15	24.44	69.63	48.60	949
w EC	−379.90	−22.00	7.94	−5.14	25.76	69.06	46.99	1,107
wo EC	−438.70	−16.53	8.04	−2.01	24.76	72.00	43.33	1,290

Table 7.4: *Summary of the speech rates in the non-mediated dialogue data.*

However no significant difference was observed for the setting, with or without eye-contact (EC) during the recordings ($U = 726,180,\ p = 0.4716$).[2]

7.4.5 Comparison Between *On-Talk* and the Non-Mediated Dialogues

Of the groupings used in § 7.4.4, familiarity ("Familiar" and "Unfamiliar") is removed since the equivalent is not available in the computer-mediated dialogues.[3]

From the summary of the two datasets in Table 7.2, the difference between the median (*Mdn*) of the non-mediated dialogue data at 15.65%, and the median (*Mdn*) of the computer-mediated dialogue *On-Talk* data of −25.54%, was calculated as 41.19%. This value of 41.19 was added to all the *On-Talk* speech rates so that the median value of the two datasets would align. The alignment of both datasets resulted in Figure 7.3.

7.4.6 Verifications with Adjusted Values

Using the newly adjusted values of *On-Talk*, a significance test was performed between the grouping of the two datasets. For example, the male subjects of the computer-mediated dialogue data and the non-mediated dialogue data are tested. The results of these tests are listed below in Table 7.5.
As observable from Table 7.5, a significant difference was identified between the following.

- Between the female subjects of the two datasets.
- Between the male subjects of the two datasets.
- Between the setting of without eye-contact (w/o EC) of the two datasets.

[2]The settings of with or without eye-contact (EC) is regarded as the equivalent of the w/ Video and w/o Video in the ILMT-s2s corpus.
[3]The items of "w EC" and "wo EC" are the equivalent of the setting differences "w/ Video" and "w/o Video" that are described in Chapter 9

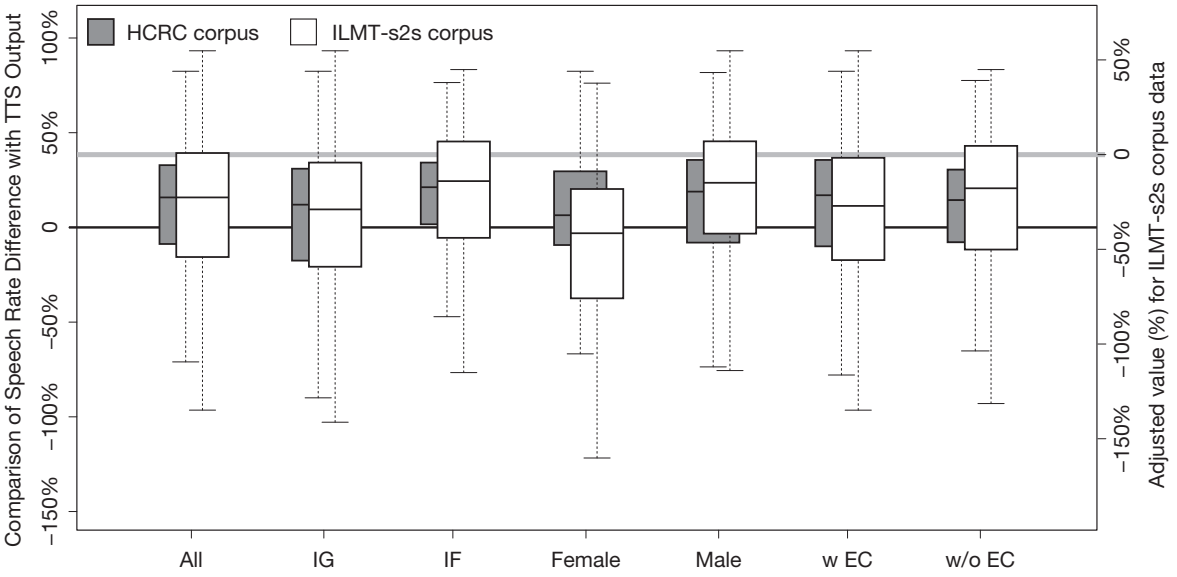


Figure 7.3: *Speech rate comparison with the values from the ILMT-s2s corpus shifted up so that the median male speech rate matches that of the HCRC Edinburgh Map Task corpus (ILMT-s2s corpus values raised by 41.19%)*

Speech Rate Comparison After Adjustment — HCRC Edinburgh Map Task corpus : ILMT-s2s corpus							
Grouping	*Mdn* (HCRC : ILMT)	*M* (HCRC : ILMT)	*SD* (HCRC : ILMT)	wilcox.test	sig.	cliff.delta	ES
Female	13.21 : −3.26	4.76 : −12.61	41.4 : 49.2	$p = 6.484e-15$	***	$\delta = 0.242$	small
Male	18.72 : 23.38	7.52 : 17.26	46.2 : 40.0	$p = 8.615e-08$	***	$\delta = 0.126$	neg.
wo EC	14.26 : 20.48	5.86 : 11.01	40.8 : 45.1	$p = 2.684e-05$	***	$\delta = 0.112$	neg.
IF	21.04 : 24.26	12.92 : 14.73	38.9 : 47.0	$p = 0.0029$	**	$\delta = 0.086$	neg.
All	15.65 : 15.65	6.06 : 7.86	43.7 : 45.2	$p = 0.1019$	–	$\delta = 0.030$	neg.
w EC	16.81 : 11.52	6.29 : 5.13	46.8 : 45.2	$p = 0.1072$	–	$\delta = 0.041$	neg.
IG	11.83 : 9.33	1.11 : 3.17	46.3 : 43.4	$p = 0.7108$	–	$\delta = 0.009$	neg.

Table 7.5: *Significance test and effect size between datasets (ILMT-s2s corpus values raised by 41.19%)*

- Between the Information Followers of the two datasets.

Though four items had a significant difference, only one item, the differences of female subjects had an effect size of small, while the others displayed only a negligible effect size.

7.5 Summary

The comparison of the non-mediated dialogue data and the computer-mediated dialogue data speech rates have provided the following contributions to the understanding of interlingual computer mediated communication.

■ Key Results.

- The speech rate of direct Human-to-Human communication (*Off-Talk Other*) in the computer-mediated dialogue dataset is similar to the speech rate of the non-mediated dialogue dataset subjects, indicating that the subjects of the computer-mediated dialogue dataset are not a fluke selection of slow speakers.

- A significant difference and a large effect size was reported for utterances spoken to the ILMT-s2s System (*On-Talk*) and those of the H2H dialogues in the non-mediated dialogue dataset. The subjects used a slower speech rate when talking to their interlocutor in the interlingual computer mediated communication.

- The role of the subject affects the speech rate of the subject in both datasets. The Information Giver uses a slower speech rate as compared to the Information Follower, with the adjusted comparison showing similar rates in decreased speech in both datasets.

- The setting difference of being able to see the interlocutor or not during the recordings does not affect the speech rate of the subjects in either datasets.

- Gender speech rate patterns differ between datasets. Subjects of the computer-mediated dialogue dataset show female subjects using a slower speech rate than male subjects within the corpus, while the subjects do not show such a big difference in the non-mediated dialogue dataset. As illustrated in Figure 7.4, and more obvious in the adjusted Figure 7.5.

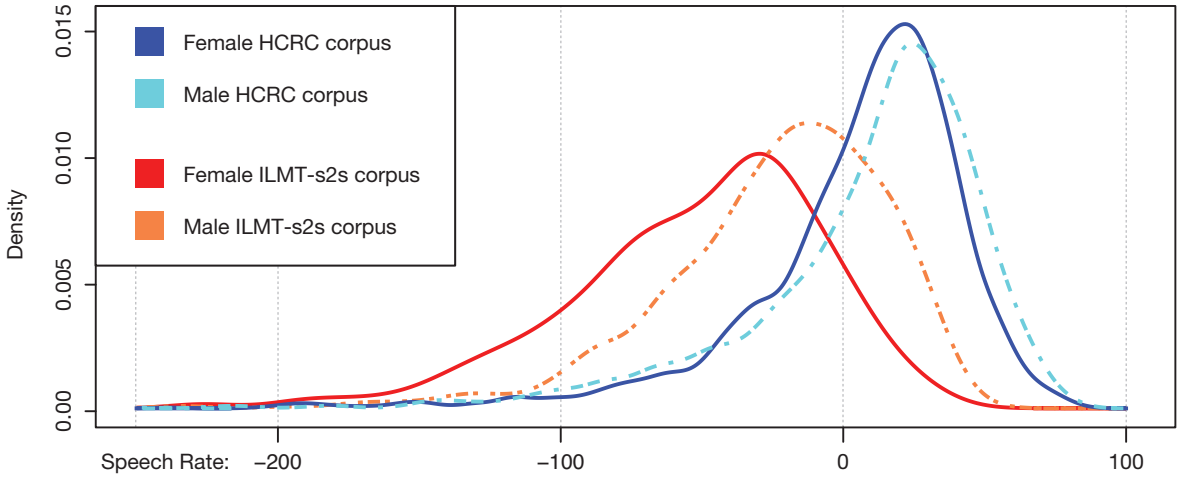


Figure 7.4: *Density plot of gender speech rates in both datasets — Without value adjestments*

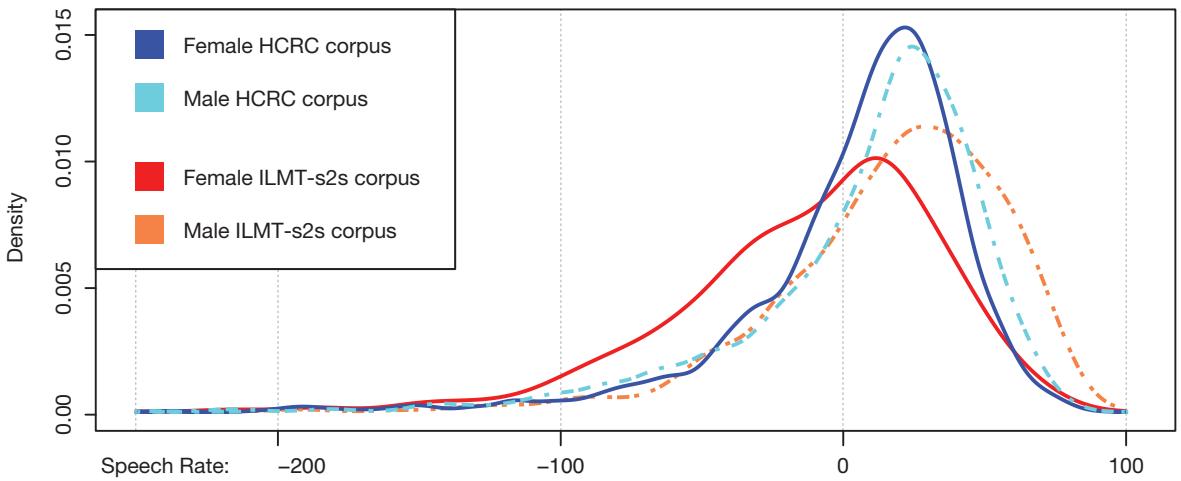


Figure 7.5: *Density plot of gender speech rates in both datasets — After ILMT-s2s corpus values raised by 41.19% to align the median of all values with the HCRC Edinburgh Map Task corpus*

7.6 Discussion

In this thesis, the subject speech rate has been analysed by looking at the mean values of the speech rate explained in Chapter 5. I have provided details of how the mean speech rate changes by role (Information Giver or Information Follower), and that it is also affected by the gender of the subject (female or male). I have also indicated that the mean speech rates for the three talk types of *On-Talk*, *Off-Talk Self* and *Off-Talk Other*, identified in the computer-mediated dialogue data differ.

However, what I have not studied is the effect of the utterance word count, on the speech rate of the subject. Yuan et al. (2006, p. 542), in their study of the Switchboard corpus (Godfrey et al., 1992) and the Chinese CallHome and CallFriend corpus (Yuan & Jurafsky, 2005) have indicated that utterances speech rates between one and seven words change drastically:

> "We can see that in both English and Chinese, there is an abrupt rise of speaking rate for the segments containing from one to seven words. For the segments having eight to about 30 words, however, the speaking rate stays level. And then, especially in English, the speaking rate rises again, but with a more gradual slope." (Yuan et al., 2006, p. 542)

A similar phenomenon was observed with the wpm measurements of the TTS output, Information Giver, and Information Follower subjects of the computer-mediated dialogue dataset as illustrated in Figure 7.6. From the boxplots in Figure 7.6 it is possible to observe that the mean wpm speech rate

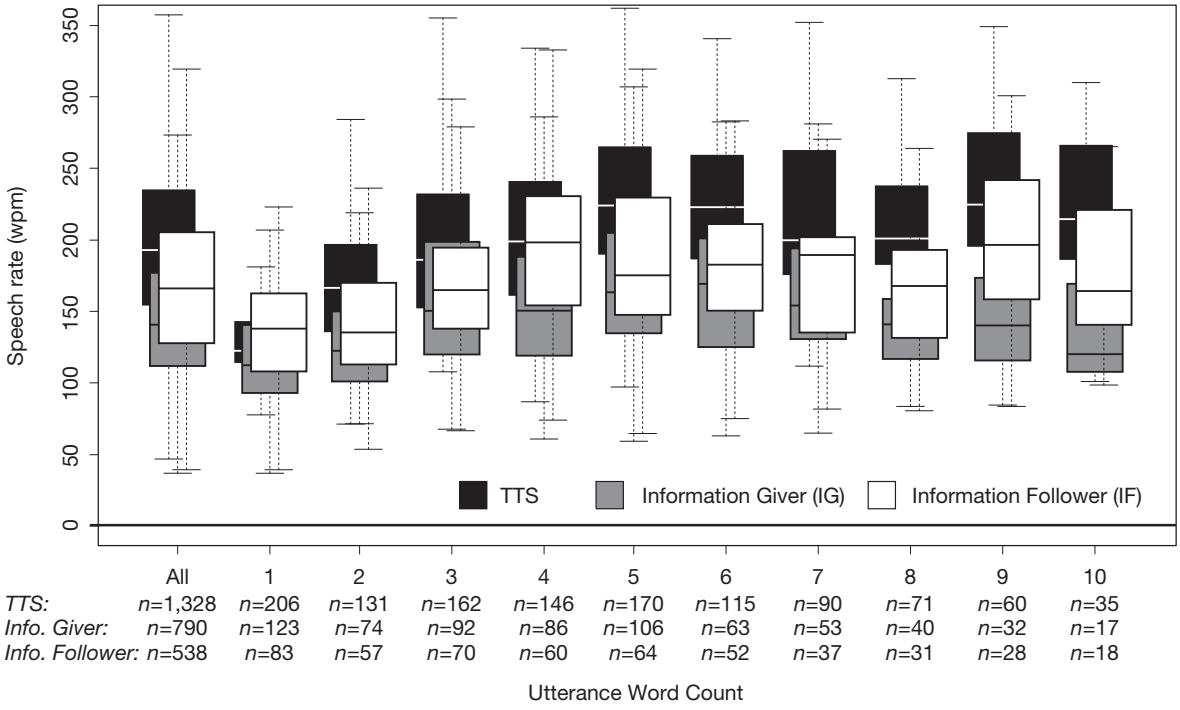


Figure 7.6: *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.*

increases up to utterances with a word count of 5 and then starts to fluctuate after that. Since the computer-mediated dialogue data has indicated that there is a speech rate difference between female and male subjects, the grouping of gender was verified and illustrated as Figure 7.7. From Figure 7.7

it is possible to observe that the TTS output and male subjects' speech rate (wpm) increase rapidly from one word to four or five words, while the female subjects' speech rate (wpm) does not increase much and remains basically flat, which is an observation that was not reported by Yuan et al. (2006). The gender difference reported by Yuan et al. (2006, pp. 543–544) was that the female subjects had a

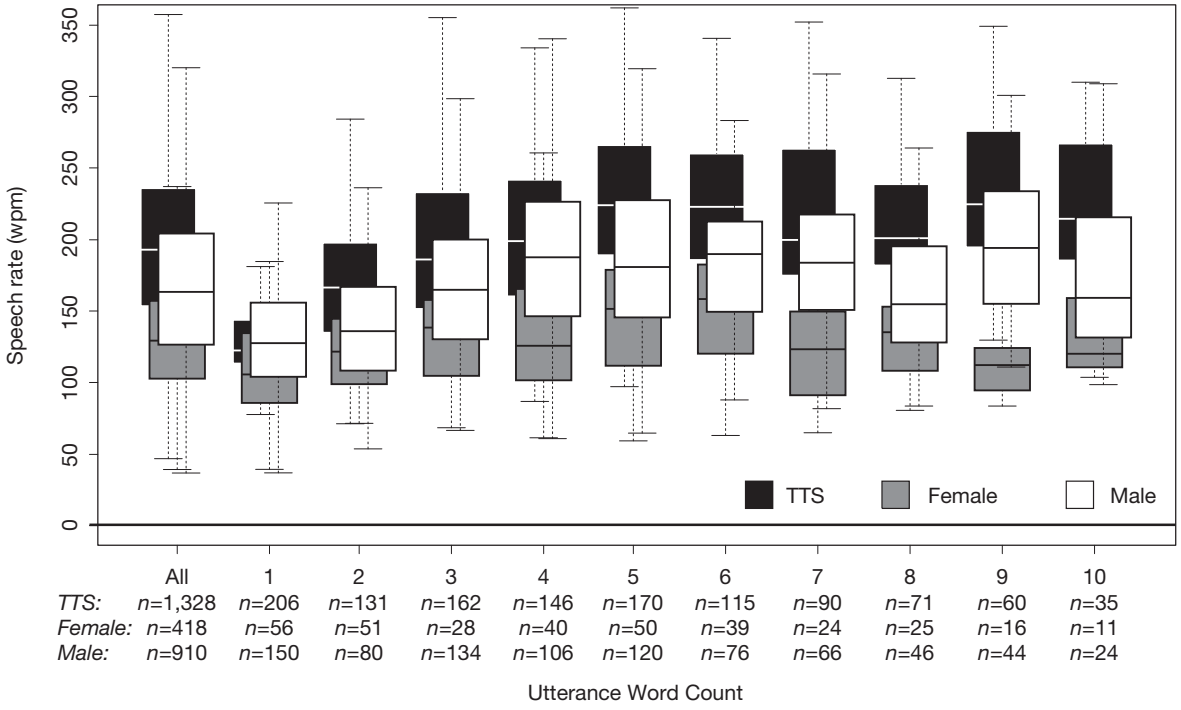


Figure 7.7: *Boxplot of the 3 talk types' multi word utterance speech rates in various groupings.*

slower speech rate, but with a minor difference to the male subjects.

> "Males tend to speak faster than females [. . .] difference between them is, however, very small, only about 4 to 5 words or characters per minute (2%), though it is statistically significant. It might be due to things that we would not normally think of as speech-rate parameters, such as differences in word-frequency distributions." (Yuan et al., 2006, pp. 543–544)

To verify if this pattern also exists in the non-mediated dialogue data, the wpm speech rate divided by gender and role were compared as illustrated in Figure 7.8, with similar curves in the graphs being observed. This indicated that the data of the non-mediated dialogue dataset also shows a curvature similar to that reported by Yuan et al. (2006) and also a curvature more similar to the TTS output than the data of the computer-mediated dialogue dataset. However the increase that Yuan et al. reported to start with utterances about 30 words long, can be seen to start from ≈12 words for both the subjects of non-mediated dialogue dataset and the TTS output. It is important not to forget that the analysis by Yuan et al. was performed on the SwitchBoard corpus (Godfrey et al., 1992) which is a collection of telephone dialogues. As previously mentioned in § 2.2.2: "Dialogue Acts and Communication Medium", I have mentioned that Morikawa & Maesako (1998, p. 149) in their reference to the publication in Japanese of Yoshida and Kakuta of a study in audio telephone

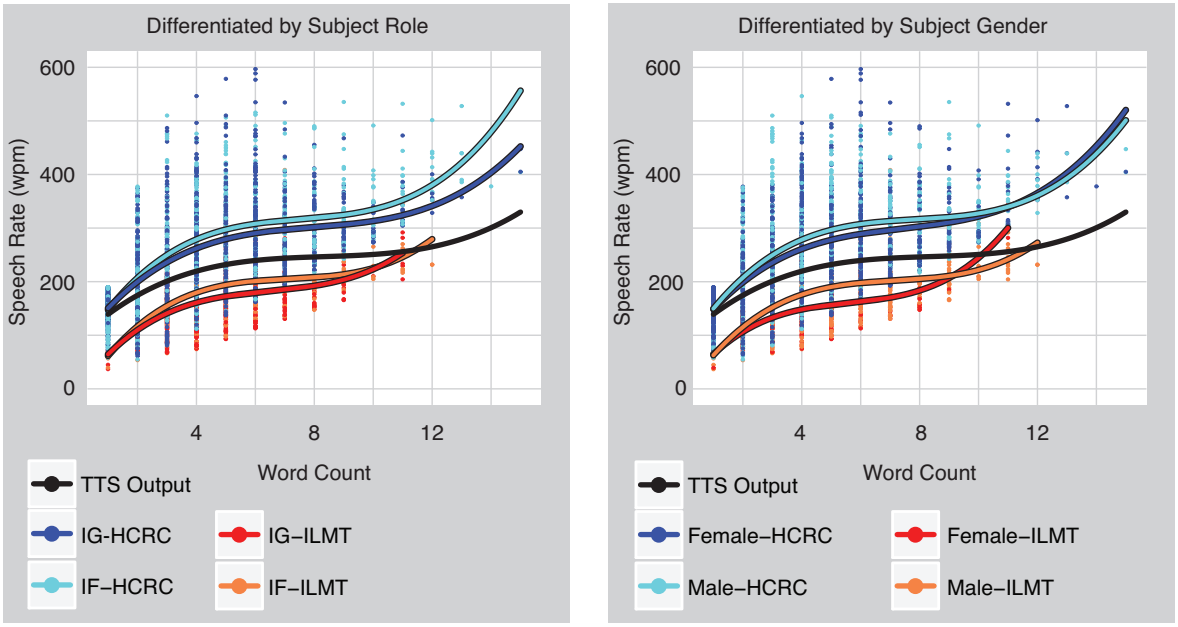


Figure 7.8: *Scatter plot of ILMT-s2s corpus utterance durations by word count*

communication, indicate that the subjects of their study consider communication over the telephone as a different alternative to Face-to-Face communication with its own communication style. Due to the different communication methods and the fundamental fact that the data being analysed is different, identical results are not expected, but the fundamental pattern is present in both the non-mediated dialogue dataset and the computer-mediated dialogue dataset.

Returning to the point of how this affects the data or results from the computer-mediated dialogue data in this thesis. The results from Yuan et al. (2006) have the potential to place doubt onto the results concerning the speech rates that I have reported in this thesis. This is because 60% of the data in the computer-mediated dialogue dataset are located between a word count of 2 words (20$^{\text{th}}$ percentile) and 8 words (80$^{\text{th}}$ percentile). This range between 2 and 8 words is where Yuan et al. (2006, p. 542) mention that there is an "abrupt rise of speaking rate". The median word count for the Information Giver and the female subjects are five and the median word count for the Information Follower and male is four. Following this theory, the speech rate of the Information Follower should have a slower wpm speech rate than the Information Giver, and male subjects too should have a slower wpm speech rate than the female subjects. From Table 7.6 the Information Giver and female subjects

ILMT-s2s corpus (English wpm speech rate)	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Count
Subject Role								
Information Giver — 4 word utt.	60.7	119.1	150.5	163.9	188.2	390.9	68.63	86
Information Follower — 4 word utt.	74.0	154.5	198.3	194.2	228.6	362.0	61.39	60
Information Giver — 5 word utt.	59.2	134.9	163.4	171.3	204.9	307.1	55.74	106
Information Follower — 5 word utt.	64.6	148.7	175.2	186.2	229.4	319.5	61.74	64
Subject Gender								
female — 4 word utt.	61.3	101.8	125.7	141.0	160.4	332.4	58.89	40
male — 4 word utt.	60.7	146.6	187.6	189.7	226.3	390.9	65.55	106
female — 5 word utt.	59.2	112.7	151.5	153.0	177.9	287.9	52.19	50
male — 5 word utt.	64.6	145.7	180.8	186.8	227.3	319.5	58.09	120

Table 7.6: *Summary and standard deviation of 4 and 5 word utterances (wpm)*

and the combination of four word Information Giver and five word Information Follower utterances follow this trend of a faster speech rate for utterances with more words, but it is not the case for the Information Follower, male subjects or the difference between four word Information Follower and five word Information Giver utterances. In both cases, the effect of the role of the subject (Information Giver being slower than the Information Follower) or the gender of the subject female being slower than the male) is stronger than the influence of the utterance word count. For this reason, the wording of "an abrupt rise of speaking rate for the segments containing from one to seven words" by Yuan et al. (2006, p. 542) would not be appropriate for the data of the computer-mediated dialogue dataset.

■ After Note.

A further point of interest is the fact that Yuan et al. (2006, pp. 544) also identified the ambiguity of the wpm measurement method caused by the different "word-frequency distributions" when describing the gender speech rate difference: "It might be due to things that we would not normally think of as speech-rate parameters, such as differences in word-frequency distributions". When Figure 7.8 which

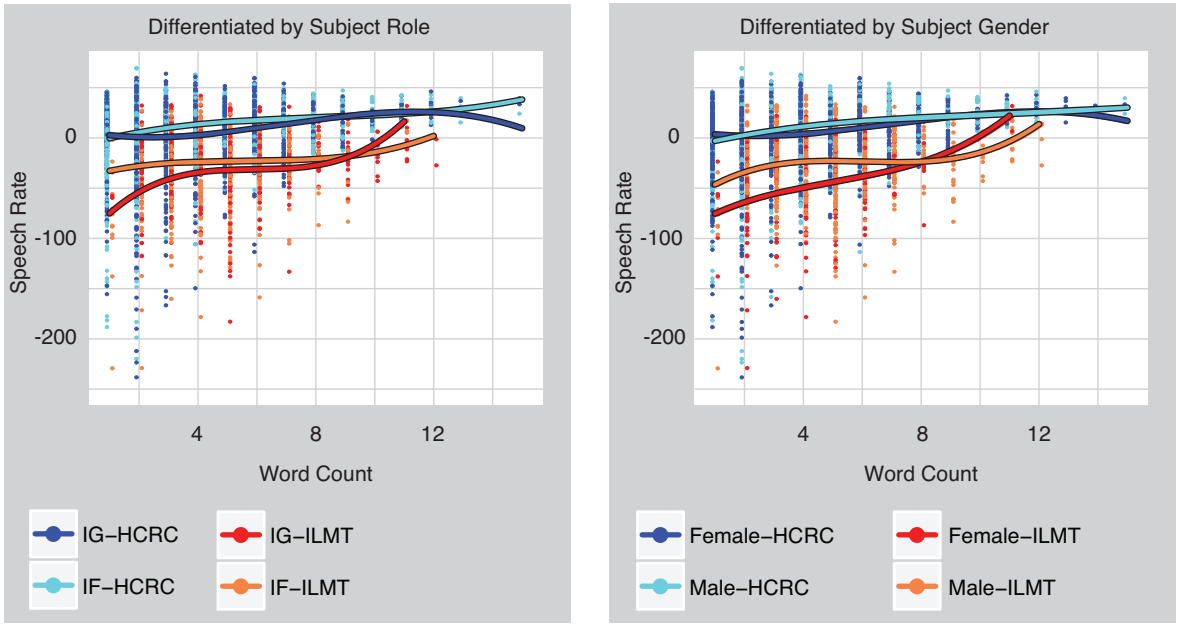


Figure 7.9: *Scatter plot of ILMT-s2s corpus utterance durations by word count*

used the wpm as the speech rate indicator is represented using the speech rate calculation method of comparing the TTS output duration with the subject utterance duration explained in § 5.1.2, it is possible to see that the speech rate of the subjects do not show "an abrupt rise" as indicated in Figure 7.9 with the lines representing the non-mediated dialogue data closer to a linear model than a non-linear model from 1 word all the way to 12 word utterances. This is both an indication of (i) how close the TTS output is to direct human communication, and (ii) the lack of "an abrupt rise" in the speech rate, with the word-frequency distributions no longer affecting the measurement values of the method (§ 5.1.2) used in this thesis.

Chapter 8

Dialogue Acts Differences Between Datasets

8.1 Introduction

The computer-mediated dialogue dataset (the ILMT-s2s corpus) specifically used the same maps as the non-mediated dialogue dataset (the HCRC Edinburgh Map Task corpus) so that a comparison could be made between the two communication styles — the comparison between direct Human-to-Human communication and machine mediated interlingual communication. The annotation of the dialogue acts (§ 4.4.3.2) of the computer-mediated dialogue data was also conducted to facilitate the comparison of the two datasets by using an annotation scheme similar to the one used in the non-mediated dialogue data.

8.2 Dataset

In this chapter, the dialogue act labels of the non-mediated dialogue dataset described in § 3.2.5.1 and the dialogue act labels of the computer-mediated dialogue dataset described in § 3.3.3.1 of Chapter 3 are used for the analysis.

8.3 Methods

The medium used to convey communication is said to alter communication. High or low latency within a communication tool will alter the frequency and duration of utterances (O'Conaill et al., 1993). To verify if this is true with the computer-mediated data that was collected for this study, the utterances are compared with those of the non-mediated dialogues.

8.3.1 Comparison of Dialogue Act Utterance Count and Word Count

The problem with comparing the utterance count and the word count of dialogues is that the quantity of the data is not the same. For this study, I look at the proportion within each dataset as a way of standardizing the differing size of data and look specifically at the differences between Information Giver and Information Follower.

8.3.2 Visualisation of the Dialogue Patterns

To understand the differing patterns of each datasets, the dialogue act word count of the two datasets are visualized. Due to the differing size of each sample, a density plot is used for the visualisation.

■ Reason for Using the Density Plot and Not a Histogram.

The reason to visualise the data is to identify patterns. The pattern I want to identify in this section is the different usage of long and short utterances, *irrelevant to the sample size*. By using a density plot, the y axis is no longer the *count* of occurrences, but the *possibility* of occurrences within the sample. The problem with the histogram of Figure 8.1, is that due to the sample size being different

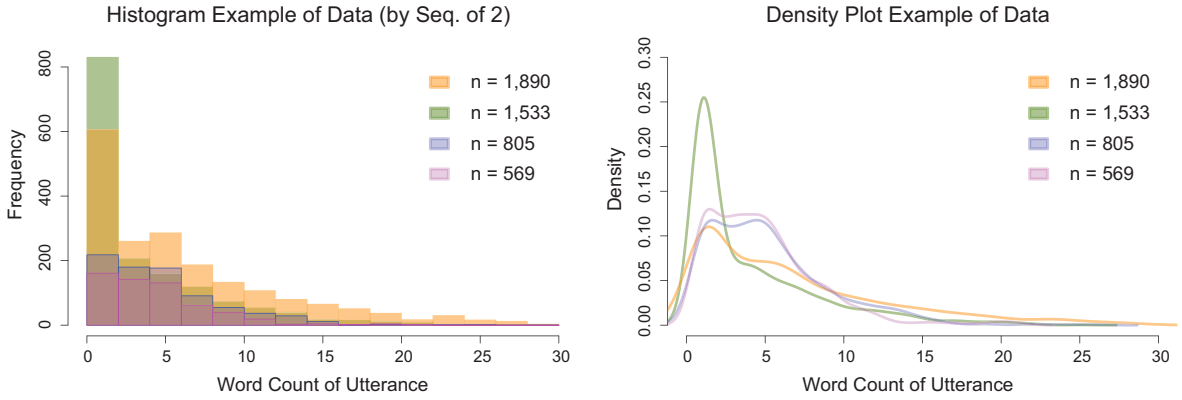


Figure 8.1: *Histogram & density plot of word count per utterance by dataset & role.*

for each item, the resulting figure does not provide a proportionate representation of each item within the data,[1] while the different sample size is standardised in the density plot and provides a figure where the distribution pattern of each item is immediately comparable with each other.

Methodological Note

The objective of the visualisation in this section is to see the patterns of the samples. It is not to see the frequency difference, in other words, which sample contains more utterances. One effect of the density plot is that the quantitive difference between each sample is eliminated, but this characteristic is part of the objective in this specific verification.

[1] Mental arithmetic of rescaling each sample is still required.

8.4 Results

Of the 128 dialogues of the non-mediated dialogue data (the HCRC Edinburgh Map Task corpus), the sixteen dialogues that use the same maps as those used in the computer-mediated dialogue data (the ILMT-s2s corpus) were selected for use in this comparison.

8.4.1 Comparison of Dialogue Act Utterance Count and Word Count

To keep the comparison in a single language, of the thirty subjects in the computer-mediated dialogues, only the data from the fifteen English speaking subjects were used. Table 8.1 summarises the data of the two datasets before they were standardised[2] and Table 8.2 summarises the data after the dialogue acts have been standardised and the Portuguese speaking subjects have been deleted from the data.

Corpus	Dialogue Count	Subject Count	*On-Talk* Dialog Act Count (IG : IF [% Diff.])		*On-Talk* Word Count (IG : IF [% Diff.])	
HCRC	16	32	3,790	(2,203 : 1,587 [−28%])	21,119	(14,735 : 6,384 [−57%])
ILMT-s2s	15	30	2,603	(1,436 : 1,167 [−19%])	14,418	(7,969 : 6,449 [−19%])

Table 8.1: *Summary of the whole of the ILMT-s2s corpus and data from Map 01 and Map 07 of the HCRC Edinburgh Map Task corpus (Off-Talk dialogue acts of the ILMT-s2s corpus have not been included in this summary)*

Corpus	Dialogue Count	Subject Count	*On-Talk* Dialog Act Count (IG : IF [% Diff.])		*On-Talk* Word Count (IG : IF [% Diff.])	
HCRC	16	32	3,468	(1,932 : 1,536 [−20%])	20,644	(14,325 : 6,319 [−56%])
ILMT-s2s	15	15	1,376	(807 : 569 [−29%])	7,410	(4,654 : 2,786 [−41%])

Table 8.2: *Summary after standardising the dialogue acts of both datasets and after the Portuguese speaking subjects were removed from the ILMT-s2s corpus*

In Table 8.2, there are ≈2.5 times the number of dialogue acts in the non-mediated dialogue data (3.468) than the computer-mediated data (1,376). Also the total word count shows that there are ≈2.8 times the number of words in the non-mediated dialogue data (20,644) than the computer-mediated data (7,410). This indicates that the subjects of the non-mediated dialogues are using more words (≈10%) in their dialogue acts than the subjects of the computer mediated map task. Table 8.3 (and to some degree Figure 8.2), illustrates in detail how the non-mediated dialogue data contains longer dialogue acts (more words per dialogue act) with the higher mean and third quarter figures.

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
HCRC Word Count	1	1	4	5.95	8	56	6.69	3,468
HCRC IG Word Count	1	1.75	5	7.42	10	56	7.73	1,932
HCRC IF Word Count	1	1	2	4.11	6	39	4.45	1,536
ILMT-s2s Word Count	1	2	4	5.39	7	178	7.34	1,376
ILMT-s2s IG Word Count	1	2	5	5.77	7	178	9.10	807
ILMT-s2s IF Word Count	1	2	4	4.84	6	24	3.52	569

Table 8.3: *Summary of dialogue act word count by dataset and role.*

[2]To standardise the scheme with the ILMT-s2s scheme, "*Ready*" and "*Uncodable*" have been removed from the annotation data of the HCRC Edinburgh Map Task corpus and the two "*Acknowledgement*" labels of the ILMT-s2s corpus were combined into one and the label of "*Interjection*" is removed.

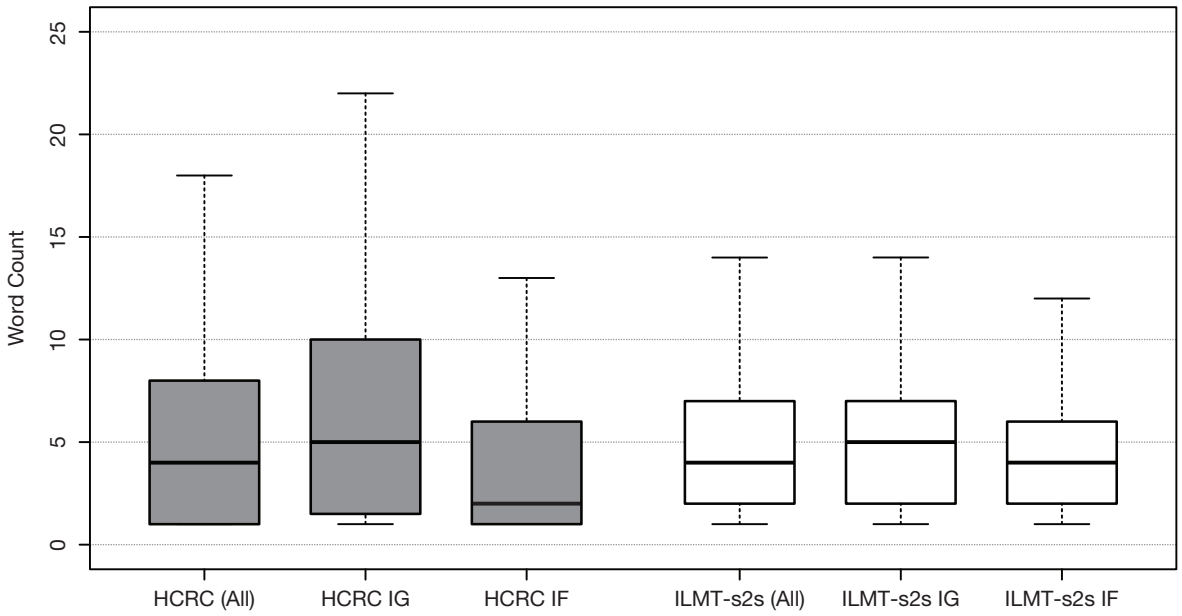


Figure 8.2: *Boxplot of dialogue act word count by dataset and role*

8.4.2 Visualisation of the Dialogue Patterns

■ Distribution of Utterances by Word Length.

Visualising the data by overlaying two density plots on top of each other, as in Figure 8.3, it is clear that the way that the subjects communicate with each other is different. Both the Information Giver

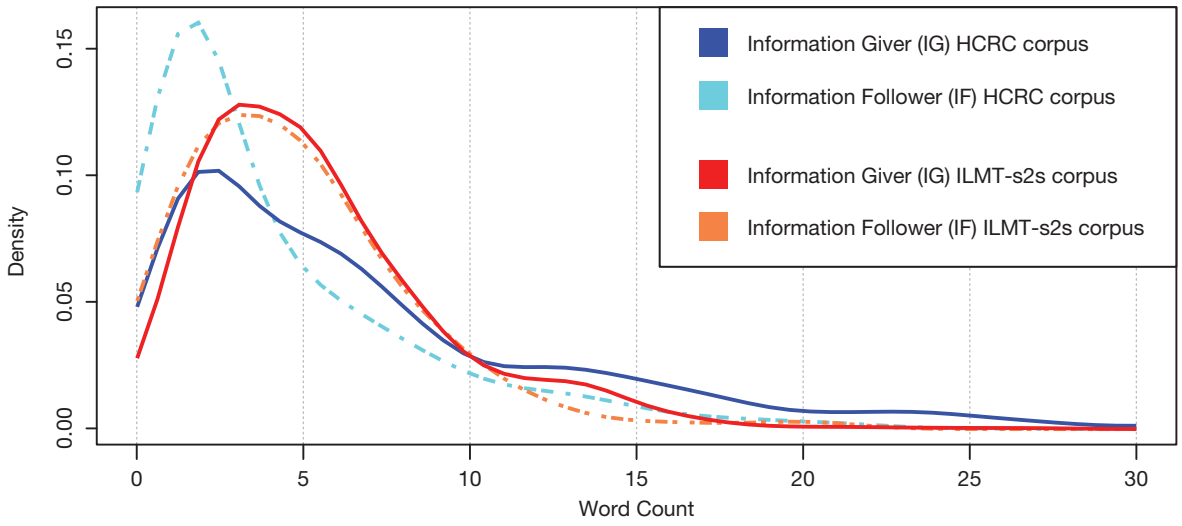


Figure 8.3: *Density plot — Difference of the frequency of x word count utterances per dialogue act per dataset/role.*

and the Information Follower follow a similar trajectory in the computer-mediated dialogues (the ILMT-s2s corpus), but differ in the non-mediated dialogues (HCRC Edinburgh Map Task corpus). The non-mediated dialogue data shows that the Information Follower has a higher proportion of short dialogue acts than the Information Giver with the steep curve placed higher than the low gentle curve of the Information Giver until the word count increases over four.

■ Proportion of Utterances by Dataset.

What is not shown in Figure 8.3 is how the utterances of each role are proportioned within each dataset. To do so, the ratio within each dataset is compared as in Figure 8.4.

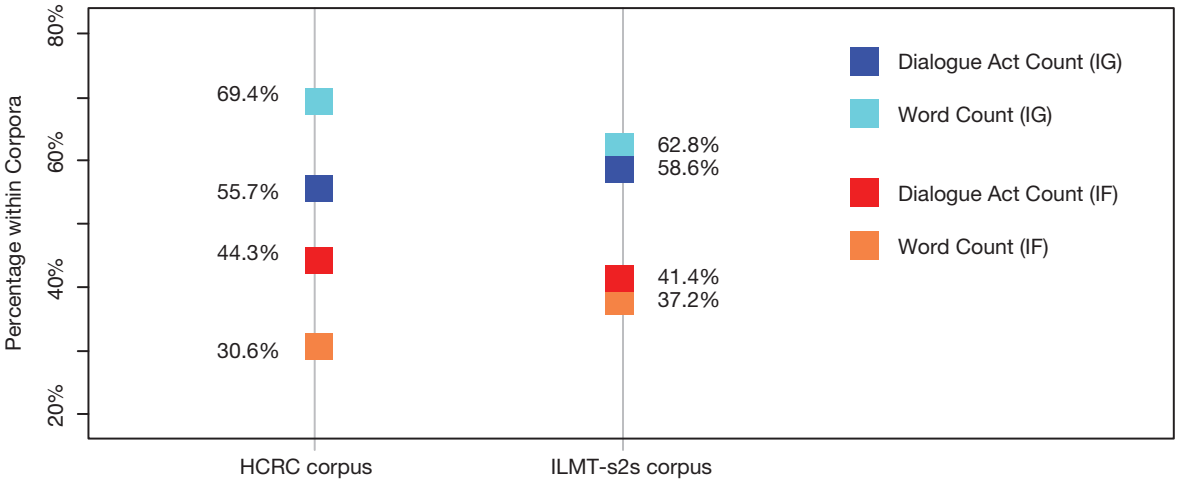


Figure 8.4: *Ratio comparison of dialogue act count and word count of roles between datasets*

■ Conclusion of the Visualisation.

By visualising the data, it is possible to identify the following.

- The distribution curve between roles *differs* in the computer-mediated dialogues.
- The proportion of utterances and word count *differ* between roles in the non-mediated dialogues:
  $IG \mid DA\ Count : Word\ Count = 55.7\% : 69.4\%$ (Diff.= 13.7%)
  $IF \mid DA\ Count : Word\ Count = 44.3\% : 30.6\%$ (Diff.= 13.7%)
- The distribution curve between roles is *similar* within the computer-mediated dialogues.
- The proportion of utterances and word count is *similar* between roles in the computer-mediated dialogues:
  $IG \mid DA\ Count : Word\ Count = 58.6\% : 62.8\%$ (Diff.= 4.2%)
  $IF \mid DA\ Count : Word\ Count = 41.4\% : 37.2\%$ (Diff.= 4.2%)

These values from the non-mediated dialogue dataset indicate that both subjects (Information Giver and Information Follower) are providing a somewhat equal (55.7% versus 45.3%) contribution by means of dialogue acts, to the communication of mutual understanding, but as can be seen from the median word count of 2 words in Table 8.3, the Information Follower is not required to utter as much detailed information to the Information Giver. The same cannot be said about the the subjects of the computer-mediated dialogue dataset. Of the 1,376 counted dialogue acts in the computer-mediated dialogue dataset, 807 were uttered by the Information Giver (58.6%) and 569 were uttered by the Information Follower (41.4%). What is different from the non-mediated dialogue dataset is the word count of these dialogues acts, the proportion between the roles is very similar. Of the 7,410 words

uttered in these dialogue acts, 4,654 words are uttered by the Information Giver (62.8%), while 2,786 words are uttered by the Information Follower (37.2%). As illustrated in Figure 8.4 the percentage of dialogue acts and word count are very similar for the computer-mediated dialogue dataset, which indicates that though the Information Giver contributed more to the communication, the Information Follower was required to utter similar quantities of information when the Information Follower deemed it necessary to communicate.

8.5 Ad-hoc Analysis

From the comparison in the previous section (§ 8.4), it was possible to identify that the subjects of the computer-mediated dialogue dataset (ILMT-s2s corpus) spoke differently to the subjects in the non-mediated dialogue dataset (HCRC Edinburgh Map Task corpus). Subjects with the role of Information Giver had more dialogue acts but used fewer words within the dialogue acts, whereas the Information Follower used less dialogue acts, but more words were used within the fewer dialogue acts. In this section, the comparison is made between the various dialogue acts of each dataset to see if the communication style changes or not.

■ Comparison of Dialogue Act Distribution.

From the bar chart in Figure 8.5 it is possible to see that in the non-mediated dialogue data there is a level of cooperation that shares the effort and improves the efficiency of the communication in the equal frequency of the "*Acknowledgement*" dialogue acts from the Information Follower and the "*Instruct*" dialogue acts from the Information Giver representing 15.40% and 15.46% respectively of all the dialogue acts within the selected data (for the detailed values refer to Table C.1 of Appendix C).

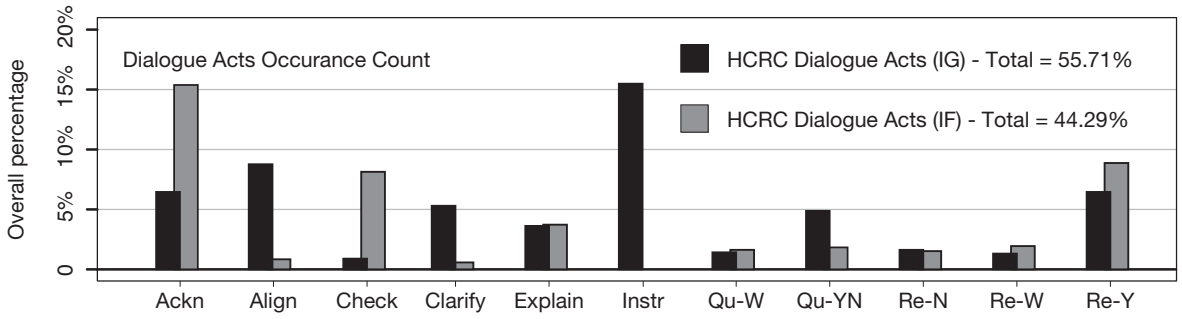


Figure 8.5: *Bar chart role ratio comparison of dialogue acts within the HCRC Edinburgh Map Task corpus*

This balance cannot be observed in the *On-Talk* of the computer-mediated dialogue data as illustrated in Figure 8.6. Contrary to the data from the non-mediated dialogue dataset, an equivalent to the "*Instruct*" dialogue acts from the Information Giver, which represents 22.09% of all the *On-Talk* dialogue acts, is not provided from the Information Follower in the computer-mediated dialogue data (for the detailed values refer to Table C.2 of Appendix C). Even if the word count for each dialogue

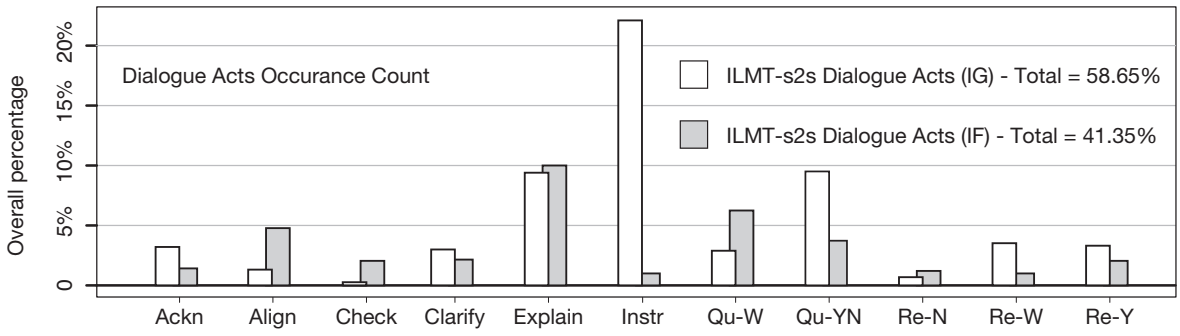


Figure 8.6: *Bar chart role ratio comparison of dialogue acts within the ILMT-s2s corpus*

act is compared, obviously since an acknowledging utterance would be shorter than an instruction, there is no sign that the Information Follower is contributing equally to the communication.

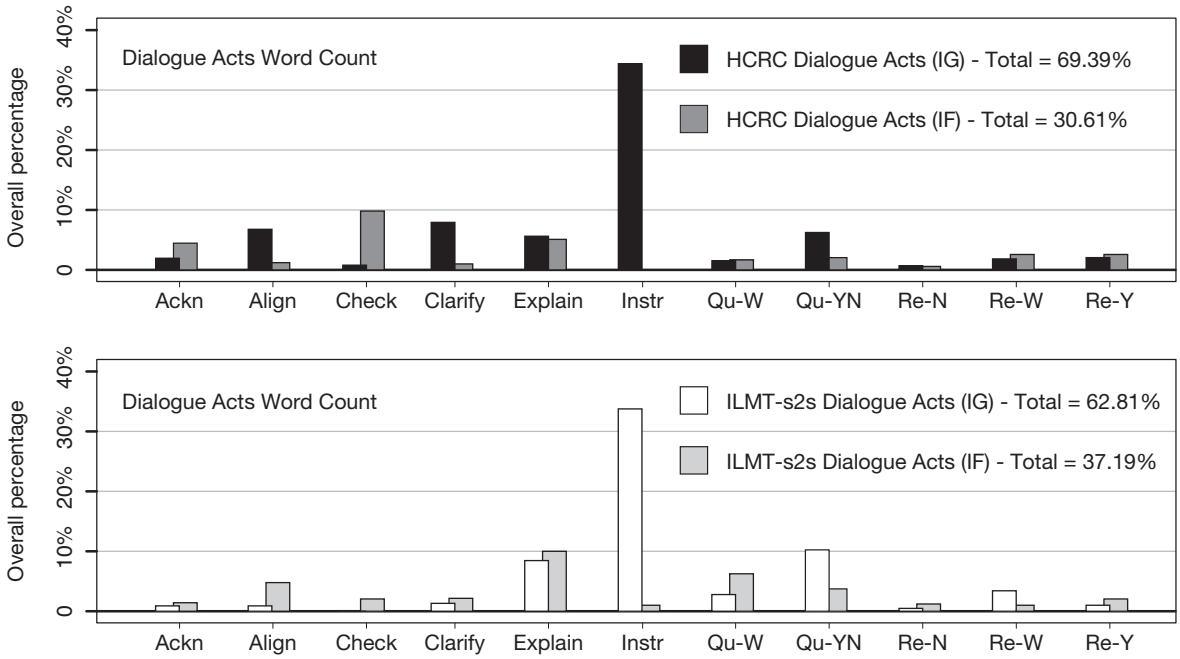


Figure 8.7: *Bar chart role ratio comparison of dialogue acts word count within datasets*

Overall, the dialogue acts representing cooperation (e.g., "*Acknowledge*", "*Align*", "*Check*" and "*Reply Y*") decrease in the computer-mediated dialogue data and a more Gricean (Davies, 2000, 2006) interaction is represented during *On-Talk* as illustrated in Figure 8.8.

8.6 Summary

Before going into details about the results of this study, it must first be noted that the dialogue act annotation of the non-mediated dialogue dataset (HCRC Edinburgh Map Task corpus) and the computer-mediated dialogue dataset (ILMT-s2s corpus) were not done by the same people and unfortunately an inter-coder agreement can not be verified since the annotators of the computer-mediated dialogue dataset did not annotate any of the non-mediated dialogue dataset files to make an inter-coder comparison possible. Therefore there remains the question about the agreement level of the

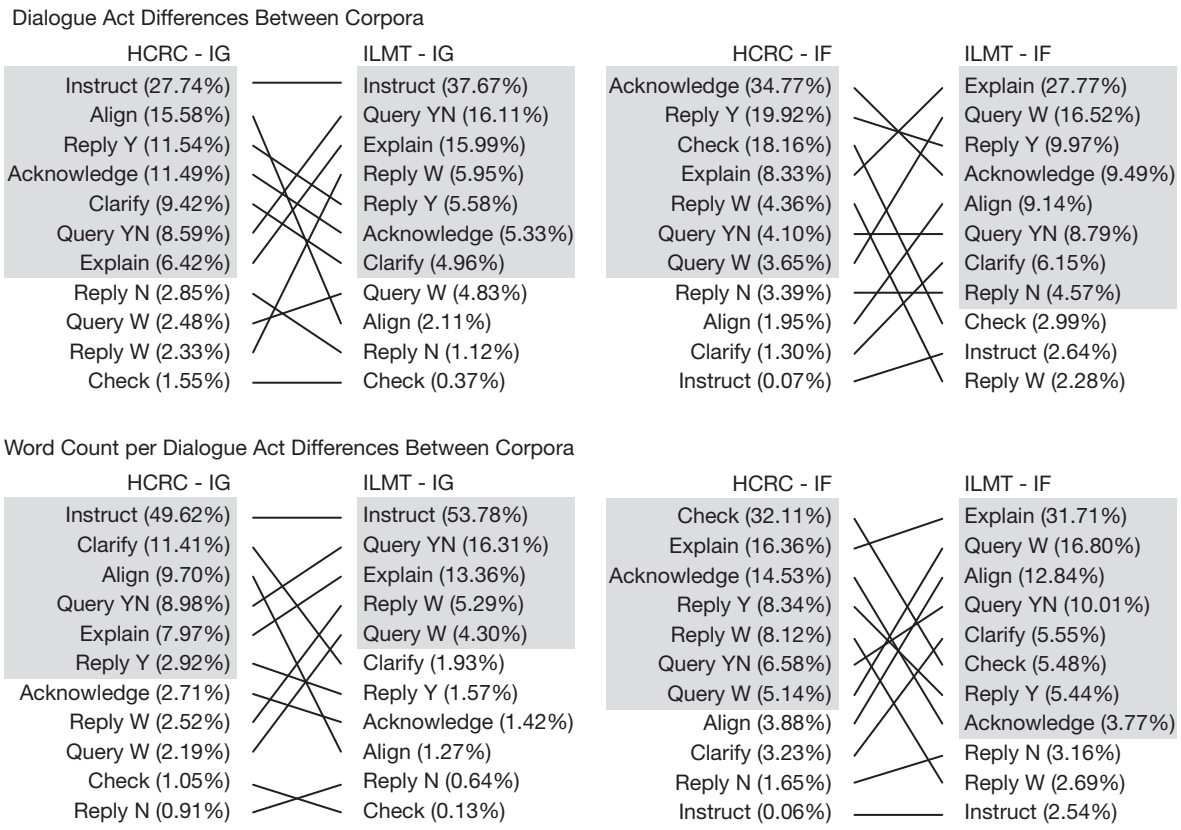


Figure 8.8: *Frequency of occurrence differences between the datasets: Grey highlighted section indicate 90% coverage.*

dialogue act annotations between the two datasets. However I am optimistic that the results presented in this thesis show an accurate illustration of computer mediated multi-language communication.

■ Key Findings.

□ Comparison of Dialogue Act Utterance Count and Word Count.

- Non-mediated dialogue dataset (HCRC Edinburgh Map Task corpus):

  The Information Giver and the Information Follower use approximately the same amount of dialogue acts between each other indicating a greater cooperation between subjects.

  This is not evident from the comparison of the word count, where the Information Giver uses ≈40% more words with the Information Giver uttering 69.39% and the Information Follower uttering 30.61% of all words in the dialogues ("S-Ratio" in Table C.1 of Appendix C) or from the boxplots in Figure 8.9 due to the different frequency and dialogue act occurring.

- Computer-mediated dialogue dataset (ILMT-s2s corpus):

  The Information Giver uses approximately 20% more dialogue acts than the Information Follower, but both use approximately the same amount of words in the uttered dialogue acts. The Information Giver uttering a median (Mdn) of 5 and a mean (M) of 5.77 words per dialogue act and the Information Follower uttering 4 and 5.39 respectively. This similarity can also be identified in the boxplot of Figure 8.10.

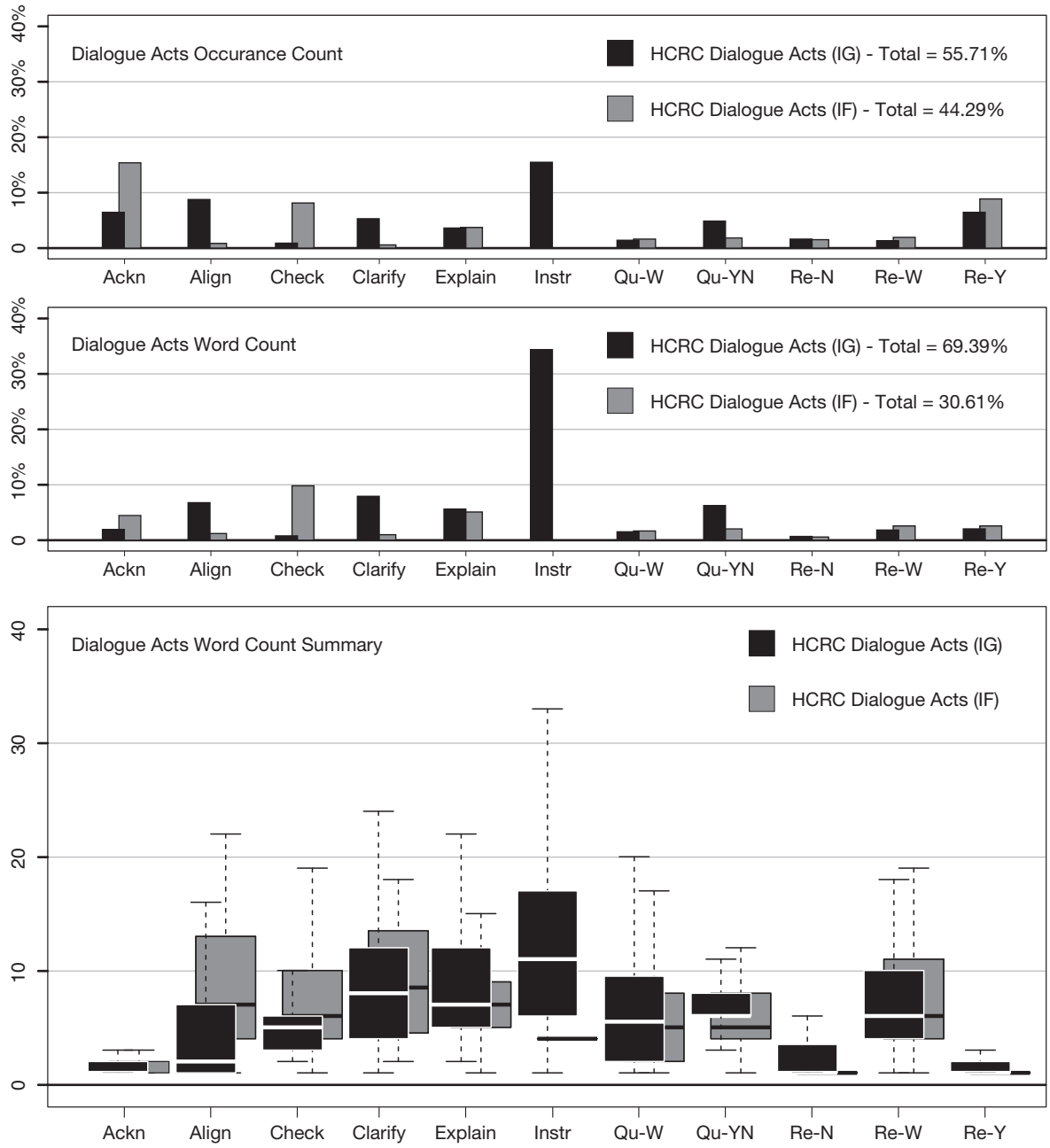


Figure 8.9: *Bar chart role ratio comparison of dialogue acts within HCRC Edinburgh Map Task corpus (top and middle) and boxplot comparison of said dataset dialogue act word count (bottom).*

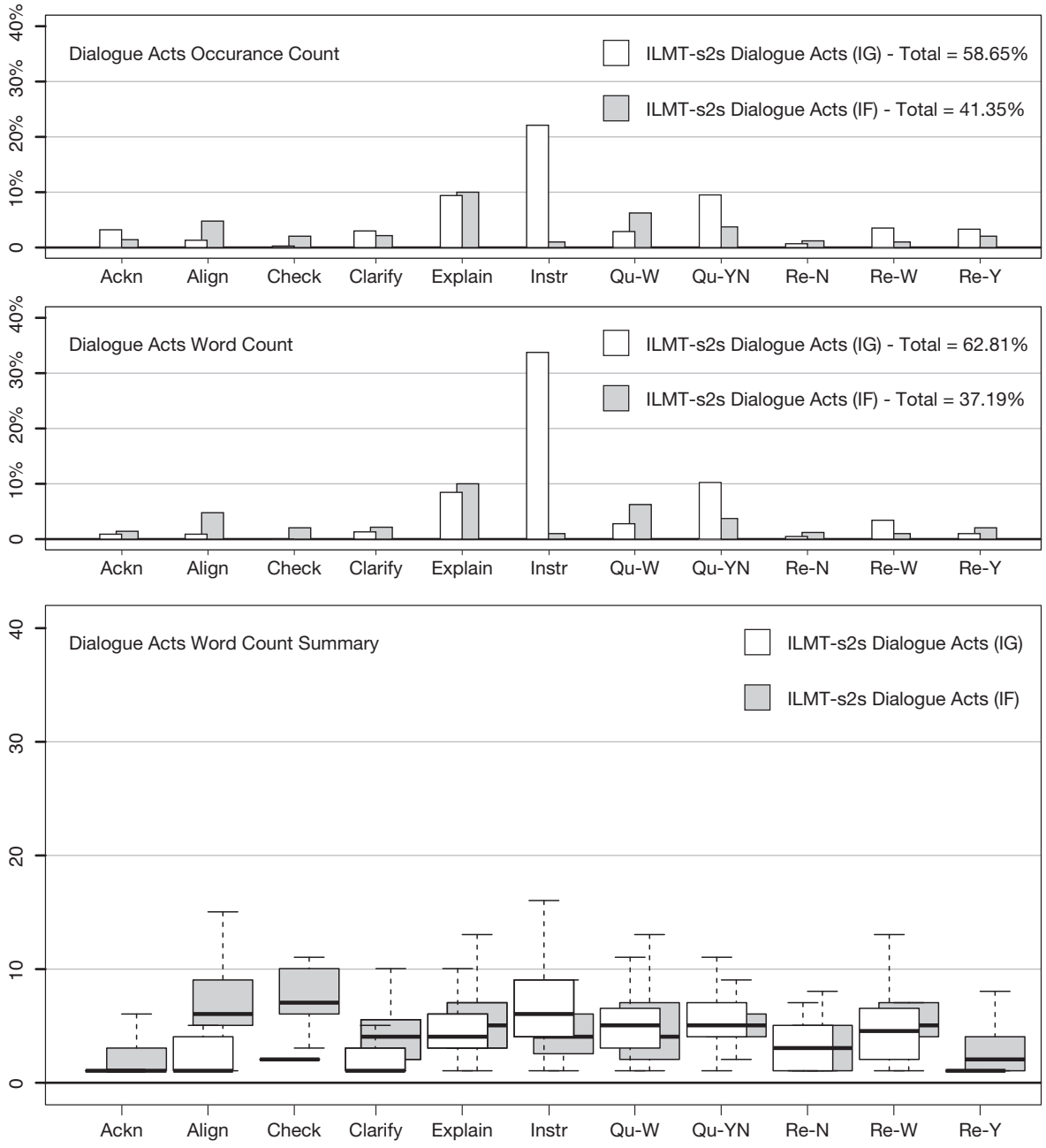


Figure 8.10: *Bar chart role ratio comparison of dialogue acts within ILMT-s2s corpus (top and middle) and boxplot comparison of said dataset dialogue act word count (bottom).*

☐ Comparison of Dialogue Act Distribution.

- Non-mediated dialogue dataset (HCRC Edinburgh Map Task corpus):

  The Information Giver and the Information Follower use approximately the same amount of dialogue acts between each other, but from the comparison of the word count, the Information Giver uses ≈40% more words.

- Computer-mediated dialogue dataset (ILMT-s2s corpus):

  The Information Giver uses ≈20% more dialogue acts than the Information Follower, but both use approximately the same amount of words in the uttered dialogue acts.

  The use of "Acknowledgements" saw a reduction in Information Follower *On-Talk* utterances. However, as stated in Figure 4.28 (§ 4.4.3: "Annotation"), utterances that were labeled as "Acknowledgements" were used by the subjects of the computer-mediated dialogue dataset as *Off-Talk*.[3]

Due to the characteristics of the computer-mediated dialogue dataset, where (i) the subject speaks to a computer system (the ILMT-s2s System), (ii) the system converts the speech to text, (iii) translates the text to the target language, and (iv) the system sends the translation to the interlocutor's computer for synthesised output, inter-cooperation between the subject and the interlocutor — that results in spreading evenly the 'cost' of communication (Newlands et al., 2003) — is no longer as efficient as would be the case in Face-to-Face communication. This can be seen in the ≈25% decrease of "*Acknowledgements*", ≈10% decrease of "*Reply Y*"[4] and ≈15% decrease of "*Check*" counter with the ≈20% increase in "*Explain*" and ≈13% increase in "*Query W*"[5] for the Information Follower subjects of the computer-mediated dialogue data.

The decrease in word count per dialogue act for the Information Giver, the increase in word count per dialogue act for the Information Follower, decrease in short backchannel dialogue acts such as "*Acknowledgements*" and "*Reply Y*" indicate to an increase in the cognitive cost of communicating (*On-Talk*) with the interlocutor. This phenomenon can be seen as a more Gricean (Davies, 2000, 2006) interaction, where, with the increase in the cognitive cost of the communication, items that facilitate but do not provide details to the information required have been removed. This outcome is similar to the communication style reported for slow video conferencing systems. As mentioned in § 2.2.2: "Dialogue Acts and Communication Medium", in a study of video conference dialogues, O'Conaill et al. (1993, p. 420) reported the users adopted a different communication style to Face-to-Face communication which is similar to the observations found with the users in the computer-mediated dialogue dataset.

1. **High Latency Video Conferencing:**

[3] Out of the 519 English *Off-Talk* dialogue acts, 324 (≈60%) were labeled as "Acknowledgements". 187 for the Information Giver and 137 for the Information Follower.

[4] *Reply Y*: Any reply to any query with a yes-no surface form which means "yes", however that is expressed.

[5] *Query W*: Any query which is not covered by the other categories — not a "Yes" or "No" query.

Listeners produced fewer backchannels and interrupted less often.

Computer-mediated dialogue dataset (ILMT-s2s corpus):

The Information Follower used fewer backchannels (with the reduction of "*Acknowledgements*" in *On-Talk*) and the system does not facilitate interruptions due to the procedural flow of "subject utterance — system mediation — output to interlocutor" enforced by the computer system.

2. **High Latency Video Conferencing:**

   Listeners were also less likely to anticipate turn endings.

   Computer-mediated dialogue dataset (ILMT-s2s corpus):

   The Information Follower is also less likely to anticipate turn endings due to the short mean word count of each Information Giver utterance (turn) and from the clear indication in the computer system's UI (§ 4.2.5: "The User Interface").

3. **High Latency Video Conferencing:**

   The result of listeners reducing interruptions and speaker feedback, combined with the general difficulty of switching speakers, was a formal lecture style of interaction, with long turns, handed over by a very deliberate process.

   Computer-mediated dialogue dataset (ILMT-s2s corpus):

   From the procedural flow that the mediating computer system provides, interruptions were repressed, the cognitive cost of producing feedback was increased and switching speakers was systematic. From this structural limitation, the Information Follower increased the mean word count of their uttarances when compared to the non-mediated dialogue dataset interactions.

8.7 Discussion

The subjects of the computer-mediated dialogue data did utter a lot of "*Acknowledgements*" dialogue acts as indicated in Figure 8.11. The reason why it seems as if there is a large reduction in the dialogue act of "*Acknowledgements*" is because they are uttered as *Off-Talk* and not *On-Talk*.

The transmission of communication items is different in the two datasets. In the direct H2H setting of the non-mediated dialogue data, every audible sound that comes out of the subject's mouth is transmitted to the interlocutor. However with the interlingual computer mediated setting that used an activation button to communicate, only utterances that the subject intends to communicate to the interlocutor are transmitted. Due to this difference, standardising the data of the two datasets to only analysing the utterances in the computer-mediated dialogue data that are transmitted to the interlocutor (*On-Talk*) is reasonable because the objective of the dialogue is task based, and the understanding of the interlocutor is critical to the completion of the task.

The analysis in this chapter is from the comparison of the two different datasets, and therefore it is not a full study of the characteristics of the subject's stand-alone adaptation with the ILMT-s2s

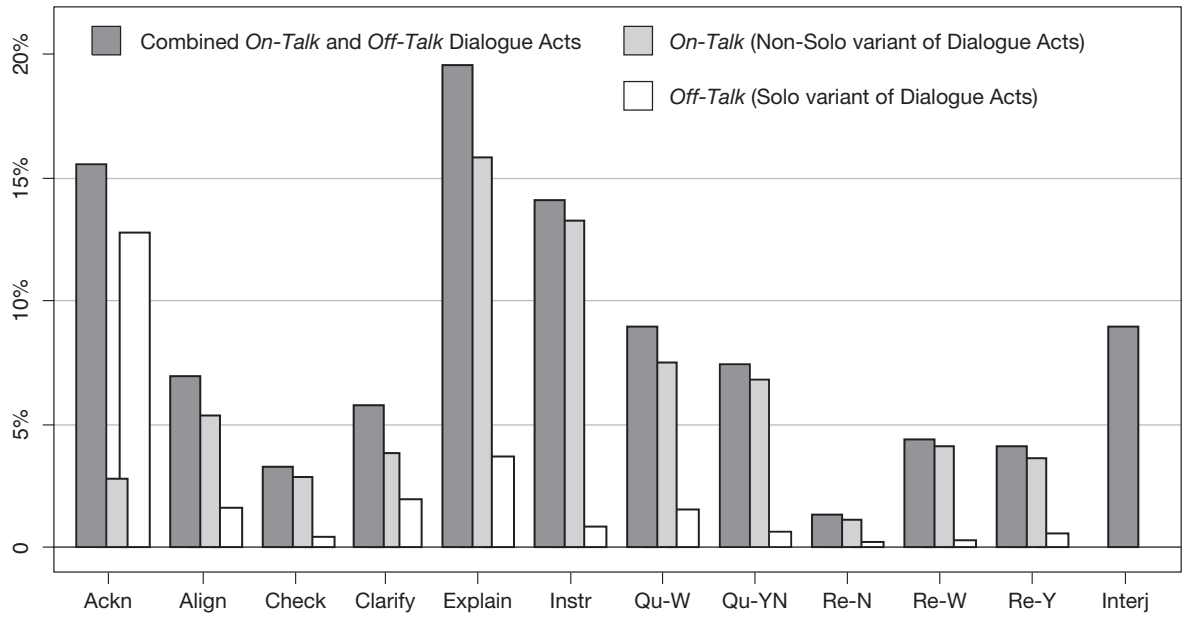


Figure 8.11: *Bar chart of dialogue acts: Total of On-Talk and Off-Talk variant's combined as well as separated — "Acknowledgement CP" / "Acknowledgement CPU" have been combined into one item*

System (a system using an activation button to transmit communication). There are two point of interest that were not analysed in-time for the completion of this thesis.

(i) To look into, not only *On-Talk* as performed here, but also the *Off-Talk* utterances of the subject and identify how the *Off-Talk* utterances are contributing the the subject's communication style — e.g., "*Acknowledgements*" was uttered by the subject, but most of the time, not to the interlocutor as *On-Talk*.

(ii) To look into the linear pattern of the dialogue acts of the two datasets, by looking at the preceding and following dialogue acts surrounding each "*turn*" — in other words to look into the two elements of the computer mediated communication: First, when the subject is trying to communicate the task to the interlocutor and second, when the subject is trying to adapt their communication to the ILMT-s2s System. This will hopefully provide an insight into what kind of errors and how much information was needed to understand a given instruction, and see if there is any uniformity in the actions of the subjects of the computer-mediated dialogue data.

Another point of interest that needs further investigation from this observation is the opposing usage of the "Align" dialogue act in the computer-mediated dialogue dataset (Figure 8.8). In the non-mediated dialogue dataset the "Align" dialogue act is mainly used by the Information Giver while in the computer-mediated dialogue data, it is mainly used by the Information Follower. Also a recent publication by Koch, Forgas, & Matovic (2013) highlighted how the mood of subjects affected the subjects communication style, with negative moods increasing the subject's tendency to observe Grice's conversational norms. It will be interesting to keep this finding from Koch et al. (2013) in mind, while further analysing the dialogue acts of the computer-mediated dialogue data.

Chapter 9

ASR Performance & Speech Rate

9.1 Introduction

Chapter 6: "The Three Talk Types" reported the findings surrounding the three talk types identified in the computer-mediated dialogue data. The three talk types, which are differentiated by the direction of the dialogue (who the subject is talking to), were identified in the dataset data and are labeled as *On-Talk*,[1] *Off-Talk Self*,[2] and *Off-Talk Other*.[3] Of the three talk types, the mean speech rate of *On-Talk* was the slowest, followed by *Off-Talk Self*, with the fastest being *Off-Talk Other*.

In this section, I study the speech rate of the utterances labeled as *On-Talk*, with relation to the Automatic Speech Recognition (ASR) performance, to see if the WER (ASR performance) is affecting the speech rate of the subject. As reported by Oviatt et al. (1998, p. 419) when a subject is interacting with a speech based computer dialogues system and encounters an error, the subjects would adapt to the situation as described:

> "First, they *increase linguistic contrast* through alteration of input modes and lexical content over repeated correction attempts. Second, when correcting with verbatim speech, they *increase hyperarticulation* by lengthening speech segments and pauses, and increasing the use of final falling contours. Third, when they hyperarticulate, users simultaneously *suppress linguistic variability* in their speech signal's amplitude and fundamental frequency." (Oviatt et al., 1998, p. 419)

From the overview of the data illustrated in Figure 9.2 (page 164), there is a high probability that the ASR performance does affects the subject's speech rate. To verify if this phenomena applies to the data of the computer-mediated dialogue data, the error rate of the S2T/ASR system was calculated and compared with the speech rate of the subjects.

[1] Talking to their interlocutor through the ILMT-s2s System.
[2] Talking to themselves.
[3] Talking Face-to-Face to someone.

9.2 Dataset

In this chapter, the speech rate difference labels described in § 3.3.3.4 and the ASR performance labels described in § 3.3.3.5 of Chapter 3 are used for the analysis.

9.3 Methods

9.3.1 Calculating the ASR Performance (WER & *F*-Score)

The ILMT-s2s System's ASR performance was assessed by aligning the human transcriptions[4] (§ 4.4.1) and the ASR result transcript to compute the Word Error Rate (WER), Precision, Recall, and F-Score of each utterance (McCowan, Moore, et al., 2005).

9.3.2 Visualisation of the Speech Rate and ASR Performance

Using the values of the speech rate (§ 4.4.4: "Calculating the Speech Rate") and ASR performance (§ 4.4.5: "Calculating the ASR Performance (WER & F-Score)"), a graph was created to displays the flow of change.

The emphasis of this study is to look into the speech rate and ASR performance. The difference in the duration of the dialogues is ignored, but the differing quantities of utterances directed towards the interlocutor need to be transformed into a comparable format of twenty frames per dialogue. Each utterance value was duplicated 20 times (or equivalent), and then grouped by the number of utterances there were in the dialogue (or equivalent) as indicated in the seven examples in Table 9.1 so that the total utterance count would be a unified 20.

Dialogue number (En)		D.01	D.02	D.03	D.04	D.05	D.06	D.07
Utterance Count per Dialogue		33	100	106	73	201	93	89
Duplicate each value	×	20	1	10	20	20	20	20
Total number of values		660	100	1,060	1,460	4,020	1,860	1,780
Group values (windows)	÷	33	5	53	73	201	93	89
Number of window samples		20	20	20	20	20	20	20

Table 9.1: *Sample window alteration example*

9.3.3 Speech Rate & ASR Performance Analysis

The analysis uses the values from the 20 frame window explained in § 9.3.2 "Visualisation of the Speech Rate and ASR Performance". The reason for this is to standardise the weight of the values from each subject, since the utterance count for each subject and also the role of the subject (Information Giver or Information Follower) would have an effect on the utterance count.

[4] Punctuation is removed from the transcription so as not to interfere with the alignment, and all letters are changed into small cap.

Using these values, the speech rate and the ASR performance are compared within each grouping (role, gender, language spoken, setting of the recording).

- The speech rate of each *On-Talk* utterance of *[Refer to grouping list below]* are tested to verify if there is a significant difference (*Wilcox*).
- The ASR performance of each *On-Talk* utterance of *[Refer to grouping list below]* are tested to verify if there is a significant difference (*Wilcox*).

Grouping list:

1. Gender (female or male)
2. Role (Information Giver or Information Follower)
3. Setting (w/ Video or w/o Video)
4. Language Spoken (English or Portuguese)

9.4 Results

9.4.1 Calculating the ASR Performance (WER & *F*-Score)

The ASR performance was calculated on the ≈2,300 *On-Talk*[5] utterances of the computer-mediated dialogue dataset using the formula explained in § 4.4.5. The distribution of the ASR performance values (Reverse *F*-Score) is illustrated as a density plot in Figure 9.1 with the summary of the values listed in Table 9.2. The WER values are also listed in Table 9.3 as a reference.

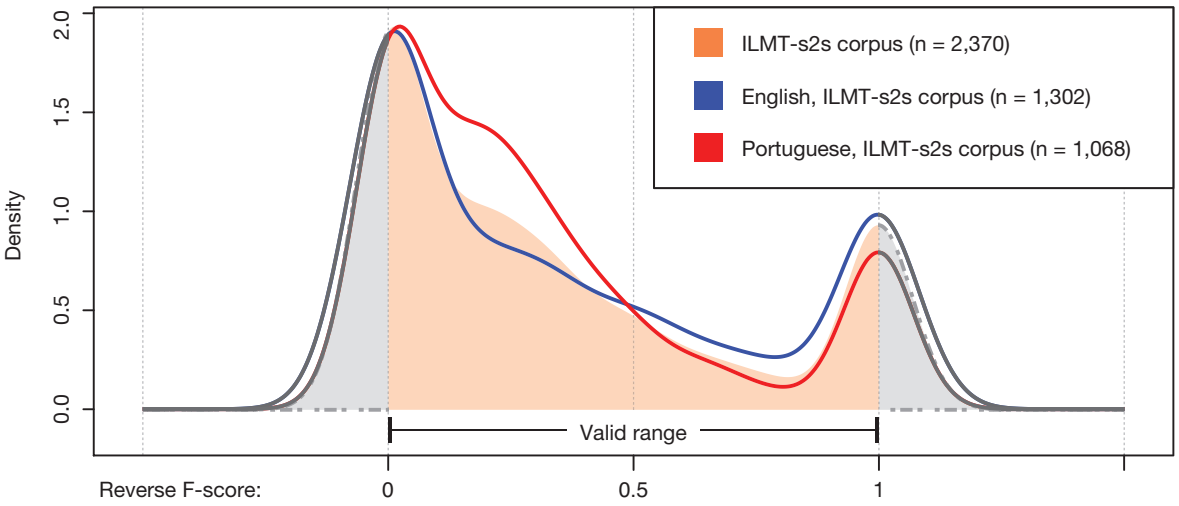


Figure 9.1: *ASR performance density plot of English and Portuguese dialogue data*

ASR Performance	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
ILMT-s2s	0.00	0.00	0.22	0.34	0.56	1.00	0.36	2,370
ILMT-s2s (En)	0.00	0.00	0.25	0.37	0.67	1.00	0.38	1,302
ILMT-s2s (Pt)	0.00	0.00	0.20	0.31	0.43	1.00	0.34	1,068

Table 9.2: *Summary of computer-mediated dialogue F-Score values*

[5] "*On-Talk*" are utterances directed to the subject's interlocutor via the ILMT-s2s System as explained in Chapter 6.

ASR Performance	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Item Count
ILMT-s2s	0.00	0.00	0.25	0.41	0.67	10.00	0.55	2,370
ILMT-s2s (En)	0.00	0.00	0.29	0.43	0.75	10.00	0.60	1,302
ILMT-s2s (Pt)	0.00	0.00	0.25	0.38	0.54	8.00	0.48	1,068

Table 9.3: *Summary of computer-mediated dialogue WER values*

9.4.2 Visualisation of the Speech Rate and ASR Performance

Using the speech rate values that are calculated in § 4.4.4 and the ASR performance values calculated with the method explained in § 4.4.5, the *On-Talk* values of each dialogue are standardised into 20 frame windows and visualised as Figure 9.2.

- The top (orange) curve represents the ASR performance accuracy (Reverse *F*-Score) — 0% if the ASR result is completely correct, and 100% if it's completely wrong.
- The bottom (blue) curve represents the subject speech rate — Positive percentage points if the speech rate is faster than the reference TTS output and negative percentage points if it's slower.

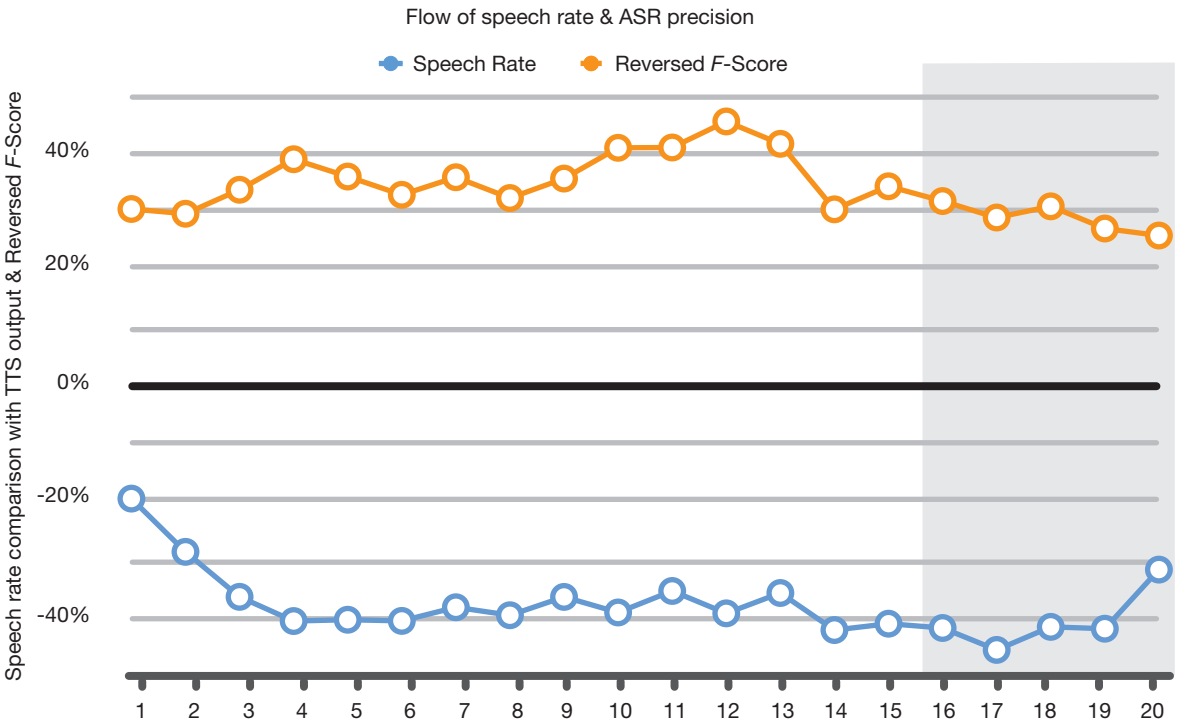


Figure 9.2: *Speech rate related and ASR accuracy for all subjects combined — Dialogue duration standardised into 20 frame windows*

	Speech Rate (%)					Reverse *F*-Score (%)				
	Min.	*Mdn*	*M*	Max.	*SD*	Min.	*Mdn*	*M*	Max.	*SD*
All dialogues	−185.80	−31.31	−37.61	28.59	33.64	0	29.77	34.16	100	23.85

Table 9.4: *Summary of 20 frame window dialogue data*

■ The Individual Dialogue Charts.

For the individual dialogue charts, refer to the following figures in § A.3 of Appendix A.

Figure A.14: Individual dialogue figures — Dialogues 1 to 4.

Figure A.15: Individual dialogue figures — Dialogues 5 to 8.

Figure A.16: Individual dialogue figures — Dialogues 9 to 12.

Figure A.17: Individual dialogue figures — Dialogues 13 to 15.

■ The Dialogue Charts by Grouping.

For the charts of the grouping within each dataset, refer to the following figures in § A.3 of Appendix A.

Figure A.3: Grouping by role, spoken language, video setting and crashes during video usage.

Figure A.4: Grouping of gender divided by role and spoken language.

Figure A.5: Grouping of gender divided by video setting and crashes during video usage.

Figure A.6: Grouping of roles divided by gender and spoken language.

Figure A.7: Grouping of roles divided by video setting and crashes during video usage.

Figure A.8: Grouping of spoken languages divided by role and gender.

Figure A.9: Grouping of spoken languages divided by video setting and crashes during video usage.

Figure A.10: Grouping of video setting divided by role and spoken language.

Figure A.11: Grouping of video setting divided by gender and crashes during video usage.

Figure A.12: Grouping of dialogues with systems crashes divided by role and spoken language.

Figure A.13: Grouping of dialogues with systems crashes divided by gender.

9.4.3 Speech Rate & ASR Performance Analysis

The analysis in this section used the values from the 20 frame window explained in § 4.4.6. The reason for this was to standardise the weight of the values from each subject. As illustrated in Figure 9.3, and summarised in Table 9.5 the *On-Talk* utterance count for each subject varied greatly and the difference of the subject's role as an Information Giver or Information Follower also created a difference with the Information Giver uttering, on average, 11 % more utterances than the Information Follower.

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*	Count
Utterance count — Both IG & IF	76	129	171	162.6	192.5	262	51.35	2,439
Utterance count — IG	34	64	94	90.4	100.5	201	39.27	1,356
Utterance count — IF	33	54.5	73	72.2	90.5	117	24.88	1,083

Table 9.5: *ILMT-s2s corpus On-Talk utterance count summary*

For this reason, I did not want the data patterns from certain subjects, or roles, that had more utterances to overwhelm the patterns of the subjects, or roles that had fewer utterances.

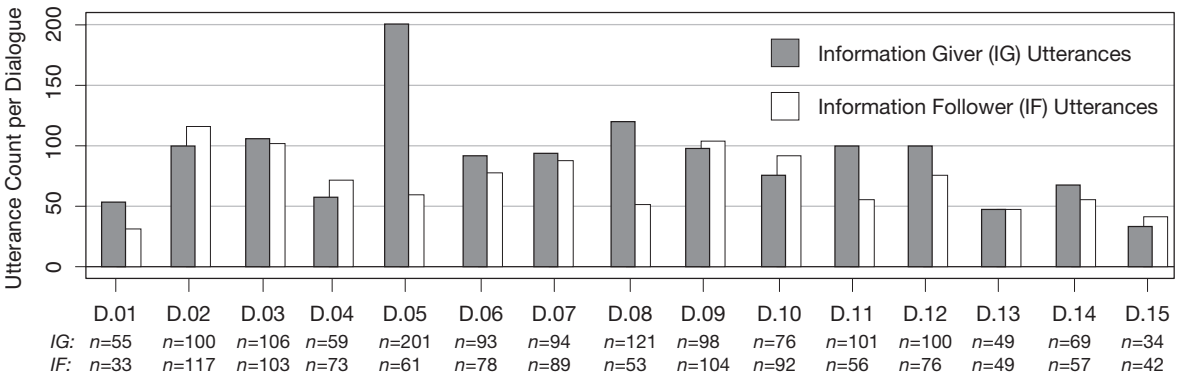


Figure 9.3: *ILMT-s2s corpus On-Talk utterance count per dialogue*

■ Summary of the Data by Grouping.

The computer-mediated dialogue data is summarised by; (i) the gender of the subject, (ii) the role of the subject, (iii) the language spoken by the subject, and (iv) the setting the subject was in.

As listed in Table 9.6 and highlighted in the "Diff." row, a large mean (M) difference between the speech rates of *On-Talk* is identified in (i) the gender of the subject, and (ii) the role of the subject, but the difference within groups for the ASR performance are small.

	Speech Rate Difference (%)				Reverse F-Score			
Category	M	SD	M	SD	M	SD	M	SD
Gender (♀ : ♂)	−47.89	34.81	−25.86	27.99	34.32	24.31	33.97	23.35
Diff.	M: 22.03		SD: 6.81		M: 0.35		SD: 0.96	
Role (IG : IF)	−47.94	33.07	−27.27	30.96	32.81	21.62	35.51	25.86
Diff.	M: 20.67		SD: 2.11		M: 2.70		SD: 4.24	
Language (Pt : En)	−41.35	33.50	−33.87	33.40	33.72	21.93	34.60	25.66
Diff.	M: 7.48		SD: 0.10		M: 0.88		SD: 3.73	
Setting (w/o Vid : w/ Vid)	−38.03	35.11	−37.24	32.34	29.18	23.58	38.52	23.26
Diff.	M: 0.79		SD: 2.77		M: 9.34		SD: 0.31	

Table 9.6: *Speech Rate Adaptation to the ILMT-s2s System by Group (20 frame window)*

■ Significance Test of the Speech Rate by Grouping.

Using the data summarised in Table 9.6, a significance test was performed within the groupings. For example, the male subjects and the female subjects of the computer-mediated dialogue data are tested. The results of these tests are listed below in Table 9.7.

Speech Rate Comparison						
Grouping	Mdn	M	SD	wilcox.test	sig.	cliff.delta	ES
Role (IG : IF)	−41 : −22	−48 : −27	33 : 31	$p < 2.2e-16$	***	$\delta = 0.412$	med.
Gender (♀ : ♂)	−40 : −23	−48 : −26	35 : 28	$p < 2.2e-16$	***	$\delta = 0.391$	med.
Lang. (Pt : En)	−35 : −28	−41 : −34	34 : 33	$p = 0.0020$	**	$\delta = 0.146$	neg.
Setting (w/o Vid : w/ Vid)	−31 : −31	−38 : −37	35 : 32	$p = 0.9797$	–	$\delta = 0.001$	neg.

Table 9.7: *Speech rate significance test and effect size between between groupings (20 frame window)*

The results indicated that for the speech rate, (1) gender, (2) role, and (3) spoken languages have a significant difference, but only the differences between (1) gender, and (2) role have a meaningful effect size of medium, as reported in Table 9.7.

■ Significance Test of the ASR Performance (Reverse F-Score) by Grouping.

Using the data summarised in Table 9.6, a significance test was performed within the groupings. For example, the male subjects and the female subjects of the computer-mediated dialogue data are tested. The results of these tests are listed below in Table 9.8.

ASR Performance (Reverse F-Score) Comparison							
Grouping	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	ES
Setting (w/o Vid : w/ Vid)	22 : 36	29 : 39	24 : 23	$p = 6.603e-08$	***	$\delta = 0.255$	small
Role (IG : IF)	28 : 33	33 : 36	22 : 26	$p = 0.4103$	–	$\delta = 0.039$	neg.
Lang. (Pt : En)	30 : 30	34 : 35	22 : 26	$p = 0.9296$	–	$\delta = 0.004$	neg.
Gender (♀ : ♂)	30 : 29	34 : 34	24 : 23	$p = 0.9751$	–	$\delta = 0.001$	neg.

Table 9.8: *ASR performance significance test and effect size between between groupings (20 frame window)*

For the ASR performance, a significant difference is only observed in the setting difference of w/o Video and w/ Video, but with only a small effect size difference, as reported in Table 9.8.

■ Visual Verification of Gender Differences.

The difference in the subject gender indicates that the female subjects are speaking at a slower mean speech rate of −47.89% than their male counterparts at −25.86% which is a difference of ≈22, but the ASR performance is basically the same, with only a mean ASR performance difference of ≈0.35, as indicated in Table 9.6. This difference in speech rate and similarity in the ASR performance can also be seen from Figure 9.4, which plots the mean of all subjects divided by gender. Gender values

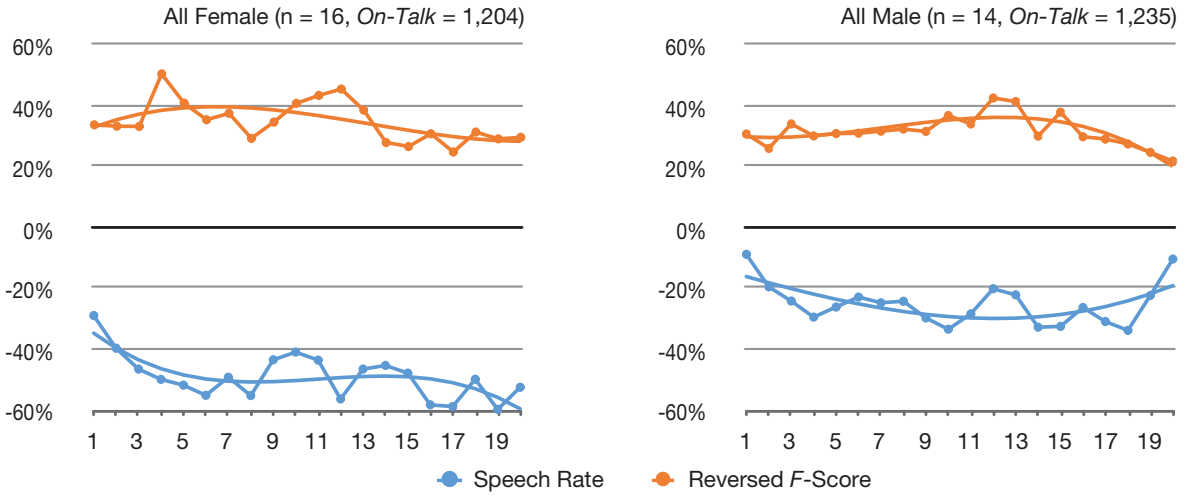


Figure 9.4: *Plot of subject's speech rate and Reverse F-Score by gender (20 frame window)*

further divided by role, spoken language, and setting can be found in Figure A.4 and Figure A.5 of Appendix A with the female speech rate being consistently lower than male values, but with ASR performance remaining similar.

■ Visual Verification of Role Differences.

As with the difference between gender, similar differences are found for the two roles of the subject. The mean speech rate of the Information Giver is lower at −47.94% than that of the Information Follower at −25.86% which is a difference of ≈20, without much difference in ASR performance with a difference of ≈3 as illustrated in Figure 9.5 and listed in Table 9.6. Role values further divided

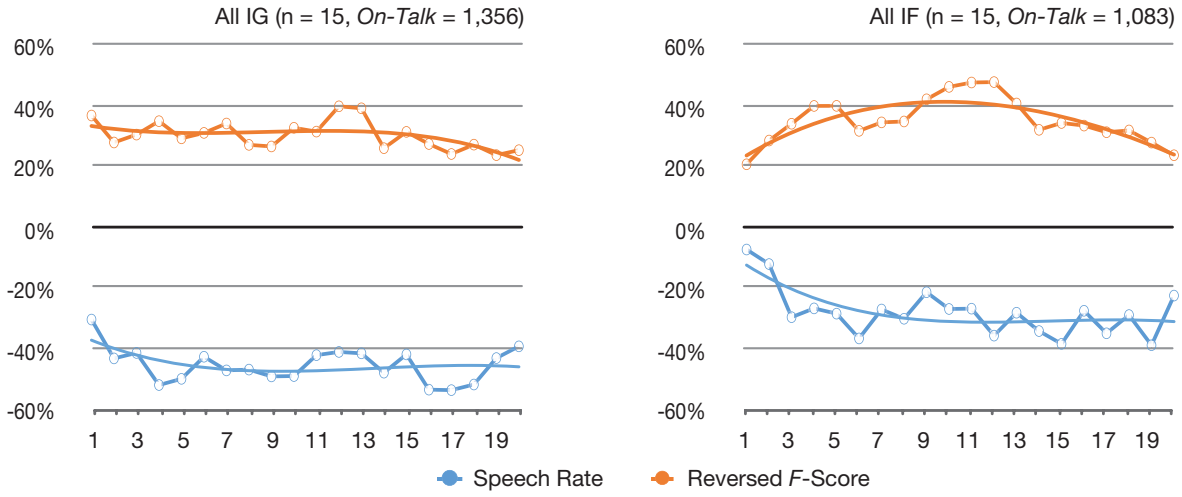


Figure 9.5: *Plot of subject's speech rate and Reverse F-Score by role (20 frame window)*

by gender, spoken language, and setting can be found in Figure A.6 and Figure A.7 of Appendix A with the Information Giver being consistently lower than the speech rate values of the Information Follower, but with ASR performance being fairly similar.

■ Visual Verification of Setting Differences.

Different from gender and role, the mean speech rate between the two settings of w/o Video and w/ Video are similar, but the ASR performance shows the biggest difference at ≈9.3. The mean speech rate of the two settings, w/ Video and w/o Video, are both close to −37.5%, while the ASR performance differs by ≈9.3 points with the setting of w/o Video displaying a better ASR performance of 29.2 while w/ Video is only 38.5, as listed in Table 9.6 and illustrated in Figure 9.6. Setting values further divided by role, gender, and spoken language can be found in Figure A.10 and Figure A.11 of Appendix A with the w/o Video setting consistently being lower than the ASR performance of the w/ Video, but with speech rate being somewhat similar.

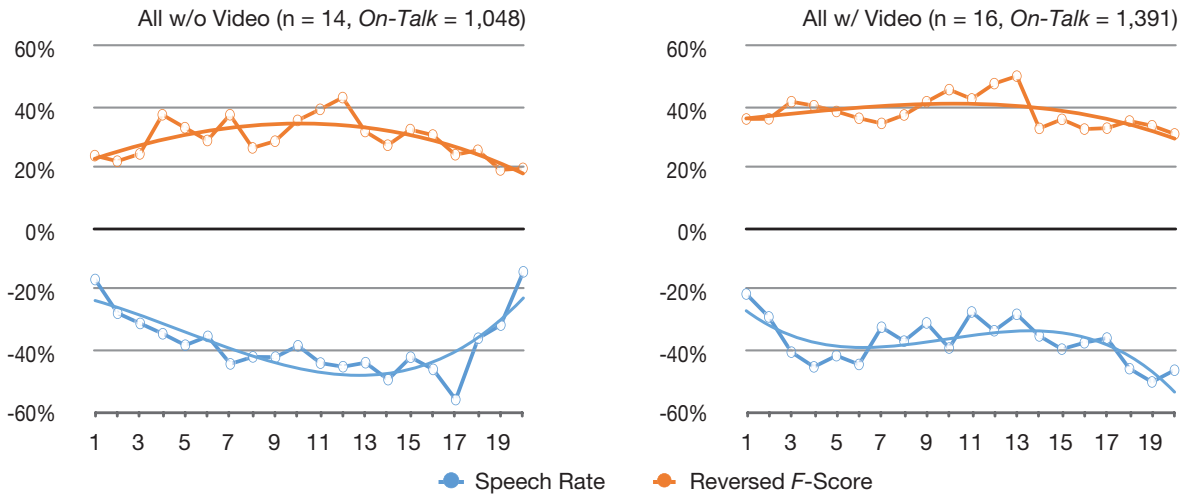


Figure 9.6: *Plot of subject's speech rate and Reverse F-Score by setting (20 frame window)*

■ Visual Verification: Summary of Differences.

A summary of the various groupings of the speech rate is illustrated in Figure 9.7, and a summary of the ASR performance in Figure 9.8. The figures are divided into two, with the combined values indicated on the left side of the figure and a breakdown of the groupings available on the right.

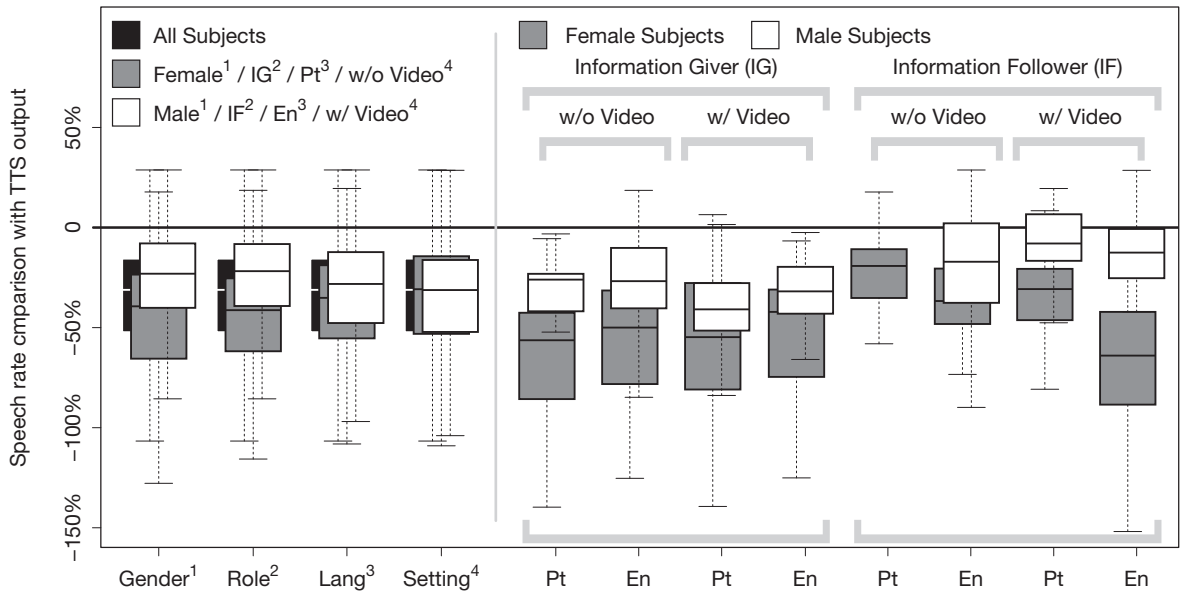


Figure 9.7: *Boxplot of speech rate adaptation to the ILMT-s2s System by group (20 frame window)*

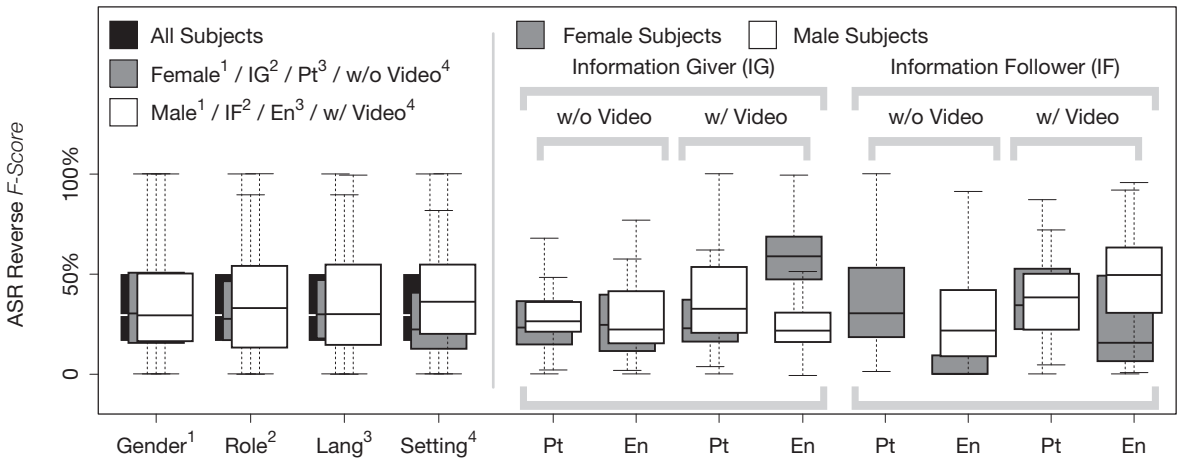


Figure 9.8: *Boxplot of Reverse F-Score by group (20 frame window)*

9.5 Ad-hoc Analysis

9.5.1 Removal of Dialogues with System Crashes

During the data collection there were four dialogues (dialogues 02, 03, 08 and 10) that experienced system crashes as indicated with the "c" markings under "Dialogue" in Table A.2 of Appendix A. A Mann-Whitney U test revealed that the speech rate of the dialogues with a system crash (w/ Crash) and without a system crash (w/o Crash) have no significant difference (Mann-Whitney U test: $W = 31,822$, $p = 0.0721$), and a negligible effect size (Cliff's δ estimate $= 0.0960$). The dialogues that included the crashes were all during the recording setting of w/ Video, so the data from the recording setting of w/o Video was removed and tested again. The Mann-Whitney U test revealed that once the data from the recording setting of w/o Video was removed the speech rate of the dialogues with a system crash (w/ Crash) and without a system crash (w/o Crash) have a significant difference (Mann-Whitney U test: $W = 10,995$, $p = 0.0292$), but still only a negligible effect size (Cliff's δ estimate $= 0.1410$). The distribution of speech rates with the differentiation of dialogues with and without crashes is illustrated in Figure 9.9.

The dialogues that included a system crash were balanced for both role and gender, but not for the spoken language, as indicated in Table 9.9. If the "w/o Crash" dialogues are removed, there are

w/ Video Setting Dialogues			
Groupings	All w/ Video (♀ : ♂)	w/ Video with Crash (♀ : ♂)	**w/ Video without Crash (♀ : ♂)**
IG — En	2 : 2	2 : 1	0 : 1
IG — Pt	2 : 2	0 : 1	2 : 1
Subtotal IG	4 : 4	2 : 2	2 : 2
IF — En	1 : 3	0 : 1	1 : 2
IF — Pt	3 : 1	2 : 1	1 : 0
Subtotal IF	4 : 4	2 : 2	2 : 2
Total	8 : 8	4 : 4	4 : 4

Table 9.9: *Proportion of system crashes within the ILMT-s2s corpus*

two Portuguese female Information Giver subjects, but no English female subjects without a crash (w/o Crash) and also only two English male Information Follower subjects, but no Portuguese male

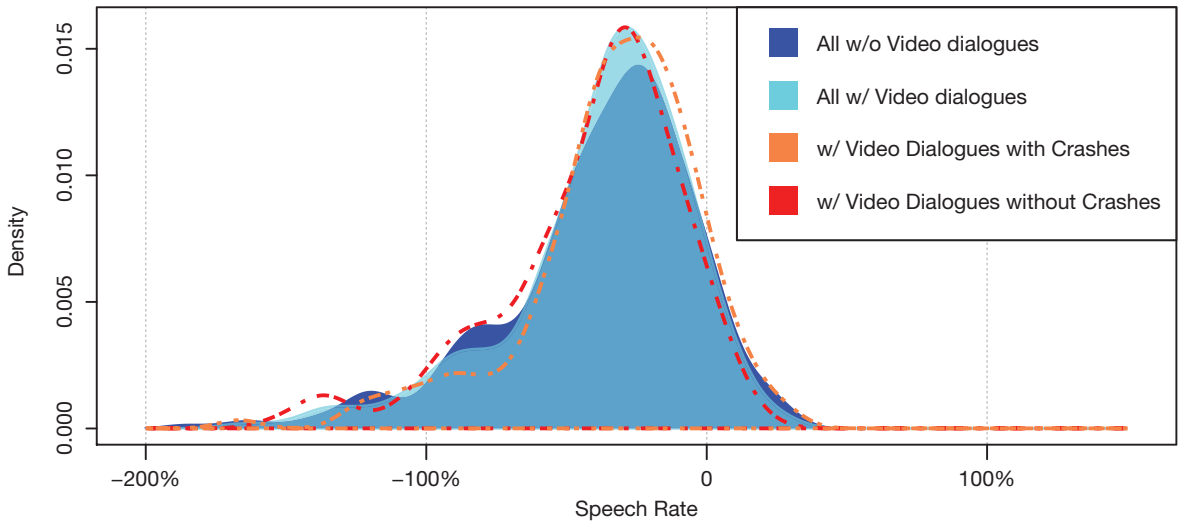


Figure 9.9: *Speech rate density plot of dialogues w/ and w/o crashes (20 frame window)*

Information Follower subjects.

For this reason, the speech rate of all female Information Giver subjects, and the speech rate of all male Information Follower subjects was compared.

- For the female subjects:

  A Mann-Whitney U test on all English speaking female subjects and Portuguese speaking female subjects in the "w/ Video" setting resulted in no significant difference ($W = 825$, $p = 0.8136$) between the English and Portuguese speaking female subjects.

- For the male subjects:

  A Mann-Whitney U test on all English and Portuguese speaking male subjects in the "w/ Video" setting resulted in no significant difference ($W = 473$, $p = 0.1599$) between the English and Portuguese speaking male subjects.

With these results and also from the similarities of the plot in Figure 9.7, the removal of the dialogues with a system crash (w/ Crash) from the data will not distort the overall pattern of the results of gender and role as indicated in the similar values of Table 9.10.

	Speech Rate Difference (%)				Reverse F-Score			
Category	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender (♀ : ♂) All Dialogues	−47.89	34.81	−25.86	27.99	34.32	24.31	33.97	23.35
Diff.	−0.10	−0.92	−2.69	−1.15	4.60	1.81	2.11	0.13
Gender (♀ : ♂) w/o Crash	−47.99	35.73	−28.55	29.24	29.72	22.50	31.86	23.22
Role (IG : IF) All Dialogues	−47.94	33.07	−27.27	30.96	32.81	21.62	35.51	25.86
Diff.	−2.24	−0.80	−0.86	−0.14	5.10	4.27	1.84	1.08
Role (IG : IF) w/o Crash	−50.18	33.87	−28.13	31.10	27.71	17.35	33.67	26.94
Language (Pt : En) All D.	−41.35	33.50	−33.87	33.40	33.72	21.93	34.60	25.66
Diff.	−4.05	−0.27	−0.96	−0.35	1.36	1.27	5.58	0.92
Language (Pt : En) w/o Crash	−45.40	33.77	−32.91	33.75	32.36	20.66	29.02	24.74
Setting (w/o Vid : w/ Vid) All	−38.03	35.11	−37.24	32.34	29.18	23.58	38.52	23.26
Diff.	0.00	0.00	−3.88	−0.51	0.00	0.00	5.18	1.99
Setting (w/o Vid : w/ Vid) w/o Crash	−38.03	35.11	−41.12	32.85	29.18	23.58	33.34	21.27

Table 9.10: *Comparison of speech rate adaptation to the ILMT-s2s System by groups before and after dialogues with system crashes are removed (20 frame window)*

9.5.2 Analysis Results After System Crash Dialogues are Removed

With the dialogues that included a system crash removed, the significance test (Mann-Whitney U test) and effect size were verified again with similar results as when the w/ Crash data was not removed. Results from the tests are listed in Table 9.11.

Speech Rate Comparison							
Grouping	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	ES
Role (IG : IF)	−43 : −23	−50 : −28	34 : 31	$p = 4.87e-15$	***	$\delta = 0.432$	med.
Gender (♀ : ♂)	−40 : −26	−48 : −29	36 : 29	$p = 3.043e-09$	***	$\delta = 0.328$	small
Lang. (Pt : En)	−39 : −28	−45 : −33	34 : 34	$p = 1.787e-05$	***	$\delta = 0.236$	small
Setting (w/o Vid : w/ Vid)	−31 : −34	−38 : −41	35 : 33	$p = 0.2368$	–	$\delta = 0.068$	neg.
ASR Performance (Reverse *F*-Score) Comparison							
Setting (w/o Vid : w/ Vid)	22 : 31	29 : 33	24 : 21	$p = 0.0097$	**	$\delta = 0.148$	small
Lang. (Pt : En)	28 : 22	32 : 29	21 : 25	$p = 0.0109$	*	$\delta = 0.140$	neg.
Role (IG : IF)	25 : 30	28 : 34	17 : 27	$p = 0.1916$	–	$\delta = 0.072$	neg.
Gender (♀ : ♂)	25 : 27	30 : 32	22 : 23	$p = 0.3104$	–	$\delta = 0.056$	neg.

Table 9.11: *Significance test and effect size between groupings (20 frame window: Without system crash dialogues)*

As a visual indication of how the data changed, boxplots illustrating the data before the w/ Crash dialogues were removed and after they were removed are provided in Figure 9.10, Figure 9.11.

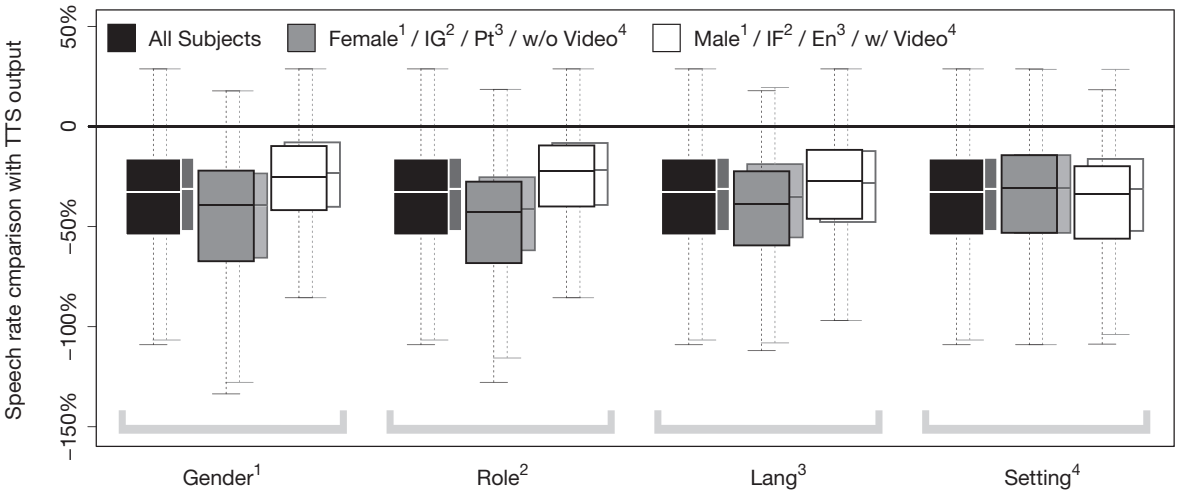


Figure 9.10: *Boxplot of speech rate adaptation to the ILMT-s2s System by group (20 frame window: Boxplots with "w/o Crash" included are displayed behind the boxplots with the "w/ Crash" removed)*

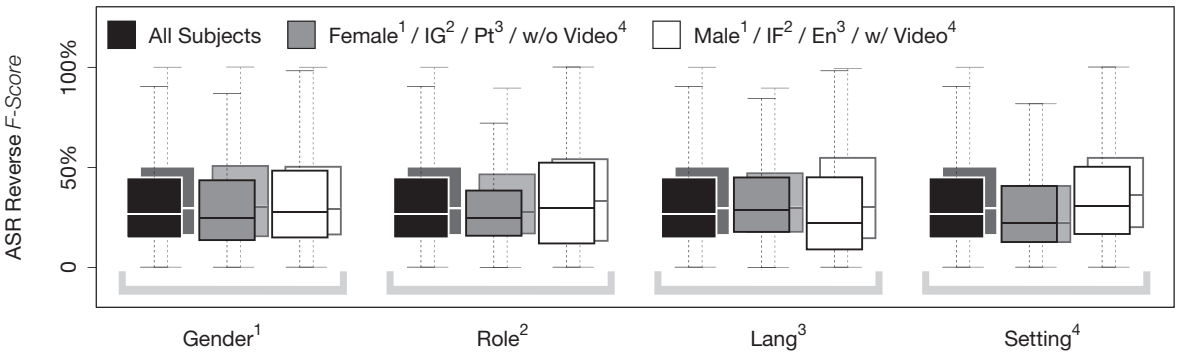


Figure 9.11: *Boxplot of ILMT-s2s System ASR Reverse F-Score by group (20 frame window: Boxplots with "w/o Crash" included are displayed behind the boxplots with the "w/ Crash" removed)*

9.6 Summary

The statistical analysis of the subjects' *On-Talk* speech rate and the ASR performance have provided the following contributions to the understanding of interlingual computer mediated communication.

■ Key Findings.

- The mean speech rate of female subjects is slower than the mean speech rate of male subjects, however, the ASR performance (Reverse F-Score) is very similar.[6]
- The mean speech rate of Information Giver subjects is slower than the mean speech rate of Information Follower subjects, however, the ASR performance (Reverse F-Score) is very similar.[7]

9.7 Discussion

A point of interest that did arise, but not studied further, is related to an indication from Branigan et al. (2010, p. 2366) about the recovery of the subjects interacting with a computer system;

> **"**Alignment is often actually stronger when people are interacting with computers than with other people. But its strength depends upon speakers' beliefs about how to achieve successful communication with computers, which in turn depends on beliefs about their capability. We hypothesize that people may not feel confident in their ability to make accurate judgements about the capability of a computer, and might therefore tend to adopt a conservative initial model of the computer. Having adopted such a model, they might be reluctant to update it dynamically, or they might require substantial evidence of the computer's abilities before updating it.**"** (Branigan et al., 2010, p. 2366)

The fifteen dialogues of the computer-mediated dialogue dataset include four dialogues that had a system crash. The point of interest with these four dialogues is the recovery of the subject's speech rate after the crash. As mentioned by Branigan et al. (2010, p. 2366) in the extracted quote above, Branigan et al. presume that "substantial evidence" is needed to change the communication strategy used by the subject towards a computer system. From a brief glance at the data from the system crash dialogues illustrated in Figure 9.12, the speech rate recovers after the system crash, which are located with the grey vertical bars. This could be related to the fact that the subject talked to the technician during the reboot of the system and maintained the faster speech rate used for *Off-Talk Other* (explained in Chapter 6), but it could also be categorised as the "substantial evidence" required to change the communication strategy. With the improvement of operating systems and software applications, the habit of rebooting a computer to fix all problems is becoming a phenomenon of the past, but it is being replaced by an increasing frequency in software and operating system updates. Though managing the user's expectations will be a major part of creating the "substantial evidence",

[6] Plots of individual speakers can be found in figures Figure A.14 to Figure A.17 of Appendix A.
[7] Plots of individual speakers can be found in figures Figure A.14 to Figure A.17 of Appendix A.

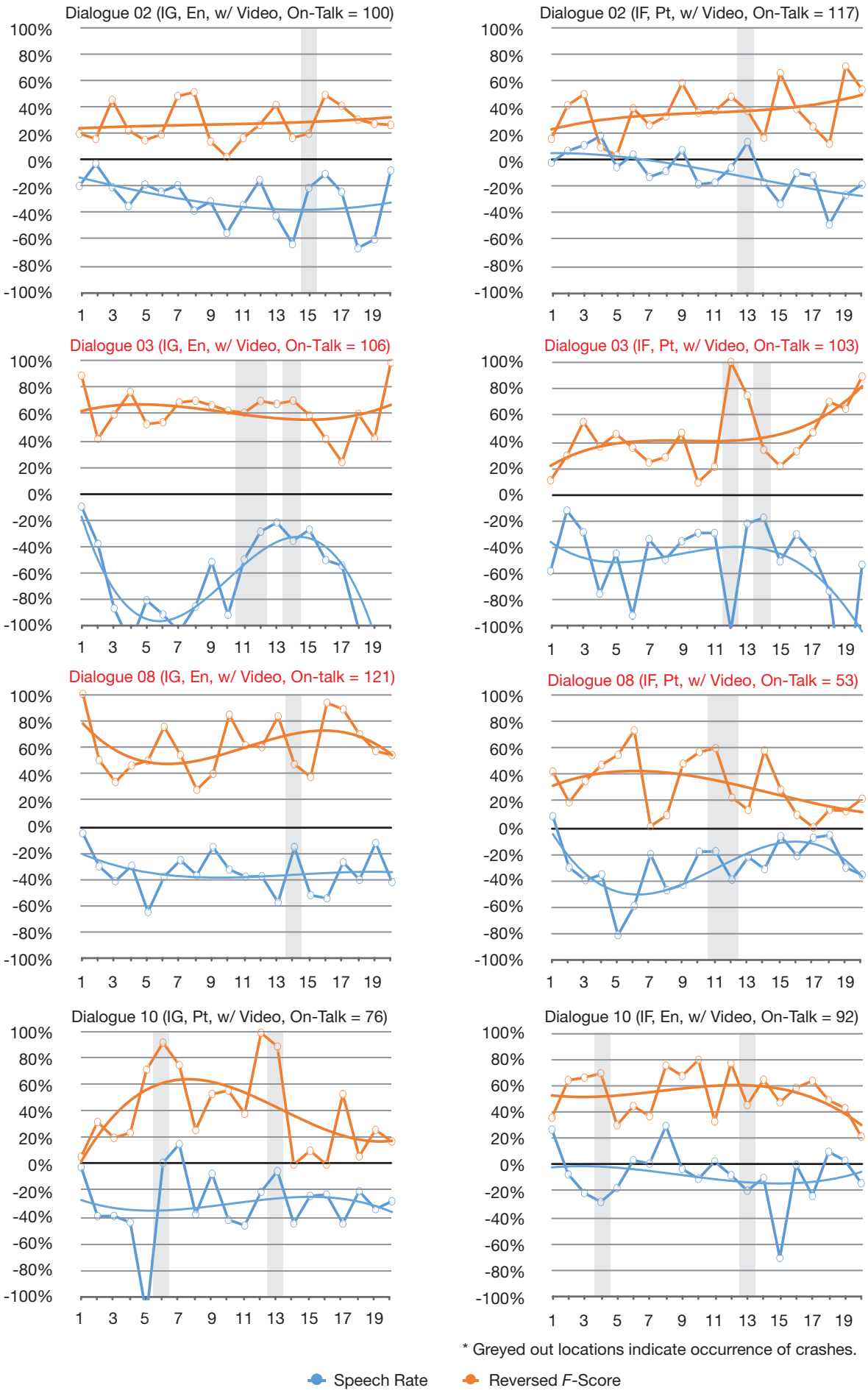


Figure 9.12: *Plots of subject's speech rate and Reverse F-Score — Dialogues with crashes*

since the perception of the user can be easily manipulated. Pearson, Hu, Branigan, Pickering, & Nass (2006, p. 1177) reported that "[when] users believe that the system is unsophisticated and restricted in capability, they adapt their language to match the system's language more than when they believe the system is relatively sophisticated and capable". This difference of "unsophisticated and restricted in capability" and "relatively sophisticated and capable" was achieved by only changing details written on the start-up screen to indicate the value ($10 and $299), relative age of the software (1987 or 2003) and reported quality of the software (Basic and Professional) while the performance of the actual software remained the same (Pearson et al., 2006, p. 1178). The presumed increase in speech rate that can be seem after the system crashes, the grey vertical bar in Figure 9.12, could be related to the presumption that the reboot of the system fixed the problems that the subjects were encountering, but this assumption may only be related to the generation that experienced the operating system of Windows 95 onwards to Windows XP.

The techniques preposed by L. Bell et al. (2003), Oviatt, Darves, & Coulston (2004), Suzuki & Katagiri (2007) where the system interacts in a different way to try and change the users' speech rate unconsciously, may be a good baseline implementation, but once the communication has started to fail, as demonstrated in this section, the speech rate did not recover. Implementing further actions that act on the user's subconscious, may also provide the opposite affect. For example when one lowers their voice to indicate to the interlocutor that they are speaking too loudly, but is replied with an even louder voice from their interlocutor who misinterpreted this cue as an mediating device problem. As a future study, it would be interesting to see if a system reboot, or equivalent action, can act as the required "substantial evidence" to improve the subject speech rate, and also how long this effect would last on the subject.

Chapter 10

ASR Performance & Perceived Emotions Analysis

In the previous four chapters, it has been identified from the computer-mediated dialogue data that;

(i) Chapter 6: I report that there are three talk types (*On-Talk*, *Off-Talk Self*, and *Off-Talk Other*) with different overall mean speech rate differences with the *On-Talk* being the slowest and *Off-Talk Other* being the fastest.

(ii) Chapter 7: I report that the subjects of the computer-mediated dialogue dataset, who's communication is mediated by a computer, use a slower speech rate than the subjects of the non-mediated dialogue dataset who talk directly to each other. Furthermore, though the patterns of speech rate difference were similar for role and setting (with or without eye-contact), the speech rate differences between gender was not similar between datasets. The female subjects of the computer mediated communication were using a slower speech rate than their male counterparts.

(iii) Chapter 8: I report that the Information Giver uses less words in their dialogue acts with more turns and the Information Follower uses more words and less turns when compared to the subject's of non-mediated dialogue dataset (the HCRC Edinburgh Map Task corpus).

(iv) Chapter 9: I report that though the speech rate of the subject *On-Talk* was affected by the ASR performance of the ILMT-s2s System, the role and the gender of the subject also affected the speech rate of the subject.

Here, I report the analysis of the Automatic Speech Recognition (ASR) performance and the perceived emotions of the subjects.

One fundamental principle that I was taught during my time in industry was not to expect one's counterpart to be the convenient user. A convenient user is someone that is good tempered, intelligent enough to overcome all of my short comings, patient with all the time in the world to study problems,

and appreciative of the slightest of aiding actions that I perform. No, a tool should never be designed with a convenient user in mind. A simple example of this is when I found a cosmetic error in a dissertation. Every hyperlink indicating a page numbering is wrong by at least 1 page from page number 46 to 156. I mentioned this to a few people and their replies were, (i) "Why does that happen?", (ii) "I know how to fix that problem", and (iii) "Why is an important document written using a software with such a major flaw? You've informed the designers to fix it haven't you. How many more people do you want to see enduring this problem/shame/embarrassment!". This is a well documented procedure with LaTeX, where the build needs to be performed before and after a feature is invoked and then built again, to obtain the desired output. Unfortunately it has not yet entered the "common knowledge" basic understanding of LaTeX users. The author of the minor error has 29 publications to their name at the time the dissertation was completed, so is experienced at writing. The person who replied "Why does that happen?", is a devout Linux and LaTeX user who will endlessly discuss the merits of using automated script and terminal command line commands in the place of a software that require a click on a graphical UI and even makes slides and posters using LaTeX. The point is, the knowledge is already available, but since the knowledge is not common, so it is the equivalent of not being available at all. So as my past adviser said, the software should provide the safety net for the user. To be able to provide a safety net, the aspects of the software that cause a negative reaction from the user needs to be studied.

Botherel & Maffiolo (2006, p. 3) indicated in a field study of two call-centres reported that "[in] the interaction between customers and commercial advisers in the case of call centres dealing with telecommunication products, the emotional attitudes mainly detected in the customers' speech are negative, linked up with difficulties they have" and Korczynski (2003, p. 57) in a previous study reported that "[enchantment] may easily turn to disillusionment in the moments when the individual customer's lack of sovereignty becomes starkly apparent. [...] This anger is vented at the closest symbol of customer's painful journey from enchantment to disillusionment: the front-line service worker". This is a prime example scenario where the tool (the Speech-to-Speech Machine Translation (S2S-MT) system) should not add further negativity into the communication and understanding where the negativity of "*Surprised*" and "*Frustrated*" occur within the dialogues of the computer-mediated dialogue data is of interest. Of course one could remove the tool altogether, but as Batliner et al. (2000, p. 1) indicated, "[the] potential market for automatic dialogue systems used in call centers for instance is growing rapidly" but also cautioned that the quality of these systems were questionable. Now with the introduction of Skype Translator (W. D. Lewis, 2015), VoiceTra (Matsuda et al., 2013) it is an expected evolution that S2S-MT systems become a more common median of communication. In this section I study the perceived emotion of the subjects during their interaction with the ILMT-s2s System while completing the map task using the perceived emotion labels of "*Surprised*", "*Frustrated*", and "*Amused*" that have been annotated (§ 4.4.3) in the computer-mediated dialogue dataset. The analysis was performed by establishing a linkage between the Cognitive states of the subjects and

their utterances appropriately, and then observing the ASR performance (by the given word error rate (WER) of those utterance) with the associated perceived emotion.

10.1 Introduction

The analysis of the perceived emotion labels was conducted to see if it is possible to quantify the user experience of the ILMT-s2s System communication. A qualitative measure could be achieved from an analysis of the user survey results (Chapter 11: "User Survey Study"), however, to counter balance the findings from the user survey analysis, this quantitive measure was added.

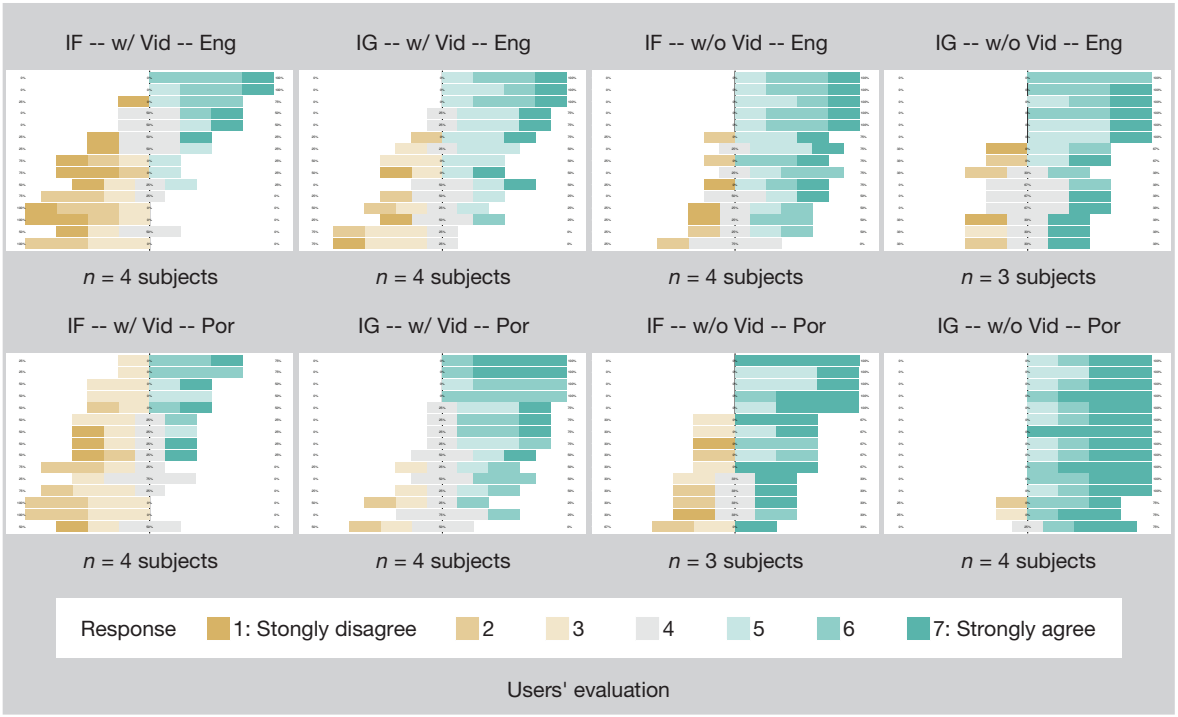


Figure 10.1: *Stacked bar chart of subject survey Likert scale replies — Role, setting and language*

As illustrated in Figure 10.1, from the analysis of the user survey, it was possible to understand that the subjects that could not see their interlocutor (w/o Vid) perceived the ILMT-s2s System better than subjects who could see their interlocutor (w/ Vid). Also, the subject with the role of Information Giver perceived the ILMT-s2s System better than subjects with the role of Information Follower. The clear distinction that was reported by the user survey, presented in Figure 10.1, is not so obvious from just a simple distribution count of the perceived emotions, as illustrated in Figure 10.2. For example, with the English speaking subjects, the subjects in the setting of w/ Video are indicating amusement more than the subject in the w/o Video setting, but they didn't perceive the system as highly. For the Portuguese speakers, there are more frustration labels in the w/ Video setting, but they are also high in the w/o Video setting too.

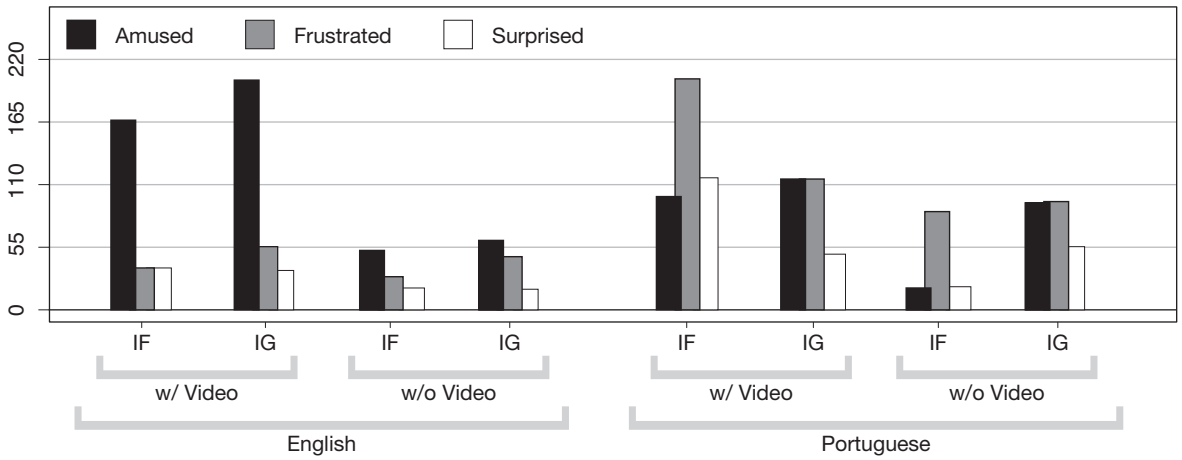


Figure 10.2: *Distribution of perceived emotions within the ILMT-s2s corpus — Summarised by groupings role, setting and language*

10.2 Dataset

In this chapter, the perceived emotion labels described in § 3.3.3.2 and the ASR performance labels described in § 3.3.3.5 of Chapter 3 are used for the analysis.

10.3 Methods

Korczynski (2003, p. 57) reported that "[enchantment] may easily turn to disillusionment in the moments when the individual customer's lack of sovereignty becomes starkly apparent. [...] This anger is vented at the closest symbol of customer's painful journey from enchantment to disillusionment: the front-line service worker". Botherel & Maffiolo (2006, p. 3) indicated that "[in] the interaction between customers and commercial advisers in the case of call centres dealing with telecommunication products, the emotional attitudes mainly detected in the customers' speech are negative, linked up with difficulties they have".

This is a prime example scenario where the communication tool should not add further negativity into the communication. Therefore, it is important to understand if negativity has occurred during the use of such a system.

10.3.1 Linking Perceived Emotion Labels with Communication Items

The perceived emotion labels (page 82) are used to identify which utterances might be causing a change to the perceived emotion of the subject. Unlike the ASR performance values or the talk type labels, the perceived emotion labels are not linked to a utterance, but independent of the utterances. From the assumption that the changes to the perceived emotion derives from the communication the perceived emotion labels are linked to the utterances of the subject and the TTS output from the communication system. All perceived emotion labels, *On-Talk* and *Off-Talk* utterances, and TTS output occurrences were listed in chronological order with the perceived emotions linked to

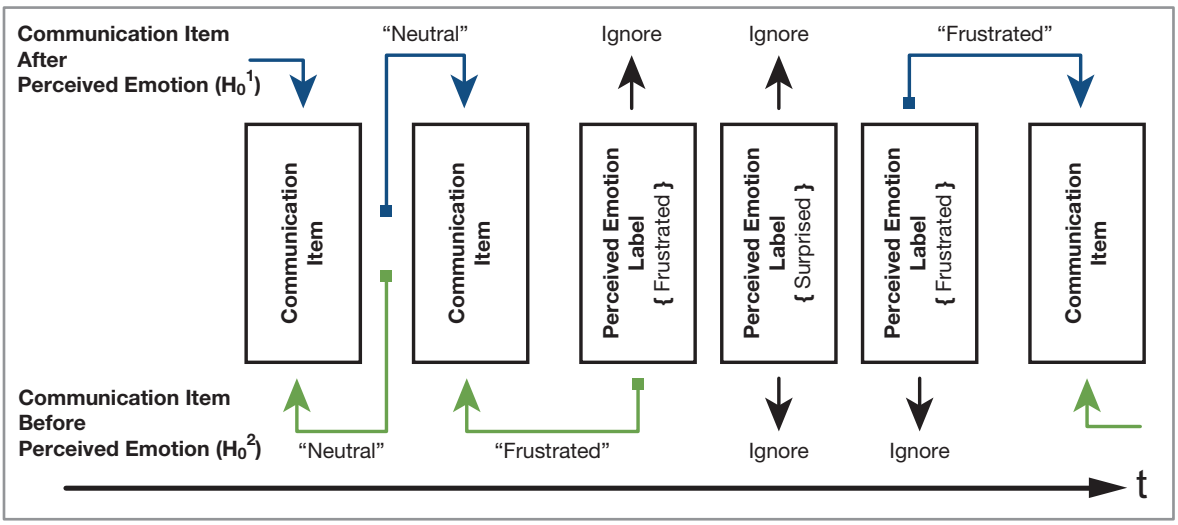


Figure 10.3: *Linkage example of perceived emotion and communication item*

the communication item immediately before and after the perceived emotion label. If no perceived emotion label is present between the communication items, the communication item is identified as neighbouring a "*Neutral*" perceived emotion as illustrated with the first two "Communication Items" of Figure 10.3. Also if perceived emotions changed without any communication item in between them, these perceived emotions are ignored as indicated with the "Perceived Emotion Label" with the arrow pointing to "Ignore" in Figure 10.3. Only perceived emotions neighbouring communication items are taken into consideration as illustrated by the two "Frustrated" perceived emotion labels of Figure 10.3. The linkage in both directions, after and before the communication item, are created to verify if the perceived emotions are affecting the ASR performance ($H_0{}^1$) or if the ASR performance is affecting the perceived emotion of the subject ($H_0{}^2$).which is explained in the next § 10.3.2.

10.3.2 Direction of the Analysis

As mentioned in § 10.3.1, the perceived emotion labels are linked to the communication item immediately before and after. The underlying hypothesis was that perceived emotions that occur immediately before a communication item, affect the ASR performance of the communication item, and this in turn affects the user's perceived emotion after seeing the ASR result. This translates into two separate null hypotheses, by comparing the *On-Talk* WER (the response variable) and perceived emotions (the independent variable) as illustrated in Figure 10.4.

$H_0{}^1$: The mean WER of utterances made *after* different perceived emotions occur is the same.

$H_0{}^2$: The mean WER of utterances made *before* different perceived emotions occur is the same.

10.3.3 Analysis by Grouping

The analysis of the perceived emotion labels are divided by setting and language. For the setting difference of being able to see the interlocutor (w/ Video) or not (w/o Video), differing communication

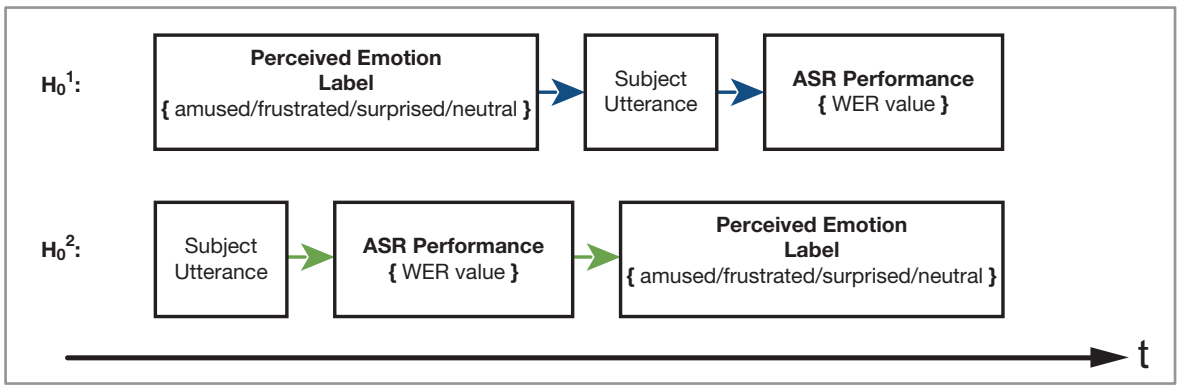


Figure 10.4: *Flow of hypothesis (H1 and H2) — WER value is the same what ever the preceding or following perceived emotion may be (amused, frustrated, or surprised).*

styles have been reported between Face-to-Face and mobile phone usage in a Japanese publication by Yoshida and Kakuta (Morikawa & Maesako, 1998, p. 149). For this reason, the ASR performance of the utterances linked to the perceived emotion labels are studied as follows:

The ASR performance of utterances linked to the four perceived emotion labels are tested to verify if there is a significant difference between *[Refer to grouping list below]* (*ANOVA*).
Grouping list:

1. Setting: w/ Eye-Contact (w/ Video)
2. Setting: w/o Eye-Contact (w/o Video)
3. w/ Eye-Contact: English Speaking Subjects
4. w/ Eye-Contact: Portuguese Speaking Subjects

10.4 Results

10.4.1 Balancing Perceived Emotion Labels' Count

The quantitative difference between the w/ Video and w/o Video count of perceived emotions illustrated in Figure 10.2 could be described as misleading due to the different duration and uneven quantity of the two settings. The mean duration of w/ Video dialogues is 45m 20s whereas the mean duration of w/o Video is 30m 58s as summarised in Table 10.1. Aside from the dialogue duration

Dialogues	Count	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
w/ Video	8	23m 2s	34m 24s	45m 48s	45m 20s	51m 35s	1h 14m 20s	16m 31s
w/o Video	7	19m 38s	24m 15s	26m 13s	30m 58s	39m 8s	44m 7s	9m 38s

Table 10.1: *Summary and standard deviation of dialogue duration — by setting*

differences, with the w/ Video being longer, the w/o Video had one less dialogue recording, which would decrease the total count of w/o Video perceived emotions. Therefore, to balance the perceived emotion label's count between both settings, the communication item (subject utterance or Text-to-Speech (TTS) output) count was used as a reference to determine the frequency of the perceived emotion label's occurrence.

10.4.2 Linking Perceived Emotion Labels with Communication Items

The perceived emotions labels (§ 4.4.3.1) of the computer-mediated dialogue data are linked to utterances of the subject as explain in § 10.3.1 and illustrated in Figure 10.5.

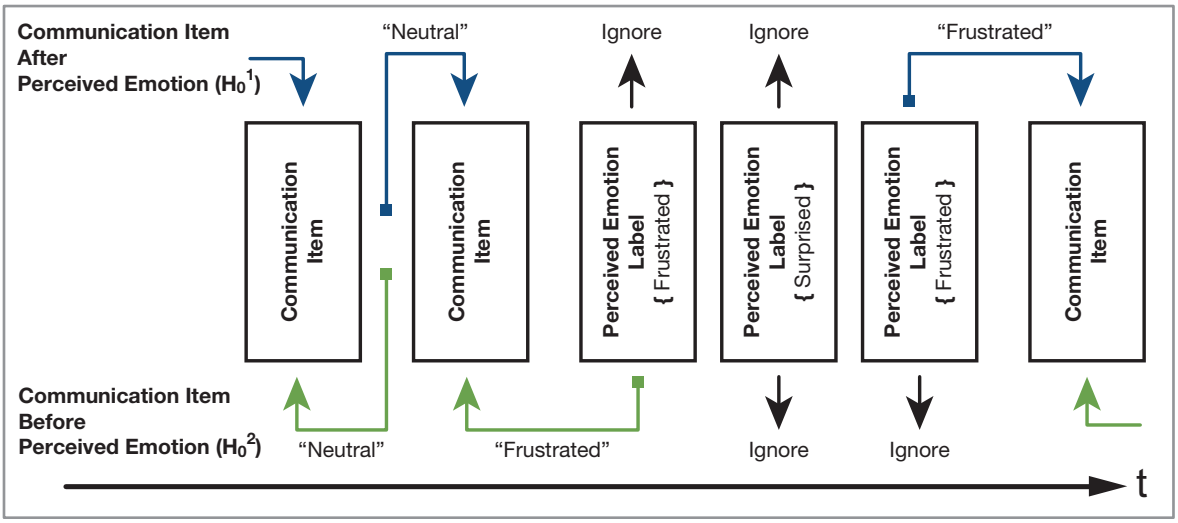


Figure 10.5: *Linkage example of perceived emotion and communication item*

Of the 1,835 labels, approximately 295 labels were ignored and approximately 1,540 were linked to communication items. The result of the bi-directional linkage ($H_0{}^1 : H_0{}^2$), is listed in Table 10.2.

Perceived emotions	Count	Total duration (approx.)
Amused	812	1 h 33 m
Linked to *On-Talk* ($H_0{}^1 : H_0{}^2$)	314 : 278	32 m : 31 m
Linked to *Off-Talk* ($H_0{}^1 : H_0{}^2$)	188 : 130	25 m : 14 m
Linked to TTS ($H_0{}^1 : H_0{}^2$)	209 : 254	27 m : 29 m
Not linked ($H_0{}^1 : H_0{}^2$)	101 : 151	9 m : 19 m
Frustrated	670	51 m
Linked to *On-Talk* ($H_0{}^1 : H_0{}^2$)	261 : 279	16 m : 22 m
Linked to *Off-Talk* ($H_0{}^1 : H_0{}^2$)	100 : 109	8 m : 7 m
Linked to TTS ($H_0{}^1 : H_0{}^2$)	200 : 160	17 m : 10 m
Not linked ($H_0{}^1 : H_0{}^2$)	109 : 122	9 m : 12 m
Surprised	353	17 m
Linked to *On-Talk* ($H_0{}^1 : H_0{}^2$)	127 : 79	7 m : 4 m
Linked to *Off-Talk* ($H_0{}^1 : H_0{}^2$)	48 : 61	2 m : 1 m
Linked to TTS ($H_0{}^1 : H_0{}^2$)	91 : 191	4 m : 11 m
Not linked ($H_0{}^1 : H_0{}^2$)	87 : 21	4 m : 1 m
Linked Subtotal ($H_0{}^1 : H_0{}^2$)	1,538 : 1,541	2 h 20 m : 2 h 9 m
Not Linked Subtotal ($H_0{}^1 : H_0{}^2$)	297 : 294	21 m : 32 m
Total	1,835	2 h 41 m

Table 10.2: *Summary of perceived emotions annotation linkage in the ILMT-s2s corpus.*

10.4.3 Direction of the Analysis

The perceived emotion labels were linked to the communication item immediately before and after as explained in § 10.4.2. The linkage in both directions, was created to verify if the perceived emotions were affecting the ASR performance ($H_0{}^1$) or if the ASR performance was affecting the perceived

emotion of the subject ($H_0{}^2$). The underlying hypothesis was that perceived emotions that occur immediately before a communication item, affect the ASR performance of the communication item, and this in turn affects the user's perceived emotion after seeing the ASR result. This translates into two separate null hypotheses, by comparing the *On-Talk* WER (the response variable) and perceived emotions (the independent variable) as illustrated in Figure 10.6.

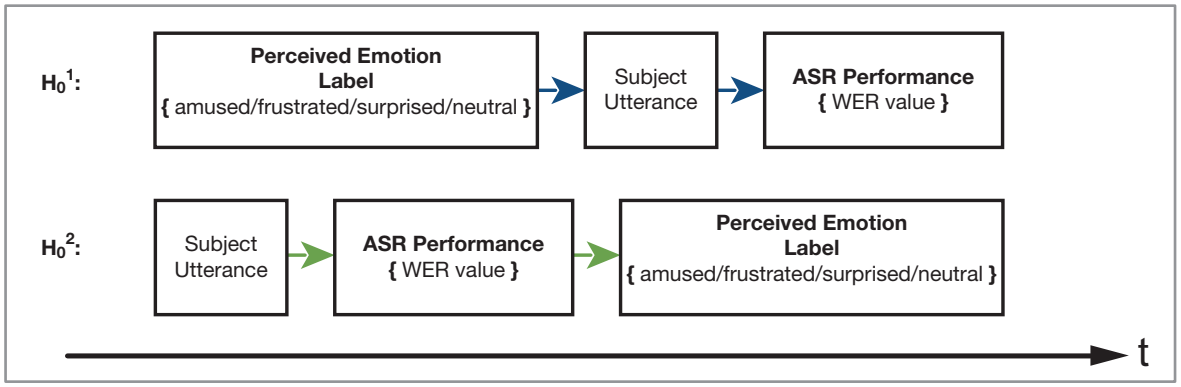


Figure 10.6: *Flow of Hypothesis (H1 and H2) — WER value is the same what ever the preceding or following cognitive state may be (Amused, Frustrated, or Surprised).*

$H_0{}^1$: The mean WER of utterances made *after* different cognitive states (*Amused* = 309, *Frustrated* = 253, *Surprised* = 124, and *Neutral* = 1,684) occur is the same for all these states.

$H_0{}^1$: No statistically significant difference (ANOVA: $F_{3,2366} = 0.395; p = 0.757$) was observed between the perceived emotions before the utterance that was sent to the ASR and the resulting ASR WER.

$H_0{}^2$: The mean WER of utterances made *before* different cognitive states (*Amused* = 271, *Frustrated* = 263, *Surprised* = 78, and *Neutral* = 1,758) occur is the same for all these states.

$H_0{}^2$: A statistically significant difference (ANOVA: $F_{3,2366} = 9.076; p = 5.67e-06$) was observed between the perceived emotions after utterances sent to the ASR and the resulting ASR WER of that utterance.

For $H_0{}^2$, Post-hoc comparisons (Tukey HSD test) revealed significant differences between *Neutral* and *Frustrated* ($p = 0.0000942$), *Neutral* and *Surprised* ($p = 0.0067592$), and *Amused* and *Frustrated* ($p = 0.0311424$).

■ Key Findings.

The analysis of the correlation between the ASR performance, and the perceived emotion of the subjects revealed a significant difference (ANOVA: $F_{3,2366} = 9.076; p = 5.67e-06$) between the perceived emotion that occur after utterances are sent to the ASR system and the resulting ASR WER of that utterance ($H_0{}^2$) — in other words, the data indicates that for communication items with the system (*On-Talk*), the ASR performance (WER) is triggering the perceived emotion and not vice versa.

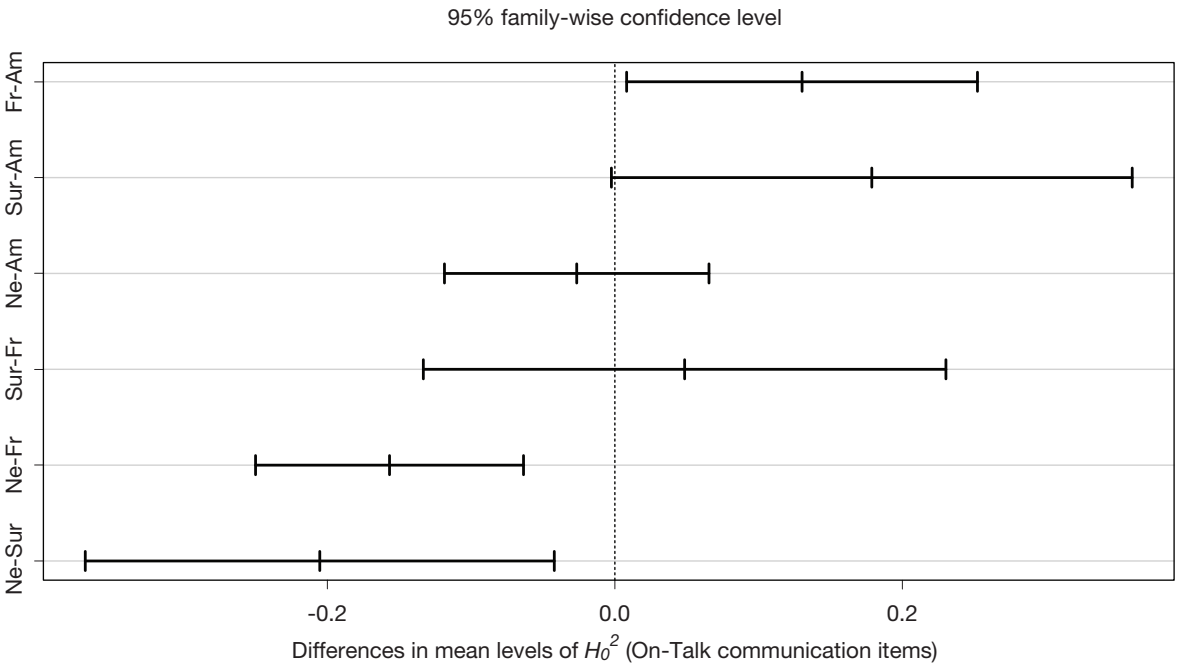


Figure 10.7: *Plot of Tukey HSD test results of* $H_0{}^2$ *ANOVA output. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)*

10.4.4 Analysis by Grouping

Since it was identified that the ASR performance was affecting the perceived emotion of the subject, so the linkage of the labels to the communication items for $H_0{}^1$ were discarded and this analysis concentrates on the data with the perceived emotion labels linked to the communication item immediately before ($H_0{}^2$), as illustrated in Figure 10.8.

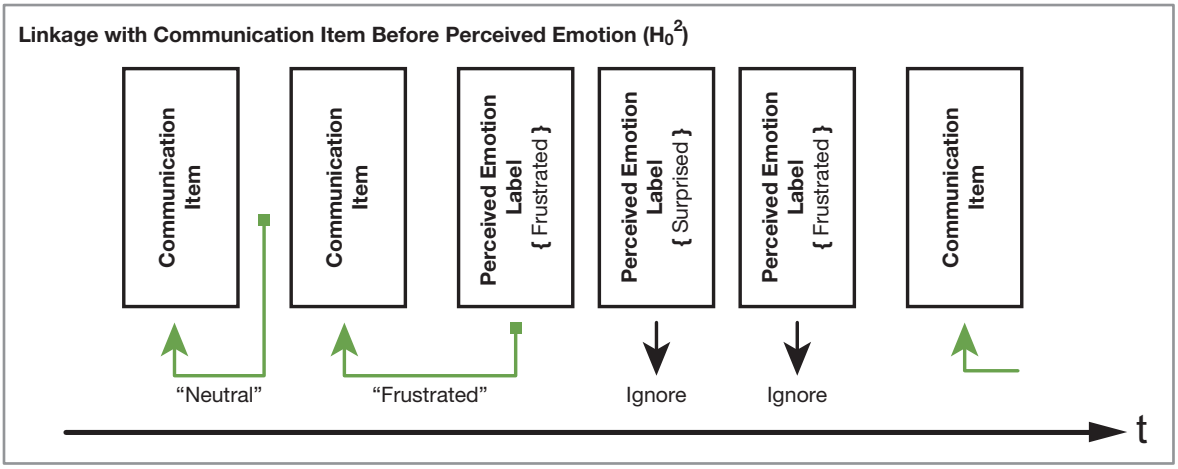


Figure 10.8: $H_0{}^2$ *linkage example of perceived emotion and communication item*

■ Visualisation of the ASR Performance & Perceived Emotion.

The data was visualized in order to understand it. A bar chart of the perceived emotion labels linked to the communication items shows an interesting pattern when it is divided into groups of language, setting and role as illustrated in Figure 10.9.

For the English speaking subjects in Figure 10.9, it is clear that the perceived emotion of *Amused*

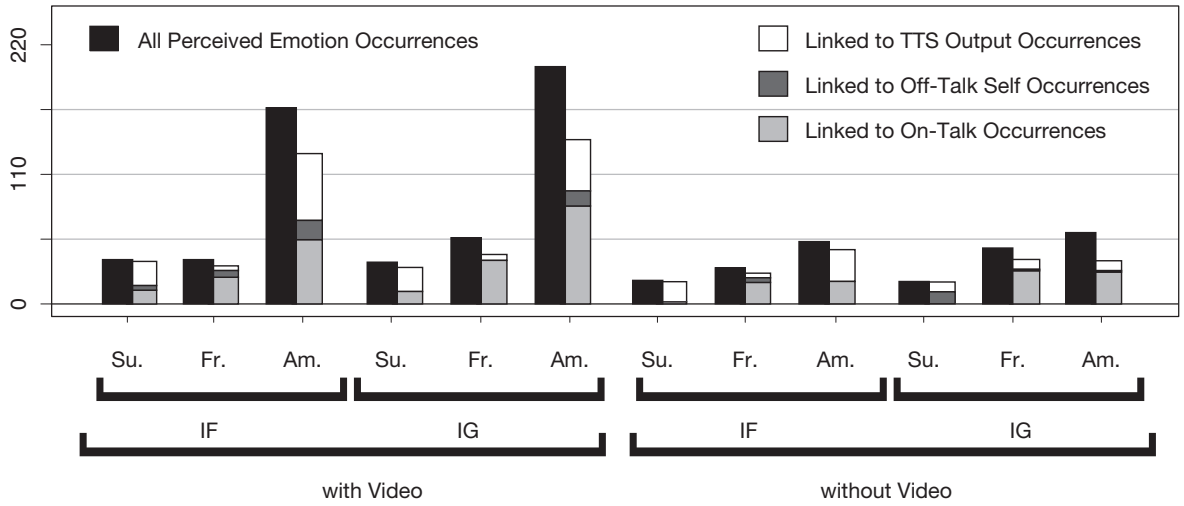


Figure 10.9: *Bar chart of English subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused) divided by role and setting*

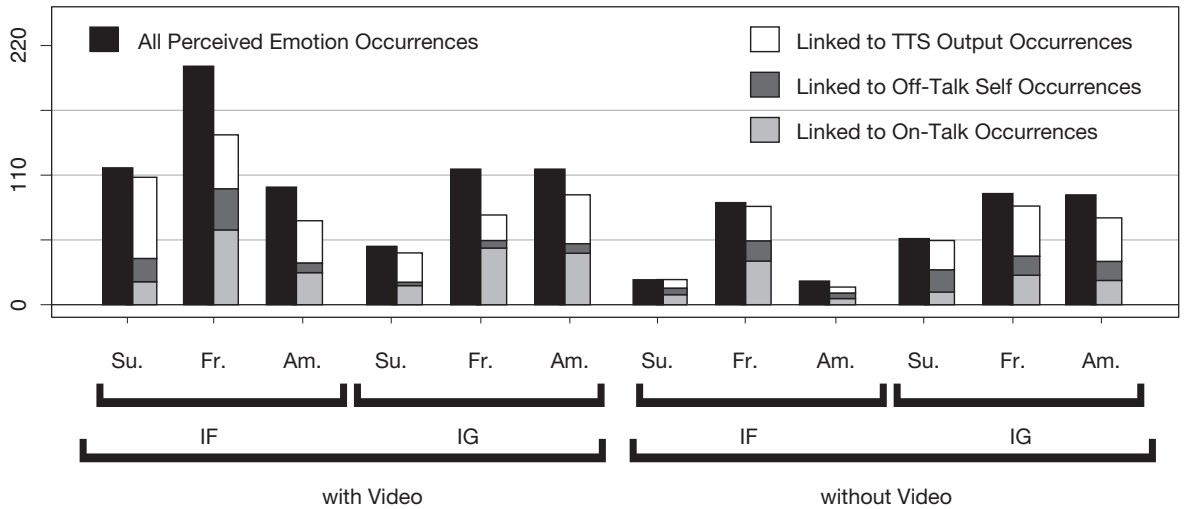


Figure 10.10: *Bar chart of Portuguese subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused) divided by role and setting*

is dominant in the w/ Video setting while in the w/o Video setting, dominance is shared between *Amused* and *Frustrated*. Looking at the Portuguese speaking subjects in Figure 10.10, it is clear that the perceived emotion of *Amused* is no longer the dominant perceived emotion, and it has been replaced by *Frustrated*. Another difference between the English and Portuguese subjects is the increase in perceived emotions that are linked to the TTS output. To understand any patterns related to the perceived emotions linked with the TTS output, the ASR performance of the interlocutor will need to be identified and linked. For this reason, this analysis has not been conducted and the following analysis is limited to the labels' relation with the ASR performance of the subject utterance.

■ ASR Performance & Perceived Emotion.

From Figure 10.9 and Figure 10.10, it is clear that there is a different pattern when the data is divided by the settings of w/ Video and w/o Video. Looking at the greater difference between median (*Mdn*) and mean (*M*) values in Table 10.3 confirms this pattern.

Setting	*Mdn*	*M*	*SD*	Count
w/o Video — Linked to Perceived Emotion	0.276	0.408	0.50	198
— *Neutral*	0.200	0.371	0.69	836
w/ Video — Linked to Perceived Emotion	0.500	0.521	0.44	401
— *Neutral*	0.250	0.390	0.45	917
All — Linked to Perceived Emotion	0.400	0.483	0.46	599
— *Neutral*	0.222	0.381	0.58	1,753

Table 10.3: *Subject's perceived emotions by setting — WER*

- A significant difference was not identified in w/o Video (ANOVA $F_{3,1030} = 0.582; p = 0.627$).
- A significant difference was identified in w/ Video (ANOVA $F_{3,1314} = 14.22; p = 3.99e-09$).

Within the w/ Video grouping, looking at the difference between median (*Mdn*) and mean (*M*) values within the grouping of w/ Video in Table 10.4, it is clear that both English and Portuguese subjects are reacting to the ASR performance.

Setting	*Mdn*	*M*	*SD*	Count
w/ Video–Pt — Linked to Perceived Emotion	0.333	0.441	0.47	204
— *Neutral*	0.214	0.351	0.41	440
w/ Video–En — Linked to Perceived Emotion	0.643	0.604	0.39	197
— *Neutral*	0.286	0.426	0.48	477
w/ Video — Linked to Perceived Emotion	0.500	0.521	0.44	401
— *Neutral*	0.250	0.390	0.45	917

Table 10.4: *Subject's perceived emotions by setting of w/ Video — WER*

- A significant difference was reported in English (ANOVA $F_{3,670} = 10.61; p = 8.06e-07$).
- A significant difference was reported in Portuguese (ANOVA $F_{3,640} = 11.52; p = 2.31e-07$).

Post-hoc comparisons (Tukey HSD test) revealed the following significant differences.

- For English subjects:

  Amused or *Neutral* with *Frustrated* ($p < 0.01$) with an effect size of medium (Cohen's d estimate: 0.622) for *Amused–Frustrated* and large (Cohen's d estimate: 0.990) for *Neutral–Frustrated* as illustrated in Figure 10.11.

- For Portuguese subjects:

  Amused or *Neutral* with *Frustrated* or *Surprised* ($p < 0.001$) with small (Cohen's d estimate: 0.355) to large (Cohen's d estimate: 1.126) effect sizes as illustrated in Figure 10.12.

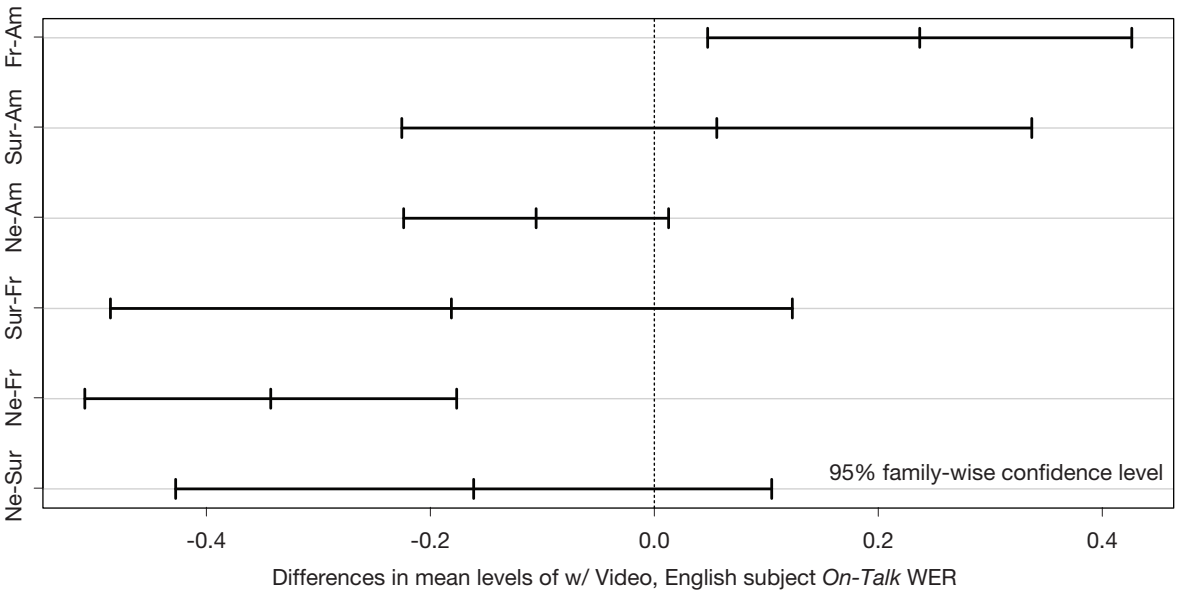


Figure 10.11: *Plot of Tukey HSD test results of "w/ Video–En–On-Talk" WER ANOVA output. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)*

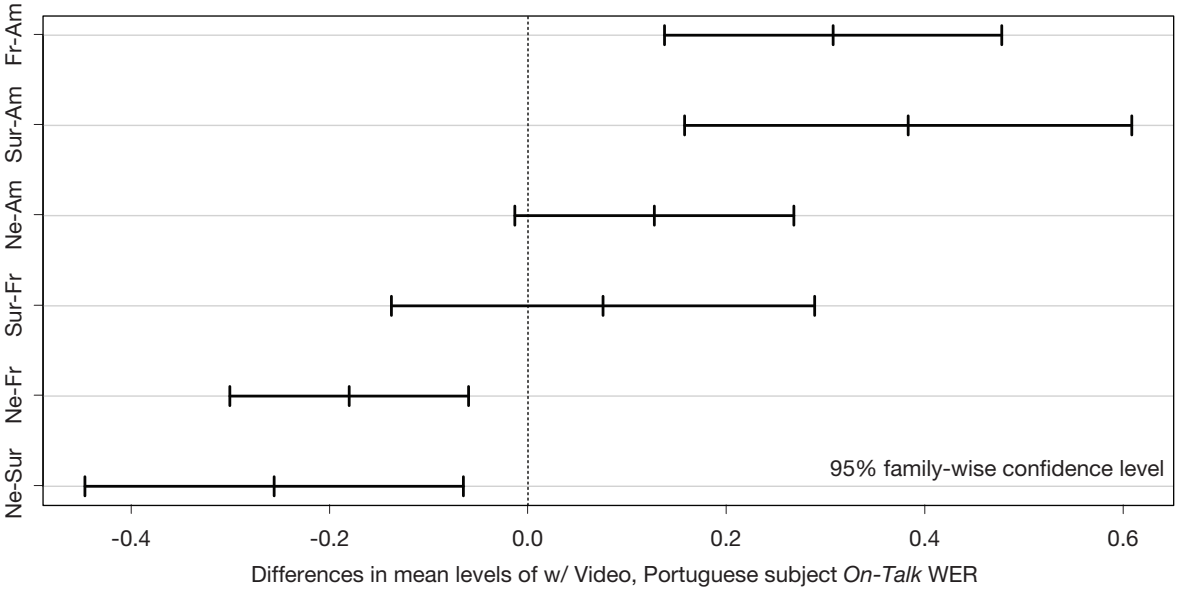


Figure 10.12: *Plot of Tukey HSD test results of "w/ Video–Pt–On-Talk" WER ANOVA output. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)*

Setting – Lang	Role	Per. Emotion	*Mdn*	*M*	*SD*	Count	Emo. (%)	Group (%)	Overall (%)
w/ Video – En	IF	Frustrated	0.875	0.751	0.31	23	11.68%	3.41%	0.98%
w/ Video – En	IF	Amused	0.523	0.523	0.38	50	25.38%	7.42%	2.13%
w/ Video – En	IF	*Surprised*	0.500	0.550	0.34	11	5.58%	1.63%	0.47%
w/ Video – En	IF	Neutral	0.400	0.479	0.55	193	—	28.64%	8.21%
w/ Video – En	IG	Frustrated	1.000	0.789	0.33	32	16.24%	4.75%	1.36%
w/ Video – En	IG	*Surprised*	0.679	0.671	0.32	9	4.57%	1.34%	0.38%
w/ Video – En	IG	Amused	0.571	0.538	0.42	72	36.55%	10.68%	3.06%
w/ Video – En	IG	Neutral	0.250	0.392	0.43	284	—	42.14%	12.07%
w/ Video–En Subtotal — Linked to Per. Em.			0.643	0.604	0.39	197	—	29.23%	8.38%
— *Neutral*			0.286	0.426	0.48	477	—	70.77%	20.28%
w/o Video – En	IF	*Surprised*	0.333	0.333	0.47	2	2.13%	0.33%	0.09%
w/o Video – En	IF	Amused	0.200	0.423	0.80	19	20.21%	3.10%	0.81%
w/o Video – En	IF	Frustrated	0.167	0.650	0.94	17	18.09%	2.78%	0.72%
w/o Video – En	IF	Neutral	0.000	0.449	1.05	213	—	34.80%	9.06%
w/o Video – En	IG	*Surprised*	0.400	0.400	NA	1	1.06%	0.16%	0.04%
w/o Video – En	IG	Amused	0.333	0.352	0.36	27	28.72%	4.41%	1.15%
w/o Video – En	IG	Neutral	0.200	0.324	0.40	305	—	49.84%	12.97%
w/o Video – En	IG	Frustrated	0.000	0.289	0.41	28	29.79%	4.58%	1.19%
w/o Video–En Subtotal — Linked to Per. Em.			0.183	0.398	0.62	94	—	15.36%	4.00%
— *Neutral*			0.148	0.375	0.74	518	—	84.64%	22.02%

Table 10.5: *English subjects' perceived emotions WER*

Setting – Lang	Role	Per. Emotion	*Mdn*	*M*	*SD*	Count	Emo. (%)	Group (%)	Overall (%)
w/ Video – Pt	IF	*Surprised*	0.625	0.645	0.39	19	9.31%	2.95%	0.81%
w/ Video – Pt	IF	Frustrated	0.429	0.480	0.41	59	28.92%	9.16%	2.51%
w/ Video – Pt	IF	Neutral	0.250	0.380	0.41	229	—	35.56%	9.74%
w/ Video – Pt	IF	Amused	0.171	0.303	0.34	26	12.75%	4.04%	1.11%
w/ Video – Pt	IG	*Surprised*	0.550	0.563	0.38	16	7.84%	2.48%	0.68%
w/ Video – Pt	IG	Frustrated	0.500	0.605	0.66	41	20.10%	6.37%	1.74%
w/ Video – Pt	IG	Neutral	0.200	0.324	0.42	211	—	32.76%	8.97%
w/ Video – Pt	IG	Amused	0.000	0.176	0.30	43	21.08%	6.68%	1.83%
w/ Video–Pt Subtotal — Linked to Per. Em.			0.333	0.441	0.47	204	—	31.68%	8.67%
— *Neutral*			0.214	0.351	0.41	440	—	68.32%	18.71%
w/o Video – Pt	IF	*Surprised*	0.625	0.662	0.48	8	7.69%	1.90%	0.34%
w/o Video – Pt	IF	Frustrated	0.297	0.439	0.38	36	34.62%	8.53%	1.53%
w/o Video – Pt	IF	Neutral	0.286	0.387	0.43	138	—	32.70%	5.87%
w/o Video – Pt	IF	Amused	0.154	0.125	0.12	5	4.81%	1.18%	0.21%
w/o Video – Pt	IG	Frustrated	0.343	0.360	0.24	24	23.08%	5.69%	1.02%
w/o Video – Pt	IG	Amused	0.268	0.380	0.38	20	19.23%	4.74%	0.85%
w/o Video – Pt	IG	*Surprised*	0.222	0.492	0.43	11	10.58%	2.61%	0.47%
w/o Video – Pt	IG	Neutral	0.222	0.347	0.70	180	—	42.65%	7.65%
w/o Video–Pt Subtotal — Linked to Per. Em.			0.290	0.417	0.36	104	—	24.64%	4.42%
— *Neutral*			0.236	0.364	0.60	318	—	75.36%	13.52%

Table 10.6: *Portuguese subjects' perceived emotions WER*

10.5 Summary

The analysis of the perceived emotions have provided the following contributions to the understanding of interlingual computer mediated communication.

■ Key Findings.

- The perceived emotions of the subjects are a result of the ASR performance. This statement is based on the fact that a significant difference was observed in the ASR performance when looking at perceived emotions linked to utterances before, but not after, the perceived emotion labels as stated on page 183 under "Direction of the Analysis".
- Of the two setting of w/ Video and w/o Video, only the data from the w/ Video had a significant difference between the ASR performance and the perceived emotions as stated in § 10.3.3.

10.6 Discussion

A point of interest that can be observed from Table 10.5 and Table 10.6, is the different usage of the perceived emotion of *Amused* between English speaking subjects and Portuguese speaking subjects in the w/ Video setting as extracted in Table 10.7.

Setting – Lang	Role	Per. Emotion	*Mdn*	*M*	*SD*	Count	Emo. (%)	Group (%)	Overall (%)
w/ Video – En	IF	Amused	0.523	0.523	0.38	50	25.38%	7.42%	2.13%
		Neutral	0.400	0.479	0.55	193	—	28.64%	8.21%
w/ Video – En	IG	Amused	0.571	0.538	0.42	72	36.55%	10.68%	3.06%
		Neutral	0.250	0.392	0.43	284	—	42.14%	12.07%
w/ Video – Pt	IF	Neutral	0.250	0.380	0.41	229	—	35.56%	9.74%
		Amused	0.171	0.303	0.34	26	12.75%	4.04%	1.11%
w/ Video – Pt	IG	Neutral	0.200	0.324	0.42	211	—	32.76%	8.97%
		Amused	0.000	0.176	0.30	43	21.08%	6.68%	1.83%

Table 10.7: *Subject's perceived emotions WER — w/ Video setting*

A clear indication of this can be seen in Figure 10.13, that looks at the WER of the utterances in the setting of w/ Video. With reference to the utterances identified as *Neutral*, the perceived emotion of *Amused* are linked to utterances with a higher WER mean (M) and median (Mdn) value than that of *Neutral* for the English speaking subjects while for the Portuguese speaking subjects, the WER mean (M) and median (Mdn) values are lower than *Neutral*. Though no significant difference was reported between the perceived emotions of *Amused* and *Neutral* in the Post-hoc comparisons (Tukey HSD test) where the Portuguese result was $p = 0.092$ and the English result was $p = 0.099$,[1] the possibility that there is difference still remains.[2] The point of interest of this observation is that while for the Portuguese speaking subjects, the perceived emotion of *Amused* could be seen as a positive state for the subject to be in, the same cannot be said for the English speaking subjects. Following

[1] If I start fishing for results and use the Reverse F-Score instead of the WER, then a significant difference is reported between *Amused* and *Neutral* for the English data ($p = 0.011$) but not the Portuguese data ($p = 0.080$).

[2] Change the 95% confidence interval to 90% and then these items are also significantly different.

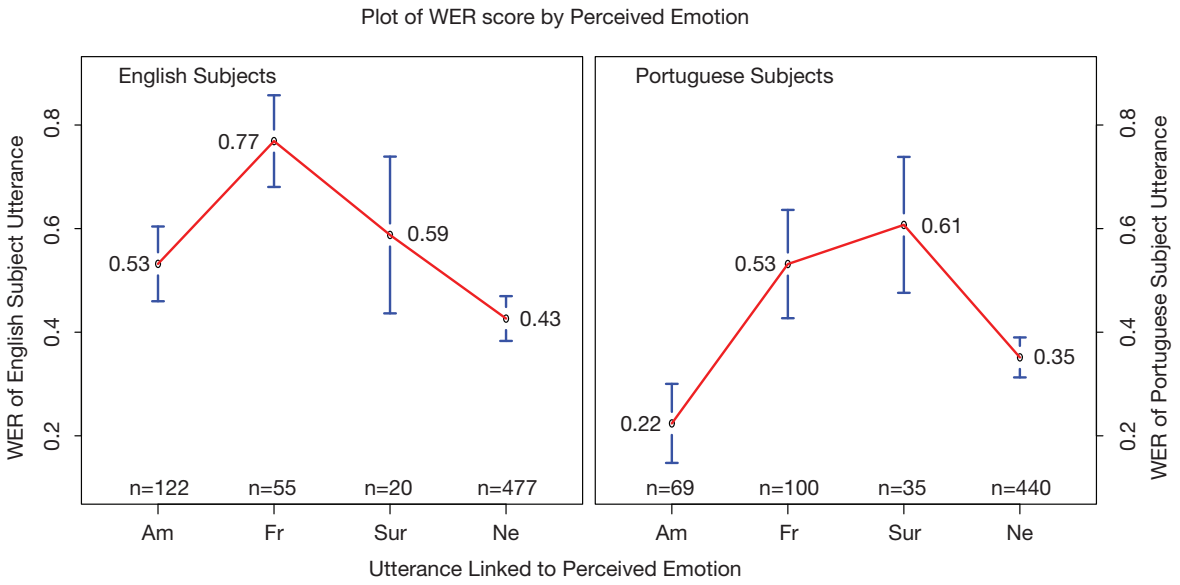


Figure 10.13: *Means plot of "w/ Video–On-Talk" perceived emotion utterance WER. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)*

the high WER of the utterances linked to the perceived emotion of *Amused*, for the English subjects, it is associated with a more serious negative difficulty in communication, indicating that a one-size-fits-all approach to understanding subjects expressing perceived emotions is not recommended. As an

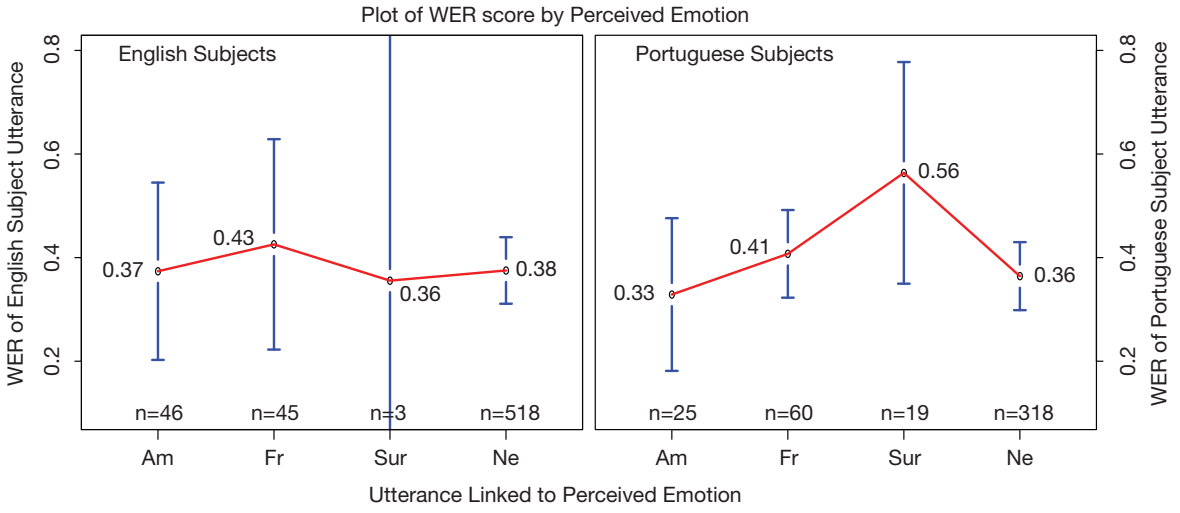


Figure 10.14: *Means plot of "w/o Video–On-Talk" perceived emotion utterance WER. (Ne: Neutral, Am: Amused, Fr: Frustrated, and Sur: Surprised)*

after-note, Figure 10.14 illustrates that within the setting of w/o Video, when the perceived emotion of *Surprised* is ignored due to its low occurrence, both English and Portuguese subjects react in a similar way, which highlights the importance of understanding subject reactions in multiple settings for multiple backgrounds.[3]

[3]This difference may not even be a Portuguese trait, but that of a subject that is not in their home country — in other words a visitor to country that the experiment was performed, and therefore, as a visitor, on better behaviour.

Chapter 11

User Survey Study

Overall, three types of analysis were performed on the data of the computer-mediated dialogue dataset: (i) analysis of the dataset data to see what the data says about the subject's, (ii) a comparison of the dataset data with data from another dataset to see how the data compares with direct Human-to-Human communication, and (iii) analysis of the user survey to try and understand what the subject's of the data collection thought. In this chapter I report the analysis of the user survey.

After the subjects finished the map task dialogue recordings, the subjects were asked to complete a user survey of fifteen statements replied using the 7 point Likert scale format and five questions replied in an open-text-field format. The analysis reported here is based upon the fifteen statement replies and provides an idea of how the experience of communicating a task oriented dialogue differed between the setting of the communication (with or without video — visual information), role of the subject (Information Giver or Information Follower), the language spoken by the subject, the gender and the map used by the subject. Some of the results of this study have already been published as an LNCS invited paper (Cerrato, Hayakawa, Campbell, & Luz, 2016) and a IEEE conference paper (Hayakawa et al., 2017a).

11.1 Introduction

Be careful of what you wish for.

The promise that the modality of a visual channel would improve communication is a delicate matter. Since the introduction of the Picturephone by the American Telephone and Telegraph Company (AT&T) in 1964, that transmitted the user's image and voice simultaneously over telephone lines (Lipartito, 2003), industry has continued to add an image channel to the conversation modality, even though the Picturephone can be categorised as a complete commercial failure. Mobile phones that were capable of sending video images with the audio channel were first introduced in 1999, and today, it is more difficult to find a mobile phone that doesn't have this function. In a study of audio only telephone communication, the fact that visual information though a video channel can be

both a positive and negative attribute to communication has already been presented by Morikawa & Maesako (1998, p. 149) in their reference to the publication in Japanese of Yoshida and Kakuta (refer to page 23 for the full quotation). Morikawa & Maesako go on to indicate that these reasonings — the advantages of "speed, available anytime, available anywhere, no visual information about the other end, and easier way to say what we want than in face-to-face conversation" and the disadvantages of "no visual information about the other end and difficulty in conveying subtle emotion" — indicate that the subjects of this study consider communication over the telephone as a different alternative to Face-to-Face communication with its own communication style. Not only the consumer market, but also academia and industry are commonly equipped with video-conferencing devices as a means to communicate over distance. From dedicated devices from Polycom Inc. to Skype of Microsoft Inc., image/video is now an easily accessible option for communication. Microsoft has even released a Machine Translation function called Skype Translator (W. D. Lewis, 2015) that translates up to 10 languages in Speech-to-Speech Machine Translation (audio and video calls) and over 50 languages in Text-to-Text form (instant messaging).

It is therefore inevitable that machine translation will be swiftly added to this modality. With the addition of Machine Translation, this will now create a conversation style that few users have adequate exposure and little research has been conducted to date (Hara & Iqbal, 2015).

It has already been well established that conversation style changes with the environment. High latency video conference systems produce fewer, but longer utterances between participants, with lower latency systems providing conversations styles that are closer to Face-to-Face communication, shorter but more frequent utterances (O'Conaill et al., 1993). It has also been mentioned that task effectiveness is mainly determined by the quality (in terms of low latency) of the audio channel, thus implying that video is of marginal importance (Finn, Sellen, & Wilbur, 1997). A previous study into video enhanced communication during medical meetings has reported that the participants of video-conferencing benefit from the video channel, but these benefits were only recognised after continuous use of the system (Kane & Luz, 2006a). The long term study by Kane & Luz reported that the medical professionals only changed their perception of the video communication after eight months as mentioned in the quoted Table 11.1 from Kane & Luz (2006a, p. 519).

Requirement	Strongly Agree	Agree	No Opinion	Disagree	Strongly Disagree
See all Radiology	74%	11%	16%	0	0
See all Pathology	58%	26%	11%	5%	0
Hear all Participants	21%	32%	42%	5%	0
See all Participants '04	5%	21%	53%	21%	0
See all Participants '05	19%	50%	19%	8%	4%

Table 11.1: *Participant's requirements expressed* prior *to teleconferencing experience and revised requirement for video* after *8 months experience.*

In this study, I look at how the addition of Machine Translation affects the user of a Speech-to-Speech Machine Translation (S2S-MT) system, with and without a video channel (w/ Video & w/o Video). The study of the computer-mediated dialogue dataset (the ILMT-s2s corpus) user survey

concludes that the subjects of the ILMT-s2s System perceived the system more favourably when they could not see their interlocutor (w/o Video), where given the role of Information Giver, and where speaking the Portuguese language. In other words, firstly, subjects that could see each other through the live video stream setting (w/ Video) and then those with the role of following the provided instructions (Information Follower) did not score the ILMT-s2s System favourably.

11.2 Dataset Used for the Analysis

In this chapter, the subject survey data described in § 3.3.4 of Chapter 3 is used for the analysis.

11.3 Methods of the Analysis

Using the replies to the subject survey (§ 4.3.6), a qualitative study is performed to identify patterns of preference towards the communication tool. This verification is done by comparing the 7 point Likert scale (evenly spaced from 1 = Strongly disagree to 7 = Strongly agree) within the individual groupings and also within combined groupings.

11.3.1 Visualisation of the 7 Point Likert Scale Results

As with the speech rate and the ASR performance, the results to the fifteen subject survey statements are visualized. This is done by using stacked bar charts and density plot to visualise the main pattern of the results and the distribution of the results respectively.

11.3.2 Analysis of the 7 Point Likert Scale Results

To obtain a general understanding of the results, the median and mean values are calculated for the fifteen 7 point Likert scale statements. Although median is recommended to be used to analyse Likert scores (Kaptein & Robertson, 2012), the means were also used to provide more meaning to the median values: By looking at the distance between the median and means value, it is easier to estimate how the results are distributed.

11.3.3 Analysis of Results Divided into Groups

To identify what might be causing differing preferences, the 7 point Likert scale results are not only studied as a whole, but also by individual groupings to verify if the differing role, setting gender, map or language spoken alone provided any difference, but also the combination of groupings too.

The 7 point Likert scale results of *[Refer to grouping list below]* are tested to verify if there is a significant difference (*Wilcox*).

Grouping list:

1. Setting (w/ Video or w/o Video)
2. Role (Information Giver or Information Follower)
3. Gender (female or male)
4. Language Spoken (English or Portuguese)
5. Map (Map 01 or Map 07)
6. Setting & Role
7. Setting, Role & Language

11.4 Results of the Analysis

11.4.1 Visualisation of the 7 Point Likert Scale Results

The stacked bar charts and density plots of the subject survey results is illustrated in Figure 11.1. From the density plots of Figure 11.1 it is possible to see a multimodal distributions in the reply to some of the statements.

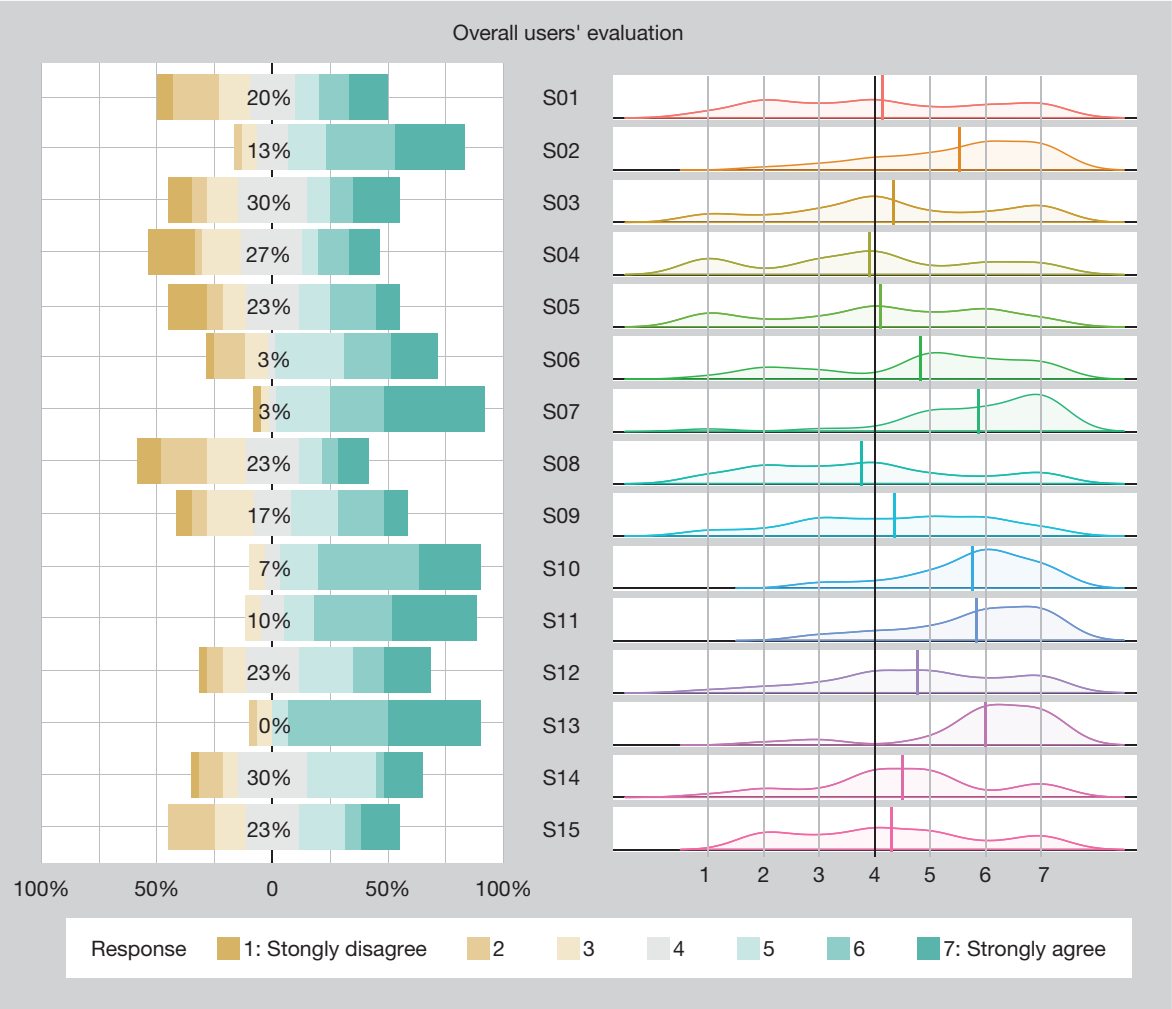


Figure 11.1: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale replies*

■ For the Subject Survey Charts Separated by Grouping, Refer to the Following Figures in Appendix A.

Figure A.23: Results differentiated by setting.

Figure A.24: Results differentiated by the role of the subject.

Figure A.25: Results differentiated by the language spoken by the subject.

Figure A.26: Results differentiated by the gender of the subject.

Figure A.27: Results differentiated by the map used by the subject.

Figure A.28: Results differentiated by the setting and role of the subject — Stacked bar charts.

Figure A.29: Results differentiated by the setting and role of the subject — Density plots.

11.4.2 Analysis of Results as a Whole — One Group

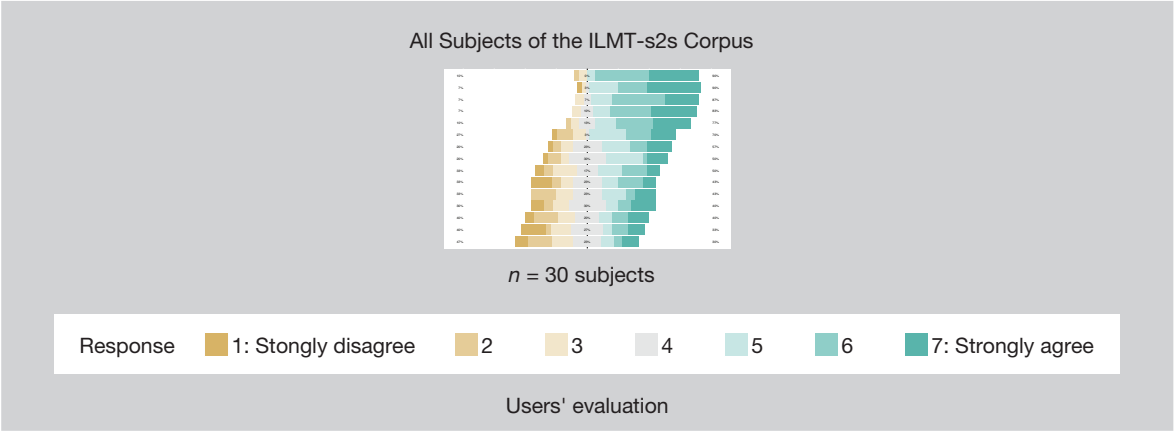


Figure 11.2: *Stacked bar chart of subject survey Likert scale replies — All Subjects*

■ Summary of All the Data as a Whole Using Median Values.

The overview of the thirty replies to the subject survey indicate that all statements return a median value of a neutral four or better than neutral as indicated in the Table 11.2.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	4.0	4.5	6.0	6.0	5.0	6.0	4.5	4.0

Table 11.2: *Median value per statement of subject survey Likert results*

■ Summary of All the Data as a Whole Using Median, Means and Standard Deviation Values.

To provide more information about the distribution of the replies, the mean (M) and standard deviation (SD) are calculated as indicated in the Table 11.3.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	4.0	4.5	6.0	6.0	5.0	6.0	4.5	4.0
M	4.13	5.53	4.33	3.90	4.10	4.83	5.87	3.77	4.37	5.77	5.83	4.77	6.00	4.50	4.30
SD	1.94	1.41	1.90	2.01	1.95	1.80	1.41	1.87	1.71	1.14	1.23	1.65	1.29	1.59	1.70

Table 11.3: *Median, mean and standard deviation value per statement of subject survey Likert results*

11.4.2.1 Key Results

The overview of the thirty replies to the subject survey indicate that all statements return a median value of a neutral four or better than neutral as indicated in the Table 11.4.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	4.0	4.5	6.0	6.0	5.0	6.0	4.5	4.0

Table 11.4: *Median value per statement of subject survey Likert results*

From this result, it is possible to say the following:

- Subjects rated the communication tool (the ILMT-s2s System; §4.2) as easy to use, with a pleasant and likeable user interface, with the expected functions, and a satisfying Text-to-Speech (TTS) output voice.

This can been seen in the five statements with a median value of 6.0, and also from the statements with a median value of 5.0 as listed below.

Statements with a median value of 6.0:

S.02: It was simple to use this system.

S.07: It was easy to learn to use this system.

S.10: The interface of the system was pleasant.

S.11: I liked using the interface of this system.

S.13: I was satisfied with the voice of this system.

Statements with a median value of 5.0:

S.06: I felt comfortable using this system.

S.12: This system has all the functions and capabilities I expected it to have.

The statements with a median value of 4.0 do not tell much about the subject's perception, since this is not a 3 point Likert scale. To provide more information about the distribution of the replies, the mean (M) and standard deviation (SD) are calculated as indicated in the Table 11.5.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	4.0	4.5	6.0	6.0	5.0	6.0	4.5	4.0
M	4.13	5.53	4.33	3.90	4.10	4.83	5.87	3.77	4.37	5.77	5.83	4.77	6.00	4.50	4.30
SD	1.94	1.41	1.90	2.01	1.95	1.80	1.41	1.87	1.71	1.14	1.23	1.65	1.29	1.59	1.70

Table 11.5: *Median, mean and standard deviation value per statement of subject survey Likert results*

Looking at the mean (M) values of Table 11.4, the negative results for statements "S.04: I was able to complete the task quickly using this system." and "S.08: I believe I could become productive quickly using this system." of 3.90 and 3.77 respectively, indicate that stronger negative feedback is provided since the median value is a neutral 4.0 and a 7 point Likert scale is used.

■ Areas of Concern with the Initial Results.

To visualise the distribution of the subject survey results, the values are plotted out.

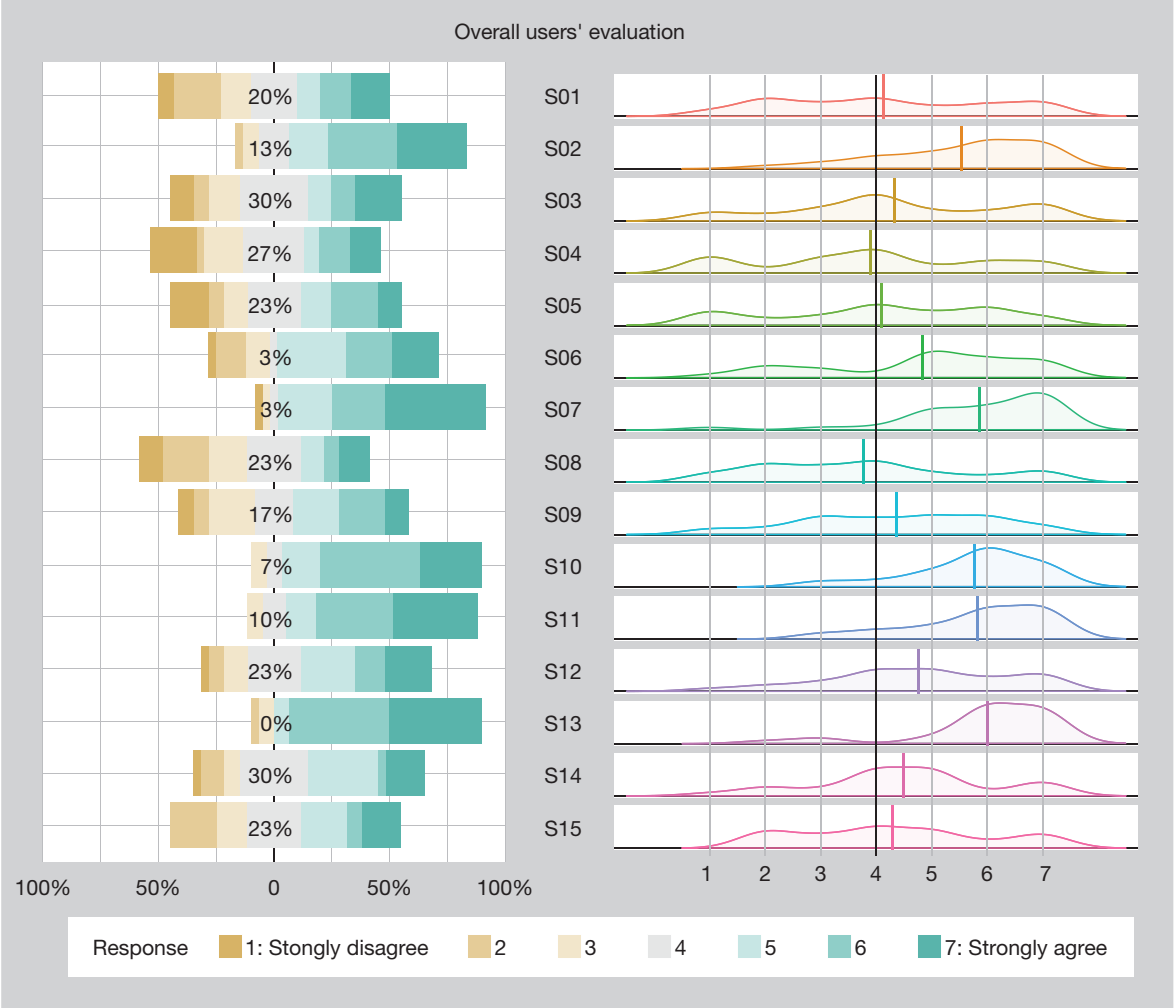


Figure 11.3: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale replies*

From the density plots of Figure 11.3 it is possible to see that all statements with a median value of 4.0 show a multimodal distributions.

Statements with a median value of 4.0:

S.01: Overall, I am satisfied with how easy it was to use this system.

S.03: I could effectively complete the tasks using this system.

S.04: I was able to complete the task quickly using this system.

S.05: I was able to efficiently complete the task using this system.

S.08: I believe I could become productive quickly using this system.

S.15: Overall, I am satisfied with this system.

This multimodal distribution indicates a need to look in more detail at the groupings of the subjects.

11.4.3 Analysis of Results Divided into Groups

Following the mixture in the curvature of the density plot in Figure 11.3 explained in the previous § 11.4.2, analysis was performed to investigate the groups of the subjects. Graphical representation of the data, summarised in Figure 11.4 and represented in more detail in Figure A.23 to Figure A.27 of Appendix A, shows that the subjects from the setting of w/o Video, the role of Information Giver, and the language of Portuguese speakers give a more positive interpretation of the communication tool (the ILMT-s2s System), however the difference between the gender or the map are not so clear.

Figure 11.4: *Stacked barchart of survey Likert scale replies — Setting, role, language, gender, & map*

The survey results are analysed for the following 7 groupings to see if there is a significant difference in the Likert score results using the Mann-Whitney-Wilcoxon test (Mann & Whitney, 1947) and also if there is a noticeable effect size difference in the Likert score results using the Cliff's δ effect size

(Cliff, 1993).[1]

Grouping list:

1. Setting (w/ Video or w/o Video)
2. Role (Information Giver or Information Follower)
3. Gender (female or male)
4. Language Spoken (English or Portuguese)
5. Map (Map 01 or Map 07)
6. Setting & Role
7. Setting, Role & Language

■ Significance Test of the Survey Results by Grouping — Setting.

The distribution of the survey results when grouped by setting are presented in Table 11.6.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
w/ Video															
Mdn	3.5	5.5	4.0	3.0	3.5	5.0	6.0	3.0	4.0	6.0	5.5	4.0	6.0	4.0	4.0
M	3.56	5.13	3.88	3.13	3.38	4.31	5.44	2.75	3.75	5.44	5.31	4.19	6.13	3.69	3.69
SD	1.79	1.50	1.89	1.36	1.63	1.74	1.63	1.13	1.44	1.31	1.40	1.64	1.02	1.49	1.40
w/o Video															
Mdn	5.0	6.0	4.5	5.5	6.0	6.0	7.0	5.0	5.5	6.0	6.5	5.5	6.0	5.0	5.0
M	4.79	6.00	4.86	4.79	4.93	5.43	6.36	4.93	5.07	6.14	6.43	5.43	5.86	5.43	5.00
SD	1.97	1.18	1.83	2.30	2.02	1.74	0.93	1.90	1.77	0.77	0.65	1.45	1.56	1.16	1.80

Table 11.6: *Median, mean and standard deviation value per statement of subject survey Likert results*

For the setting difference between w/o Video and w/ Video, eight statements resulted in a significant difference being reported. As listed in Table 11.7, significant differences were observed in statements: S04, S05, S08, S09, S11, S12, S14, and S15. Medium and large effect size were observed in statements: S01, S02, S04, S05, S06, S07, S08, S09, S11, S12, S14, and S15. Refer to Table A.17 in Appendix A for the results for all statements.

w/o Video : w/ Video — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.08	5.0 : 3.0	4.93 : 2.75	1.90 : 1.13	$p < 0.0018$	**	$\delta = 0.6384$	large
2	S.14	5.0 : 4.0	5.43 : 3.69	1.16 : 1.49	$p < 0.0019$	**	$\delta = 0.6518$	large
3	S.04	5.5 : 3.0	4.79 : 3.13	2.29 : 1.36	$p < 0.0168$	*	$\delta = 0.4911$	large
4	S.05	6.0 : 3.5	4.93 : 3.38	2.02 : 1.36	$p < 0.0227$	*	$\delta = 0.4732$	medium
5	S.11	6.5 : 5.5	6.43 : 5.31	0.65 : 1.40	$p < 0.0235$	*	$\delta = 0.4688$	medium
6	S.09	5.5 : 4.0	5.07 : 3.75	1.77 : 1.44	$p < 0.0268$	*	$\delta = 0.4598$	medium
7	S.12	5.5 : 4.0	5.43 : 4.19	1.45 : 1.64	$p < 0.0325$	*	$\delta = 0.4554$	medium
8	S.15	5.0 : 4.0	5.00 : 3.69	1.80 : 1.40	$p < 0.0401$	*	$\delta = 0.4375$	medium
9	S.01	5.0 : 3.5	4.79 : 3.56	1.97 : 1.79	$p < 0.0962$	.	$\delta = 0.3571$	medium
10	S.02	6.0 : 5.5	6.00 : 5.13	1.18 : 1.50	$p < 0.0944$	.	$\delta = 0.3527$	medium
11	S.06	6.0 : 5.0	5.43 : 4.31	1.74 : 1.74	$p < 0.0706$	.	$\delta = 0.3839$	medium
12	S.07	7.0 : 6.0	6.36 : 5.44	0.93 : 1.63	$p < 0.0750$	.	$\delta = 0.3661$	medium

Table 11.7: *Significance test and effect size between Likert scale results of "w/o Video : w/ Video"*

■ Significance Test of the Survey Results by Grouping — Role.

The distribution of the survey results when grouped by role are presented in Table 11.8.

For the role difference between Information Giver and Information Follower, two statements resulted in a significant difference being reported. As listed in Table 11.9, significant differences were

[1] Refer to § A.6.3 of Appendix A for results from the Kruskal–Wallis test, t-test, and Cohen's d and also the comparison of the Likert survey results when divided by gender and the map that was used.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Information Follower															
Mdn	3.0	6.0	4.0	4.0	4.0	5.0	6.0	3.0	4.0	6.0	6.0	4.0	6.0	4.0	3.0
M	3.53	5.27	3.67	3.60	3.47	4.27	5.53	3.07	4.00	5.53	5.53	4.33	5.67	4.07	3.67
SD	1.92	1.71	1.72	1.96	2.13	1.87	1.64	1.83	1.69	1.36	1.46	1.76	1.63	1.71	1.71
Information Giver															
Mdn	5.0	6.0	5.0	4.0	4.0	6.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0
M	4.73	5.80	5.00	4.20	4.73	5.40	6.20	4.47	4.73	6.00	6.13	5.20	6.33	4.93	4.93
SD	1.83	1.01	1.89	2.08	1.58	1.59	1.08	1.68	1.71	0.85	0.92	1.47	0.72	1.39	1.49

Table 11.8: *Median, mean and standard deviation value per statement of subject survey Likert results*

observed in statements: S08, and S15. Medium and large effect size were observed in statements: S01, S03, S06, S08, and S15. Refer to Table A.18 in Appendix A for the results for all statements.

Information Giver : Information Follower — Survey result comparison							
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.08	4.0 : 3.0	4.47 : 3.07	1.68 : 1.83	$p < 0.0318$	*	$\delta = 0.4578$	medium
2	S.15	5.0 : 3.0	4.93 : 3.67	1.49 : 1.72	$p < 0.0348$	*	$\delta = 0.4489$	medium
3	S.03	5.0 : 4.0	5.00 : 3.67	1.89 : 1.72	$p < 0.0656$	.	$\delta = 0.3911$	medium
4	S.01	5.0 : 3.0	4.73 : 3.53	1.83 : 1.92	$p < 0.0849$	.	$\delta = 0.3689$	medium
5	S.06	6.0 : 5.0	5.40 : 4.27	1.59 : 1.87	$p < 0.0895$	.	$\delta = 0.3600$	medium

Table 11.9: *Significance test and effect size between Likert scale results of "Information Giver (IG) : Information Follower (IF)"*

■ Significance Test of the Survey Results by Grouping — Language.

The distribution of the survey results when grouped by language are presented in Table 11.10.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
English															
Mdn	4.0	5.0	4.0	4.0	4.0	5.0	5.0	4.0	5.0	6.0	6.0	5.0	6.0	4.0	4.0
M	3.67	5.40	3.80	3.60	3.53	4.53	5.27	3.47	4.20	6.00	5.80	5.00	5.93	3.87	3.93
SD	1.76	1.18	1.78	2.06	1.92	2.03	1.49	1.85	1.61	0.65	1.01	1.65	1.22	1.64	1.58
Portuguese															
Mdn	5.0	6.0	5.0	4.0	5.0	5.0	7.0	4.0	4.0	6.0	6.0	5.0	7.0	5.0	4.0
M	4.60	5.67	4.87	4.20	4.67	5.13	6.47	4.07	4.53	5.53	5.87	4.53	6.07	5.13	4.67
SD	2.06	1.63	1.92	1.97	1.88	1.55	1.06	1.91	1.85	1.46	1.46	1.68	1.39	1.30	1.80

Table 11.10: *Median, mean and standard deviation value per statement of subject survey Likert results*

For the language difference between Portuguese and English, two statements resulted in a significant difference being reported. As listed in Table 11.11, significant differences were observed in statements: S07, and S14. Medium and large effect size were observed in statements: S05, S07, and S14. Refer to Table A.19 in Appendix A for the results for all statements.

Portuguese speaking subject (Pt) : English speaking subject (En) — Survey result comparison							
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.07	7.0 : 5.0	6.47 : 5.27	1.06 : 1.49	$p < 0.0044$	**	$\delta = 0.5822$	large
2	S.14	5.0 : 4.0	5.13 : 3.87	1.30 : 1.64	$p < 0.0381$	*	$\delta = 0.4356$	medium
3	S.05	5.0 : 4.0	4.67 : 3.53	1.88 : 1.92	$p < 0.1051$	–	$\delta = 0.3467$	medium

Table 11.11: *Significance test and effect size between Likert scale results of "Portuguese speaking subject (Pt) : English speaking subject (En)"*

■ Significance Test of the Survey Results by Grouping — Gender.

The distribution of the survey results when grouped by gender are presented in Table 11.12.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Female															
Mdn	4.5	6.0	4.0	4.0	4.5	5.0	7.0	4.0	3.5	6.0	6.0	4.5	6.0	5.0	4.0
M	4.1	5.4	4.4	3.8	4.3	4.6	6.1	3.9	4.3	5.8	5.9	4.4	5.7	4.8	4.3
Male															
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	3.5	5.0	6.0	6.0	5.0	6.0	4.0	4.5
M	4.1	5.6	4.3	4.0	3.9	5.1	5.6	5.6	4.5	5.7	5.8	5.2	6.4	4.1	4.4

Table 11.12: *Median and mean value per statement of subject survey Likert results*

For the gender difference between female and male, no significant differences or medium and large effect size were observed in any of the statements. Refer to Table A.20 in Appendix A for the results for all statements.

■ Significance Test of the Survey Results by Grouping — Map.

The distribution of the survey results when grouped by map are presented in Table 11.13.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Map 01															
Mdn	3.5	6.0	4.5	3.0	4.0	5.0	6.5	4.0	4.0	6.0	6.0	5.0	6.0	4.5	3.5
M	3.9	5.7	4.4	3.5	3.9	4.6	6.2	3.9	4.0	5.9	6.1	4.6	5.8	4.6	4.1
Map 07															
Mdn	4.0	6.0	4.0	4.0	4.0	5.0	6.0	3.5	5.0	6.0	6.0	5.5	6.0	4.5	4.0
M	4.3	5.4	4.3	4.3	4.3	5.0	5.6	3.7	4.7	5.7	5.6	4.9	6.2	4.4	4.5

Table 11.13: *Median and mean value per statement of subject survey Likert results*

For the map difference between Map 01 and Map 07, no significant differences or medium and large effect size were observed in any of the statements. Refer to Table A.21 in Appendix A for the results for all statements.

■ Significance Test of the Survey Results by Grouping — Setting and Role.

The distribution of the survey results when grouped by setting and role are presented in Table 11.14.

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
w/ Video — Information Follower															
Mdn	2.5	4.5	3.5	3.5	2.5	3.0	6.0	2.0	3.0	5.5	4.5	3.5	6.0	3.0	2.5
M	3.00	4.75	3.50	3.00	2.75	3.38	4.88	2.13	3.00	5.00	4.88	3.50	5.88	2.88	2.75
SD	1.85	1.83	2.00	1.31	1.75	1.51	1.89	0.83	1.31	1.51	1.64	1.77	1.25	1.13	0.89
w/ Video — Information Giver															
Mdn	4.0	6.0	4.0	3.0	4.0	5.5	6.5	3.5	4.5	6.0	6.0	5.0	6.5	4.5	4.5
M	4.13	5.50	4.25	3.25	4.00	5.25	6.00	3.38	4.50	5.88	5.75	4.88	6.38	4.50	4.63
SD	1.64	1.07	1.83	1.49	1.31	1.49	1.20	1.06	1.20	0.99	1.04	1.25	0.74	1.41	1.19
w/o Video — Information Follower															
Mdn	4.0	6.0	4.0	5.0	6.0	6.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0
M	4.14	5.86	3.86	4.29	4.29	5.29	6.29	4.14	5.14	6.14	6.29	5.29	5.43	5.43	4.71
SD	1.95	1.46	1.46	2.43	2.36	1.80	0.95	2.12	1.35	0.90	0.76	1.25	2.07	1.13	1.89
w/o Video — Information Giver															
Mdn	6.0	6.0	7.0	6.0	6.0	6.0	7.0	6.0	6.0	6.0	7.0	6.0	6.0	5.0	6.0
M	5.43	6.14	5.86	5.29	5.57	5.57	6.43	5.71	5.00	6.14	6.57	5.57	6.29	5.43	5.29
SD	1.90	0.90	1.68	2.21	1.51	1.81	0.98	1.38	2.24	0.69	0.53	1.72	0.76	1.27	1.80

Table 11.14: *Median, mean and standard deviation value per statement of subject survey Likert results*

Six sets of significance test are performed to identify the patterns in all the combinations of the groupings between setting and role.

(i) For the combination of "w/ Video & IF" with "w/o Video & IG", as listed in Table 11.15, significant differences were observed in nine statements: S01, S03, S04, S05, S06, S07, S08, S14, S15. Medium and large effect size were observed in fourteen statements: S01, S02, S03, S04, S05, S06, S07, S08, S09, S10, S11, S12, S14, S15.
Refer to Table A.22 in Appendix A for the results for all statements.

(ii) For the combination of "w/ Video & IF" with "w/o Video & IF", as listed in Table 11.16, significant differences were observed in four statements: S09, S12, S14, S15. Medium and large effect size were observed in thirteen statements: S01, S02, S04, S05, S06, S07, S08, S09, S10, S11, S12, S14, S15.
Refer to Table A.23 in Appendix A for the results for all statements.

(iii) For the combination of "w/ Video & IF" with "w/ Video & IG", as listed in Table 11.17, significant differences were observed in five statements: S06, S08, S09, S14, S15. Medium and large effect size were observed in ten statements: S01, S05, S06, S07, S08, S09, S10, S12, S14, S15.
Refer to Table A.24 in Appendix A for the results for all statements.

(iv) For the combination of "w/o Video & IG" with "w/ Video & IG", as listed in Table 11.18, significant differences were observed in one statements: S08. Medium and large effect size were observed in nine statements: S01, S02, S03, S04, S05, S08, S11, S12, S14.
Refer to Table A.25 in Appendix A for the results for all statements.

(v) For the combination of "w/o Video & IG" with "w/o Video & IF", as listed in Table 11.19, significant differences were observed in none of the statements. Medium and large effect size were observed in four statements: S01, S03, S05, S08.
Refer to Table A.26 in Appendix A for the results for all statements.

(vi) For the combination of "w/o Video & IF" with "w/ Video & IG", as listed in Table 11.20, significant differences were observed in none of the statements. A medium effect size were observed in one statement: S14.
Refer to Table A.27 in Appendix A for the results for all statements.

w/ Video – Information Follower : w/o Video – Information Giver — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.08	2.0 : 6.0	2.13 : 5.71	0.83 : 1.38	$p < 0.0013$	**	$\delta = 1.0000$	large
2	S.14	3.0 : 5.0	2.88 : 5.43	1.13 : 1.27	$p < 0.0038$	**	$\delta = 0.8929$	large
3	S.15	2.5 : 6.0	2.75 : 5.29	0.89 : 1.80	$p < 0.0150$	*	$\delta = 0.7500$	large
4	S.05	2.5 : 6.0	2.75 : 5.57	1.75 : 1.51	$p < 0.0185$	*	$\delta = 0.7321$	large
5	S.04	3.5 : 6.0	3.00 : 5.29	1.31 : 2.21	$p < 0.0279$	*	$\delta = 0.6786$	large
6	S.06	3.0 : 6.0	3.38 : 5.57	1.51 : 1.81	$p < 0.0324$	*	$\delta = 0.6607$	large
7	S.01	2.5 : 6.0	3.00 : 5.43	1.85 : 1.90	$p < 0.0447$	*	$\delta = 0.6250$	large
8	S.07	6.0 : 7.0	4.88 : 6.43	1.89 : 0.98	$p < 0.0467$	*	$\delta = 0.6071$	large
9	S.03	3.5 : 7.0	3.50 : 5.86	2.00 : 1.68	$p < 0.0499$	*	$\delta = 0.6071$	large
10	S.11	4.5 : 7.0	4.88 : 6.57	1.64 : 0.53	$p < 0.0530$	.	$\delta = 0.5893$	large
11	S.12	3.5 : 6.0	3.50 : 5.57	1.77 : 1.72	$p < 0.0608$	.	$\delta = 0.5893$	large
12	S.09	3.0 : 6.0	3.00 : 5.00	1.31 : 2.24	$p < 0.0703$	.	$\delta = 0.5714$	large
13	S.02	4.5 : 6.0	4.75 : 6.14	1.83 : 0.90	$p < 0.1374$	–	$\delta = 0.4643$	medium
14	S.10	5.5 : 6.0	5.00 : 6.14	1.51 : 0.69	$p < 0.1408$	–	$\delta = 0.4464$	medium

Table 11.15: *Significance test and effect size between Likert scale results of "w/ Video – IF : w/o Video – IG" — Part 1 of 6*

w/ Video – Information Follower : w/o Video – Information Follower — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.14	3.0 : 5.0	2.88 : 5.43	1.13 : 1.13	$p < 0.0021$	**	$\delta = 0.8393$	large
2	S.09	3.0 : 5.0	3.00 : 5.14	1.31 : 1.35	$p < 0.0163$	*	$\delta = 0.6964$	large
3	S.12	3.5 : 5.0	3.50 : 5.29	1.77 : 1.25	$p < 0.0279$	*	$\delta = 0.6786$	large
4	S.15	2.5 : 5.0	2.75 : 4.71	0.89 : 1.89	$p < 0.0374$	*	$\delta = 0.5714$	large
5	S.06	3.0 : 6.0	3.38 : 5.29	1.51 : 1.80	$p < 0.0525$	.	$\delta = 0.5357$	large
6	S.08	2.0 : 4.0	2.13 : 4.14	0.83 : 2.12	$p < 0.0596$	.	$\delta = 0.5000$	large
7	S.07	6.0 : 7.0	4.88 : 6.29	1.89 : 0.95	$p < 0.0899$	.	$\delta = 0.4464$	medium
8	S.11	4.5 : 6.0	4.88 : 6.29	1.64 : 0.76	$p < 0.1073$	–	$\delta = 0.5000$	large
9	S.10	5.5 : 6.0	5.00 : 6.14	1.51 : 0.90	$p < 0.1515$	–	$\delta = 0.4464$	medium
10	S.05	2.5 : 6.0	2.75 : 4.29	1.75 : 2.36	$p < 0.1527$	–	$\delta = 0.3750$	medium
11	S.04	3.5 : 5.0	3.00 : 4.29	1.31 : 2.43	$p < 0.1705$	–	$\delta = 0.3571$	medium
12	S.02	4.5 : 6.0	4.75 : 5.86	1.83 : 1.46	$p < 0.2598$	–	$\delta = 0.3571$	medium
13	S.01	2.5 : 4.0	3.00 : 4.14	1.85 : 1.95	$p < 0.2602$	–	$\delta = 0.3571$	medium

Table 11.16: *Significance test and effect size between Likert scale results of "w/ Video : w/o Video of IF" — Part 2 of 6*

w/ Video – Information Follower : w/ Video – Information Giver — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.15	2.5 : 4.5	2.75 : 4.63	0.89 : 1.19	$p < 0.0058$	**	$\delta = 0.8125$	large
2	S.14	3.0 : 4.5	2.88 : 4.50	1.13 : 1.41	$p < 0.0226$	*	$\delta = 0.6719$	large
3	S.06	3.0 : 5.5	3.38 : 5.25	1.51 : 1.49	$p < 0.0260$	*	$\delta = 0.6563$	large
4	S.08	2.0 : 3.5	2.13 : 3.38	0.83 : 1.06	$p < 0.0341$	*	$\delta = 0.6250$	large
5	S.09	3.0 : 4.5	3.00 : 4.50	1.31 : 1.20	$p < 0.0476$	*	$\delta = 0.5938$	large
6	S.12	3.5 : 5.0	3.50 : 4.88	1.77 : 1.25	$p < 0.0681$	.	$\delta = 0.5469$	large
7	S.01	2.5 : 4.0	3.00 : 4.13	1.85 : 1.64	$p < 0.1501$	–	$\delta = 0.4531$	medium
8	S.05	2.5 : 4.0	2.75 : 4.00	1.75 : 1.31	$p < 0.1648$	–	$\delta = 0.4219$	medium
9	S.07	6.0 : 6.5	4.88 : 6.00	1.89 : 1.20	$p < 0.1908$	–	$\delta = 0.3906$	medium
10	S.10	5.5 : 6.0	5.00 : 5.88	1.31 : 0.99	$p < 0.2470$	–	$\delta = 0.3438$	medium

Table 11.17: *Significance test and effect size between Likert scale results of "IF : IG of w/ Video" — Part 3 of 6*

w/o Video – Information Giver : w/ Video – Information Giver — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.08	6.0 : 3.5	5.71 : 3.38	1.38 : 1.06	$p < 0.0091$	**	$\delta = 0.7321$	large
2	S.04	6.0 : 3.0	5.29 : 3.25	2.21 : 1.49	$p < 0.0531$	.	$\delta = 0.5536$	large
3	S.05	6.0 : 4.0	5.57 : 4.00	1.51 : 1.31	$p < 0.0752$	.	$\delta = 0.5000$	large
4	S.11	7.0 : 6.0	6.57 : 5.75	0.53 : 1.04	$p < 0.1074$	–	$\delta = 0.4821$	large
5	S.03	7.0 : 4.0	5.86 : 4.25	1.68 : 1.83	$p < 0.1213$	–	$\delta = 0.4821$	large
6	S.01	6.0 : 4.0	5.43 : 4.13	1.90 : 1.64	$p < 0.1411$	–	$\delta = 0.4107$	medium
7	S.12	6.0 : 5.0	5.57 : 4.88	1.72 : 1.25	$p < 0.2138$	–	$\delta = 0.3929$	medium
8	S.14	5.0 : 4.5	5.43 : 4.50	1.27 : 1.41	$p < 0.2521$	–	$\delta = 0.3571$	medium
9	S.02	6.0 : 6.0	6.14 : 5.50	0.90 : 1.07	$p < 0.2751$	–	$\delta = 0.3393$	medium

Table 11.18: *Significance test and effect size between Likert scale results of "w/o Video : w/ Video of IG" — Part 4 of 6*

w/o Video – Information Giver : w/o Video – Information Follower — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.03	7.0 : 4.0	5.86 : 3.86	1.68 : 1.46	$p < 0.0495$	*	$\delta = 0.6327$	large
2	S.08	6.0 : 4.0	5.71 : 4.14	1.38 : 2.12	$p < 0.1690$	–	$\delta = 0.4490$	medium
3	S.01	6.0 : 4.0	5.43 : 4.14	1.90 : 1.95	$p < 0.2412$	–	$\delta = 0.3878$	medium
4	S.05	6.0 : 6.0	5.57 : 4.29	1.51 : 2.36	$p < 0.2583$	–	$\delta = 0.3673$	medium

Table 11.19: *Significance test and effect size between Likert scale results of "IG : IF of w/o Video" — Part 5 of 6*

w/o Video – Information Follower : w/ Video – Information Giver — Survey result comparison								
No.	Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
1	S.14	5.0 : 4.5	5.43 : 4.50	1.13 : 1.41	$p < 0.1735$	–	$\delta = 0.4107$	medium

Table 11.20: *Significance test and effect size between Likert scale results of "w/o Video – IF : w/ Video – IG" — Part 6 of 6*

■ Key Results.

From the lack of significantly different Likert score statement results and the lack of medium and large effect size results, it is possible to conclude that the difference of the gender of the subject and the map used by the subjects during the computer-mediated dialogue data collection were not factors to consider in the analysis of the user survey. This is also backed by the similar median and mean values of the different gender and map (Table A.15 and Table A.16 of Appendix A) and also the graphical representation (Figure A.26 and Figure A.27 of Appendix A) of the 2 groups.

However the same cannot be said about the groupings of "Setting (w/o Video or w/ Video)", "Role (IG or IF)", and "Language (Pt or En)". Though not always evident in the median value of the Likert results, the mean values show that the subject of "w/ Video", "IF" and "En" (Table A.16 of Appendix A) displayed stronger dissatisfaction towards the ILMT-s2s System.

11.4.3.1 Subject Setting Comparison

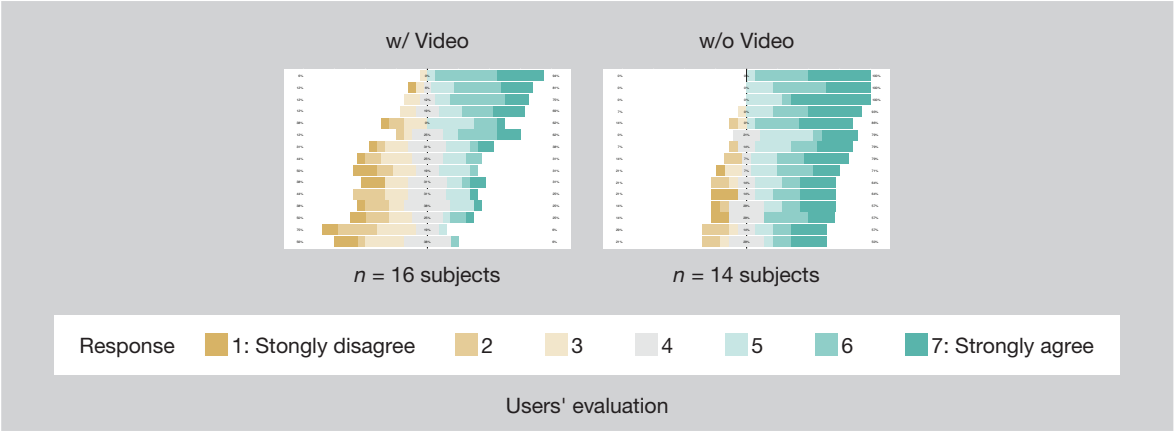


Figure 11.5: *Stacked bar chart of subject survey Likert scale replies — Setting*

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
w/ Video															
Mdn	3.5	5.5	4.0	3.0	3.5	5.0	6.0	3.0	4.0	6.0	5.5	4.0	6.0	4.0	4.0
M	3.56	5.13	3.88	3.13	3.38	4.31	5.44	2.75	3.75	5.44	5.31	4.19	6.13	3.69	3.69
SD	1.79	1.50	1.89	1.36	1.63	1.74	1.63	1.13	1.44	1.31	1.40	1.64	1.02	1.49	1.40
w/o Video															
Mdn	5.0	6.0	4.5	5.5	6.0	6.0	7.0	5.0	5.5	6.0	6.5	5.5	6.0	5.0	5.0
M	4.79	6.00	4.86	4.79	4.93	5.43	6.36	4.93	5.07	6.14	6.43	5.43	5.86	5.43	5.00
SD	1.97	1.18	1.83	2.30	2.02	1.74	0.93	1.90	1.77	0.77	0.65	1.45	1.56	1.16	1.80

Table 11.21: *Median, mean and standard deviation value per statement of subject survey Likert results*

■ Key Results.

When the subjects of the computer-mediated dialogue data collection are divided by the setting with or without live video streaming, since there are only 3 statements that did not provide a significant difference or a medium or large effect size, it would be natural to say that the subjects' perceived experience of the ILMT-s2s System was clearly different (Table A.17 of Appendix A). The three statements that didn't show a significant difference were S03 "effectiveness of system", S10 "interface of the system" and S13 "voice of system". Statement S10 and S13 both result in a high median (*Mdn*) value of 6.0 for both settings, but the interesting point is that statement S03 both show an average median (*Mdn*) value of 4.0 (w/ Vid) and 4.5 (w/o Vid).

- The lack of significant difference for statement S03 indicates that the "effectiveness of system" was not defined by the setting difference while the perception of the system was different.

w/ Video $(Mdn : M : SD) = 4.0 : 4.25 : 1.76$

w/o Video $(Mdn : M : SD) = 6.0 : 5.43 : 1.65$

11.4.3.2 Subject Role Comparison

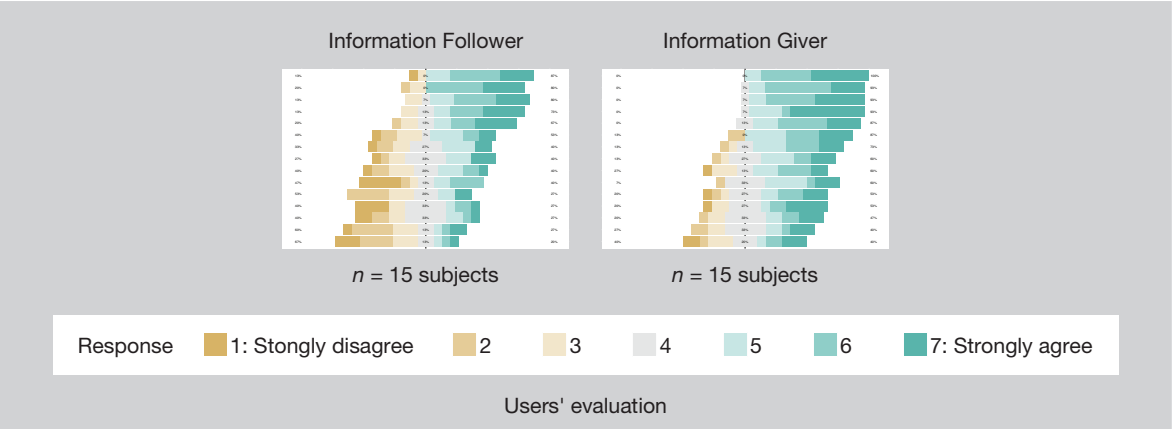


Figure 11.6: *Stacked bar chart of subject survey Likert scale replies — Role*

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
Information Follower															
Mdn	3.0	6.0	4.0	4.0	4.0	5.0	6.0	3.0	4.0	6.0	6.0	4.0	6.0	4.0	3.0
M	3.53	5.27	3.67	3.60	3.47	4.27	5.53	3.07	4.00	5.53	5.53	4.33	5.67	4.07	3.67
SD	1.92	1.71	1.72	1.96	2.13	1.87	1.64	1.83	1.69	1.36	1.46	1.76	1.63	1.71	1.71
Information Giver															
Mdn	5.0	6.0	5.0	4.0	4.0	6.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0
M	4.73	5.80	5.00	4.20	4.73	5.40	6.20	4.47	4.73	6.00	6.13	5.20	6.33	4.93	4.93
SD	1.83	1.01	1.89	2.08	1.58	1.59	1.08	1.68	1.71	0.85	0.92	1.47	0.72	1.39	1.49

Table 11.22: *Median, mean and standard deviation value per statement of subject survey Likert results*

■ Key Results.

When the subjects of the computer-mediated dialogue data collection are divided by role, a significant difference and a medium to large effect size is present for statements S08 “future productivity”, and S15 “Overall satisfaction” and a medium effect size present for S01 “ease of using the system as a whole”, S03 “effectiveness of system”, S06 “comfortable using the system” (Table A.18 of Appendix A).

- With the Information Follower subjects providing a lower Likert score for statements related to the usability of the ILMT-s2s System, further investigation is needed to verify if there were differing requirements based on the role of the subject — In other words, was the Information Follower receiving clear understandable instructions to perform the task? Was the Information Follower able to communicate his/her needs to the Information Giver? Did the dialogue structure change because of the ILMT-s2s System when compared to communication that is not mediated by a computer system?

Information Follower $(Mdn : M : SD) = 4.0 : 4.35 : 1.91$

Information Giver $(Mdn : M : SD) = 6.0 : 5.25 : 1.57$

11.4.3.3 Subject Language Comparison

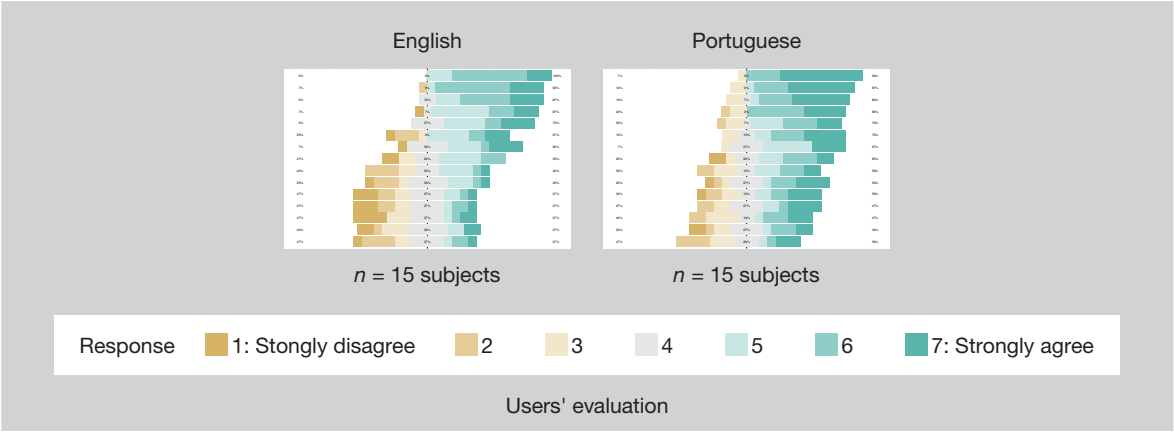


Figure 11.7: *Stacked bar chart of subject survey Likert scale replies — Language*

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
English															
Mdn	4.0	5.0	4.0	4.0	4.0	5.0	5.0	4.0	5.0	6.0	6.0	5.0	6.0	4.0	4.0
M	3.67	5.40	3.80	3.60	3.53	4.53	5.27	3.47	4.20	6.00	5.80	5.00	5.93	3.87	3.93
SD	1.76	1.18	1.78	2.06	1.92	2.03	1.49	1.85	1.61	0.65	1.01	1.65	1.22	1.64	1.58
Portuguese															
Mdn	5.0	6.0	5.0	4.0	5.0	5.0	7.0	4.0	4.0	6.0	6.0	5.0	7.0	5.0	4.0
M	4.60	5.67	4.87	4.20	4.67	5.13	6.47	4.07	4.53	5.53	5.87	4.53	6.07	5.13	4.67
SD	2.06	1.63	1.92	1.97	1.88	1.55	1.06	1.91	1.85	1.46	1.46	1.68	1.39	1.30	1.80

Table 11.23: *Median, mean and standard deviation value per statement of subject survey Likert results*

■ Key Results.

When the subjects of the computer-mediated dialogue data collection are divided by language, a significant difference and a medium to large effect size is present for statements S07 "easy to learn how to use", and S14 "satisfaction of the system's output" and a medium effect size present for S05 "efficiency" (Table A.19 of Appendix A).

- With the English speaking subjects providing a lower Likert score for statements related to the performance of the ILMT-s2s System, further investigation is needed to verify if there was a difference in the performance of the ILMT-s2s System — In other words, did the Automatic Speech Recognition (ASR) perform better and provide a lower WER for the Portuguese speakers or did the Machine Translation (MT) provide better translation?

Eng $(Mdn : M : SD) = 5.0 : 4.53 : 1.81$

Por $(Mdn : M : SD) = 6.0 : 5.07 : 1.77$

11.4.3.4 Subject Setting and Role Comparison

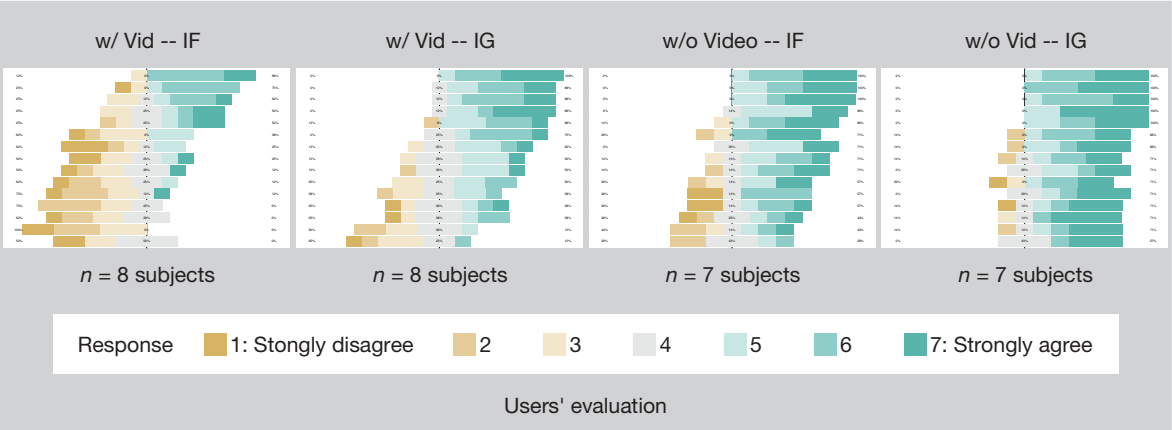


Figure 11.8: *Stacked bar chart of subject survey Likert scale replies — Setting and role*

	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
w/ Video — Information Follower															
Mdn	2.5	4.5	3.5	3.5	2.5	3.0	6.0	2.0	3.0	5.5	4.5	3.5	6.0	3.0	2.5
M	3.00	4.75	3.50	3.00	2.75	3.38	4.88	2.13	3.00	5.00	4.88	3.50	5.88	2.88	2.75
SD	1.85	1.83	2.00	1.31	1.75	1.51	1.89	0.83	1.31	1.51	1.64	1.77	1.25	1.13	0.89
w/ Video — Information Giver															
Mdn	4.0	6.0	4.0	3.0	4.0	5.5	6.5	3.5	4.5	6.0	6.0	5.0	6.5	4.5	4.5
M	4.13	5.50	4.25	3.25	4.00	5.25	6.00	3.38	4.50	5.88	5.75	4.88	6.38	4.50	4.63
SD	1.64	1.07	1.83	1.49	1.31	1.49	1.20	1.06	1.20	0.99	1.04	1.25	0.74	1.41	1.19
w/o Video — Information Follower															
Mdn	4.0	6.0	4.0	5.0	6.0	6.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0
M	4.14	5.86	3.86	4.29	4.29	5.29	6.29	4.14	5.14	6.14	6.29	5.29	5.43	5.43	4.71
SD	1.95	1.46	1.46	2.43	2.36	1.80	0.95	2.12	1.35	0.90	0.76	1.25	2.07	1.13	1.89
w/o Video — Information Giver															
Mdn	6.0	6.0	7.0	6.0	6.0	6.0	7.0	6.0	6.0	6.0	7.0	6.0	6.0	5.0	6.0
M	5.43	6.14	5.86	5.29	5.57	5.57	6.43	5.71	5.00	6.14	6.57	5.57	6.29	5.43	5.29
SD	1.90	0.90	1.68	2.21	1.51	1.81	0.98	1.38	2.24	0.69	0.53	1.72	0.76	1.27	1.80

Table 11.24: *Median, mean and standard deviation value per statement of subject survey Likert results*

■ Key Results.

To study how the setting and role affect the subjects of the computer-mediated dialogue data collection, the Likert results were further divided into groups of "w/o Vid – IG", "w/o Vid – IF", "w/ Vid – IG", and "w/ Vid – IF". This division resulted in Figure A.28 (stacked bar chart) and Figure A.29 (density plot) of Appendix A that illustrate the following increase in median and mean Likert score results:

w/ Video – IF $(Mdn:\ M) = 3.0 : 3.68 <$ w/ Video – IG $(Mdn:\ M) = 5.0 : 4.82 <$

w/o Video – IF $(Mdn:\ M) = 5.0 : 5.10 <$ w/o Video – IG $(Mdn:\ M) = 6.0 : 5.75$

When the setting and role were compared seperately, the statements with a significant difference or a medium to large effect size were as follows:

- Setting (w/ Vid or w/o Vid) — Refer to Table A.17 of Appendix A for the p-values and δ scores. Significant differences were observed in statements: S04, S05, S08, S09, S11, S12, S14, and S15. Medium and large ES were observed in statements: S01, S02, S04, S05, S06, S07, S08, S09, S11,

S12, S14, and S15.

- Role (IG or IF) — Refer to Table A.18 of Appendix A for the p-values and δ scores.
  Significant differences were observed in statements: S08, and S15.
  Medium and large ES were observed in statements: S01, S03, S06, S08, and S15.

Though obvious from Figure A.28 and Figure A.29 of Appendix A, by studying the setting and role combined it is possible to see, numerically, that the overall satisfaction (statement S15 in Table 11.24) of the ILMT-s2s System is strongly affected by the setting of w/ Video and the role of Information Follower.

- Setting (w/o Vid or w/ Vid – IG) — Refer to Table A.25 of Appendix A for the p-values and δ scores.
  Significant differences were observed in statements: S08.
  Medium and large ES were observed in statements: S01, S02, S03, S04, S05, S08, S11, S12, and S14.
- Setting (w/o Vid or w/ Vid – IF) — Refer to Table A.23 of Appendix A for the p-values and δ scores.
  Significant differences were observed in statements: S09, S12, S14 and S15.
  Medium and large ES were observed in statements: S01, S02, S04, S05, S06, S07, S08, S09, S10, S11, S12, S14, and S15.

The difference of the w/o Video and w/ Video for the Information Giver is that the addition of the live video stream has given the impression to the subject that the task is slow and not productive — the impression that the subject could become productive using this system in the future (statement S08) has reduced from a positive median of 6.0 to a slightly negative 3.5 (Table A.25 of Appendix A). This subject impression is also noticed in the large effect size of statement S04 "I was able to complete the task quickly" with an even greater fall from a median of 6.0 to 3.0. The difference for the Information Follower on the other hand is related more to the difficulties of the communications with lower scores in S09 "When I made a mistake I could recover easily and quickly", S12 "The system has all the functions and capabilities I expected", S14 "I was satisfied with the output". These lower scores do not represent productivity or speed, but difficulties to understand and express to the interlocutor.

- Role (IG or IF – w/o Video) — Refer to Table A.26 of Appendix A for the p-values and δ scores.
  Significant differences were observed in statements: S03.
  Medium and large ES were observed in statements: S01, S03, S05, and S08.
- Role (IG or IF – w/ Video) — Refer to Table A.24 of Appendix A for the p-values and δ scores.
  Significant differences were observed in statements: S06, S08, S09, S14, and S15.
  Medium and large ES were observed in statements: S01, S05, S06, S07, S08, S09, S10, S12, S14, and S15.

Studying the difference from the angle of the role w/o Video, we see that the Information Follower feels the difficulty of communication more than the Information Giver with a neutral median compared to the high 7.0 of the Information Giver. With the addition of the live video stream, the dissatisfaction of the Information Follower increases drastically with the Information Follower expressing negativity for comfortable, productive, ease of recovery, expectations, output satisfaction and overall satisfaction.

11.4.3.5 Subject Role, Setting and Language Comparison

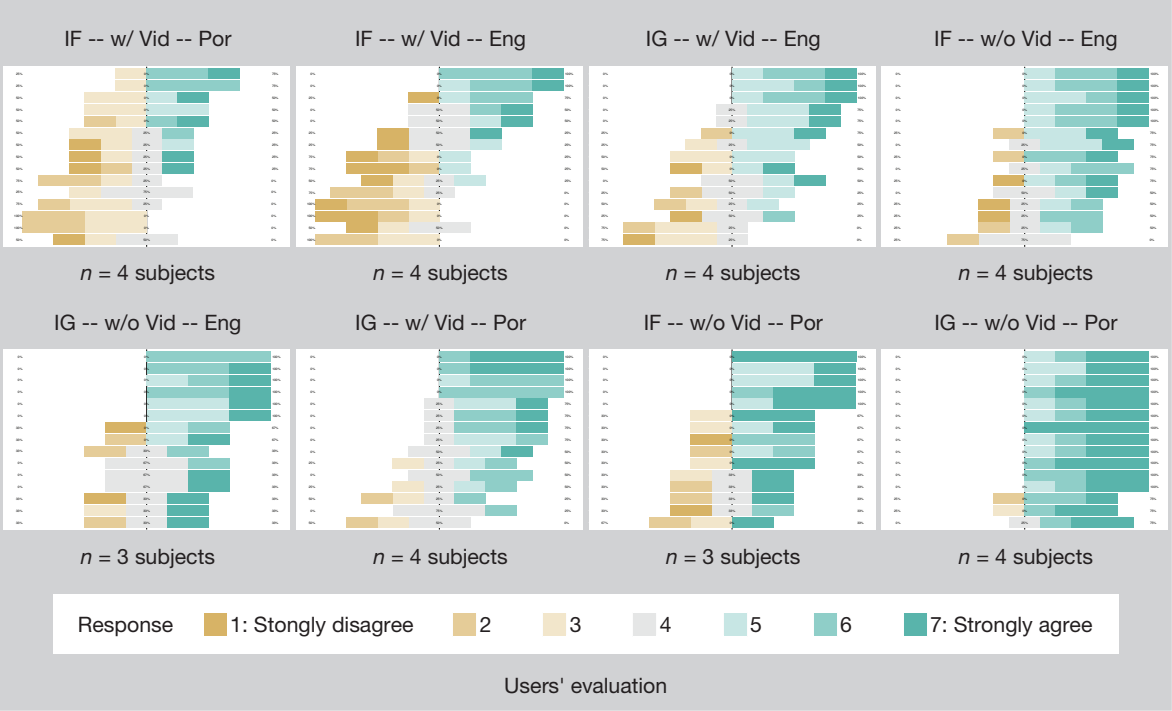


Figure 11.9: *Stacked bar chart of subject survey Likert scale replies — Role, setting and language*

■ Key Results.

Adding the language differences to the combined grouping in § 11.4.3.4, doesn't change the ordering of the setting and role preferences, but it does show that apart from the Portuguese speakers in the setting of w/ Vid and role of Information Follower, they all perceived the system more highly than their english speaking counterparts.

IF – w/ Video – Por $(Mdn: M) = 3.0 : 3.72 <$ IF – w/ Video – Eng $(Mdn: M) = 4.0 : 3.65 <$ IG – w/ Video – Eng $(Mdn: M) = 5.0 : 4.47 <$ IF – w/o Video – Eng $(Mdn: M) = 5.0 : 5.07 <$ IG – w/o Video – Eng $(Mdn: M) = 5.0 : 5.09 <$ IG – w/ Video – Por $(Mdn: M) = 5.5 : 5.17 <$ IF – w/o Video – Por $(Mdn: M) = 6.0 : 5.16 <$ IG – w/o Video – Por $(Mdn: M) = 7.0 : 6.25$

11.5 Summary

Be careful of what you wish for.

I started the introduction of this chapter with these words because the desire to add a video/image channel to the communication is strong. Since the introduction of the Picturephone by the American Telephone and Telegraph Company (AT&T) in 1964 (Lipartito, 2003), industry has continued to add an image channel to the conversation modality, and today, it is more difficult to find a mobile phone that doesn't have this function. The desire to add the video/image has been answered, but the common understanding of how to optimise the video/image channel is not common knowledge to the user yet, at least not the subjects of the computer-mediated dialogue data collection.

■ Key Results.

- ILMT-s2s System is perceived more positively when used in the setting without the live video streaming ("*w/o Video*")
- ILMT-s2s System is perceived more positively by the Information Giver
- ILMT-s2s System is perceived more positively by Portuguese speakers

From the results of the user survey, the subjects of the computer-mediated dialogue dataset have indicated that the availability of the video channel is not, overall, a positive addition to the communication tool. However, the point of interest from this analysis is the fact that there was no significant difference to the statement regarding the "effectiveness of the system" when the results where divided by setting (w/ Vid or w/o Vid), but that there was a significant difference for this statement when the subjects were divided by role (Information Giver or Information Follower). Finn et al. (1997) have already reported that the task effectiveness is mainly determined by the quality (in terms of low latency) of the audio channel, thus implying that video is of marginal importance. The results from the computer-mediated dialogue data do not object the importance of the audio channel, but highlights the difference that the role can apply to the subject. The different demands that the two different roles of Information Giver and Information Follower implied, with the Information Follower perceiving the system as less effective. A study into video enhanced communication in medical meetings reported that video enhanced the diverse needs of communication, but only after eight months of use (Kane & Luz, 2006b), However, there have been studies that report that in communication without eye contact, subjects blindly trust that the interlocutor has understood their instructions (Veinott, Olson, Olson, & Fu, 1999).

The fact that the subjects preceived the ILMT-s2s System better in the setting of w/o Video does not mean that the subjects did not appreciate the possibility to see their interlocutor. It only indicates that the resulting experience was not as positive as when they could not see the interlocutor. Though not directly asked to the subjects in the user survey, comments left by the subjects suggesting that the video channel was used as a feedback channel: "it's interesting because as a native speaker you

can deduce what someone meant to say even if it comes out wrong, but the other person can't. There was no way for her to know that the word for temple ('templo') was very similar to the word for time ('tempo'), so whenever 'time' came up she looked immensely confused. It was funny but also frustrating." or "More or less, it helped that I was seeing the other person face, so I knew by his facial expressions when some words weren't matching". If Face-to-Face communication is seen as just another modality of communication along side that of w/ Video and w/o Video used in the computer-mediated dialogue data collection, it is more understandable they these patterns occur. If the subject who uses the w/ Video setting interacts with the system with the expectation that the ILMT-s2s System will imitate Face-to-Face communication but with the enhancement of an interpreter, this inevitably compares the strength of a modality and conversation style that one is familiar with to one that only seems familiar, but in reality is very much different and new to the subject.

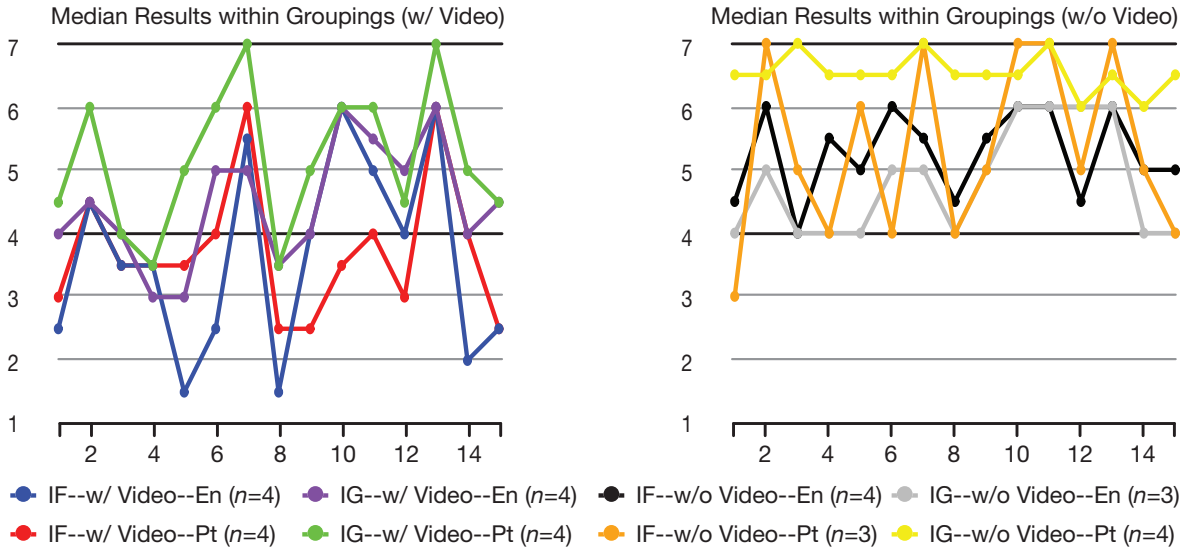


Figure 11.10: *7 point Likert score result grouping median comparison — w/ Video & w/o Video*

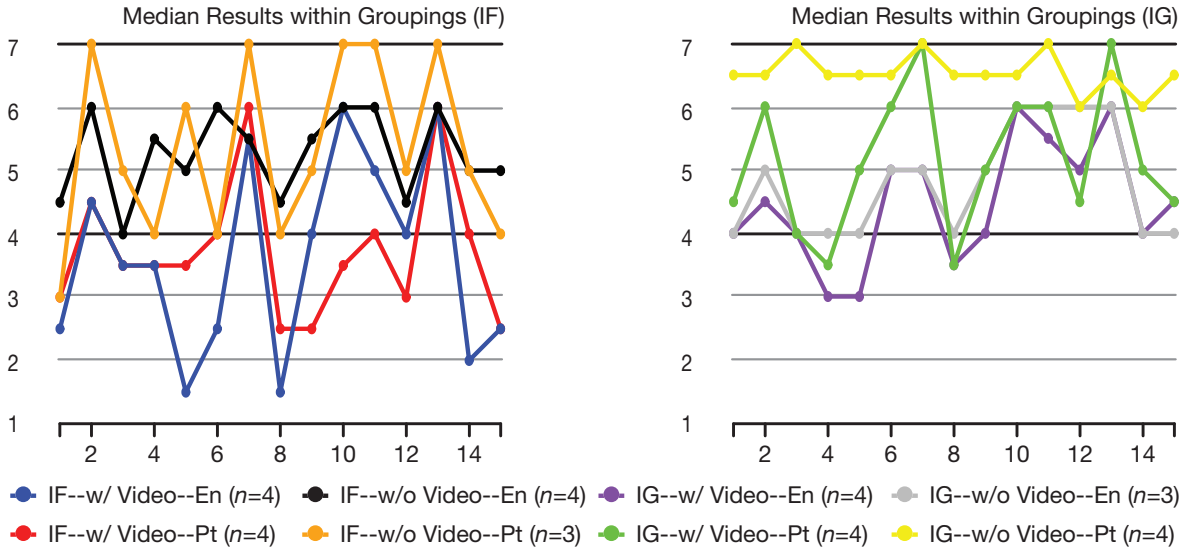


Figure 11.11: *7 point Likert score result grouping median comparison — IF & IG*

Chapter 12

Discussions

A summary of the main chapters in this thesis can be found at the end of each chapter as listed below.

- Chapter 5 "Speech Rate Difference" summary ➔ Pages 115 to 116
- Chapter 6 "The Three Talk Types" summary ➔ Pages 130 to 132
- Chapter 7 "Speech Rate Differences Between Datasets" summary ➔ Pages 142 to 147
- Chapter 8 "Dialogue Acts Differences Between Datasets" summary ➔ Pages 154 to 160
- Chapter 9 "ASR Performance & Speech Rate" summary ➔ Pages 173 to 175
- Chapter 10 "ASR Performance & Perceived Emotions Analysis" summary ➔ Pages 189 to 190
- Chapter 11 "User Survey Study" summary ➔ Pages 211 to 212

Here I summarise how the results in these chapters have answered the research question. Three questions were asked in § 1.2.2 and the analysis of the computer-mediated dialogue dataset and the comparison of the computer-mediated dialogue dataset with the non-mediated dialogue dataset have contributed to the following.

12.1 Question One

■ How are the subjects adapting to the interlingual computer mediated communication style?

The simplest answer to this question would be, as mentioned in Chapter 8: "Dialogue Acts Differences Between Datasets", that they are adapting to it as if they were using a high latency video-conferencing system. When the subjects talk to their interlocutor using the Speech-to-Speech Machine Translation (S2S-MT) system they use a slower rate of speech than when they are talking directly to a fellow human (Chapter 6). Though academic literature is split (Borsel & Maesschalck, 2008, p. 681) on the subject of gender speech rate differences, the subjects of the computer-mediated dialogue dataset demonstrated a gender difference in the adaptation to the S2S-MT system with the female subjects talking slower

to the S2S-MT system than the male subjects of the dataset even though this reduction in the female speech rate did not help them improve the Automatic Speech Recognition (ASR) performance (Chapter 9). As with the gender difference, a speech rate difference was also identified with the role of the subject, with the Information Giver talking slower than the Information Follower (Chapter 9). The instructions from the subjects with the role of Information Giver are short and the replies from their interlocutor, the Information Follower are longer but less frequent (Chapter 8).

12.2 Question Two

- How does the communication style compare with the direct Human-to-Human communication style of the non-mediated dialogue dataset?

The comparison of the interlingual computer mediated communication (the ILMT-s2s corpus) and the direct Human-to-Human communication (the HCRC Edinburgh Map Task corpus) indicated that though during direct Human-to-Human communication the subjects of both datasets used approximately the same speech rate, when the subjects of the computer-mediated dialogue dataset were talking to their interlocutor using the S2S-MT system, they used a speech rate slower than the subjects of the non-mediated dialogue dataset (Chapter 7). Also, when this speech rate difference was nullified, the speech rate differences between the role of the subjects, where the Information Giver speaks slower than the Information Follower, was similar in both datasets, however a gender difference that was not present in the non-mediated dialogue dataset was observed in the computer-mediated dialogue dataset, with the female subjects talking slower than the male subjects (Chapter 7). Furthermore, the dialogue acts of the two datasets showed clear difference. The communication style used for the interlingual computer-mediated dialogues was closer to high latency video-conferencing communication with a reduction in communicated backchannels, shorter utterances for the Information Giver and longer utterances from the Information Follower as indicated in Figure 12.1, indicating that the cost (required effort)[1] of communication had risen and used with more consideration (Chapter 8).

12.3 Question Three

- Why were the subjects of the computer-mediated dialogue data collection adapting, if, they are at all adapting to the interlingual computer mediated interaction?

The subjects of the computer-mediated dialogue dataset did adapt to the ILMT-s2s System. From the written comments in the user survey (listed in section § A.5.4 of the Appendix A), there were

[1] To quote Newlands et al. (2003, p. 344), "The use of precise instruction giving in the CMC context also seems to be a sensible way of adapting to the restrictive nature of this text-based communicative context. It may take less collaborative effort to ground precise instructions in this written context than the rather underspecified messages often found in spoken interactions; this would reduce the number of turns (or cycles of presentation and acceptance phases) required to establish mutual understanding. In contrast, in spoken interactions the use of highly specified instructions might be counter-productive, as they would require several repetitions before the information could be assimilated correctly. Very precise oral instructions could also be deemed to violate Grice's Cooperative Principle and the associated conversational maxims of quantity, quality, relation and manner (Grice, 1975)."

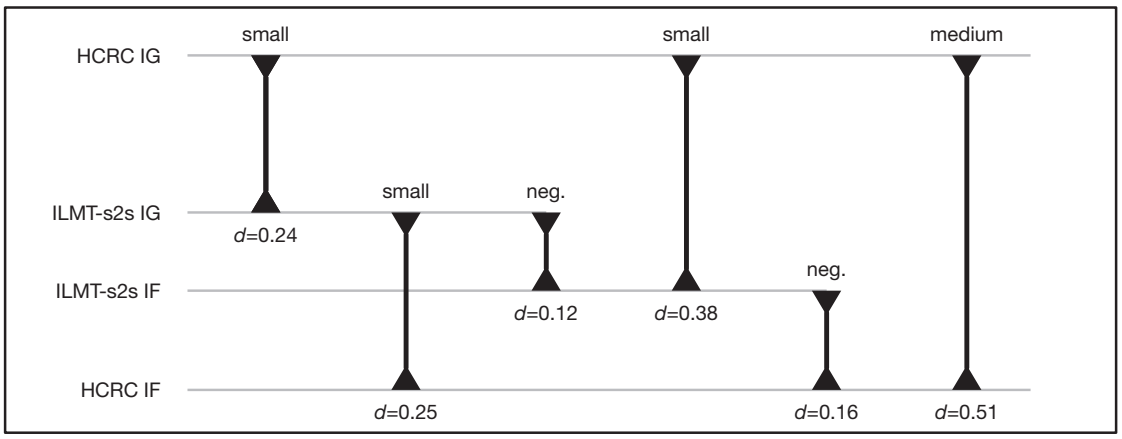


Figure 12.1: *Effect size (Cohen's d) of difference of word count per dialogue act per dataset/role.*

indications that the subjects changed their style of speech (by trying to be clearer, slower — hyperarticulation) to adapt to the ILMT-s2s System and improve the ASR performance. As lightly mentioned in § 9.4.3, Dr. Saturnino Luz found that the speech rate of the subjects would decrease after an ASR error for five utterances. However, when I looked at when the reduction in speech rate started, of the first five *On-Talk* utterances, of the fourteen speakers who start with a one hundred percent correct ASR result, the reduction in speech rate was before the first ASR error as mentioned in § 6.5: "Discussion". This means that hyperarticulation within the computer-mediated dialogue data is not only caused by the ILMT-s2s System, but also may be from the thought process needed for the map task.

The comparison of the computer-mediated dialogue dataset with the non-mediated dialogue dataset makes it clear that there is a difference between the two datasets (Chapter 7 and Chapter 8), but there is no definitive answer to the question of why. From the user survey results, there was a clear indication that the subjects using the w/ Video setting (were able to see their interlocutor) and had the role of Information Follower, didn't perceive the ILMT-s2s System as highly as subjects using the w/o Video setting (couldn't see their interlocutor), or had the role of Information Giver, providing the instructions. However, for the role of the subject there was no significant difference in the ASR performance, and even though there was a significant difference for the setting, the effect size was small.

When the perceived emotion labels are grouped by language, setting, and role and compared with the user survey results, though it is clear from the user survey results that the subjects that didn't use the w/ Video setting graded the S2S-MT system more highly, the perceived emotions do not portray the same story. The perceived emotion of *Amused* is substantially higher in the w/ Video setting for English subjects, making up ≈62% of the perceived emotion labels within these dialogues, but yet the satisfaction of the S2S-MT system is low (Chapter 11: "User Survey Study"). In the w/o Video setting the perceived emotion of *Amused* only accounts for ≈49% of the perceived emotions linked to *On-Talk*. In Chapter 10, the difference between English and Portuguese speaking subjects

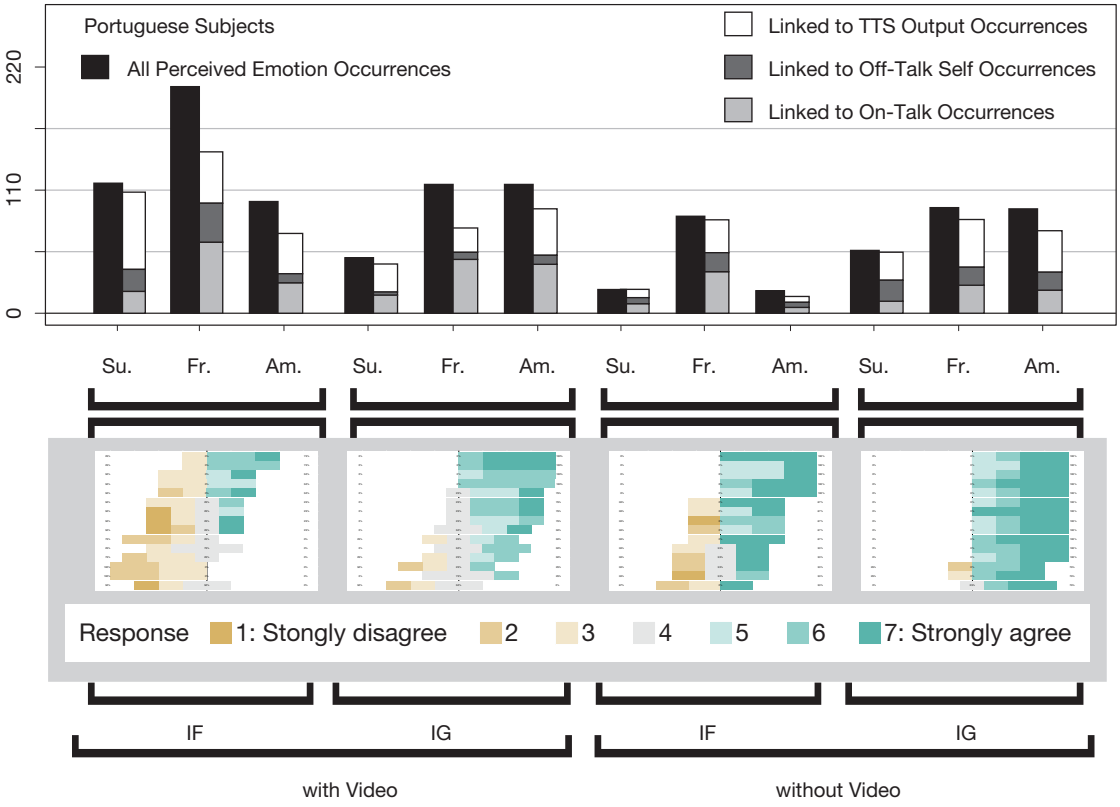


Figure 12.2: *Bar chart of Portuguese subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused), with a stacked bar chart of each grouping's user survey Likert score results*

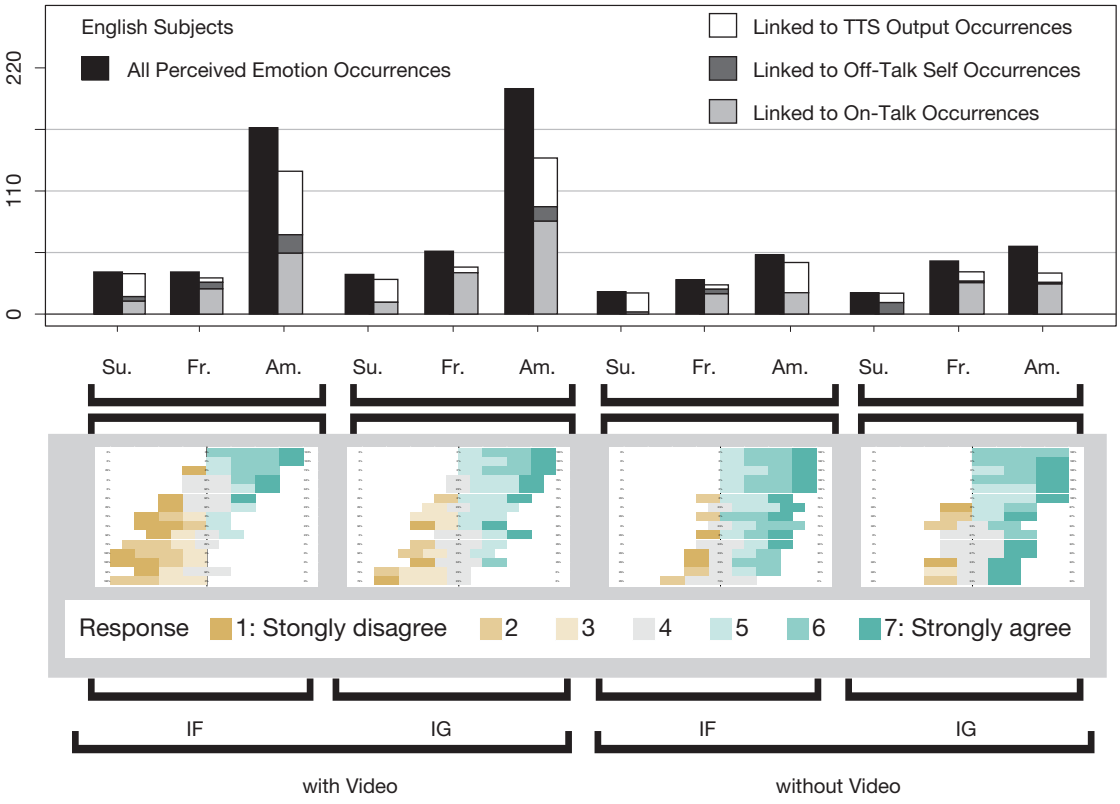


Figure 12.3: *Bar chart of English subject's perceived emotion count (Su.: Surprised, Fr.: Frustrated, Am.: Amused), with a stacked bar chart of each grouping's user survey Likert score results*

was highlighted with English speaking subjects expressing all three perceived emotions with WER means that are higher than utterances of a neutral perceived emotions — this included the perceived emotion of *Amused*. This indicates that the higher proportion of *Amused* didn't relate to a better perception of the S2S-MT system, nor did it result in a lower or higher speech rate. Since there is no clear answer to the "why", maybe the importance should be to eliminate the problem without knowing why it was caused. The problem that the female subjects used a slower speech rate than the male subjects: an effort that did not reward the female subjects with a better ASR performance. The problem that the subjects clearly prefer the communication modality where they cannot see the interlocutor: manage the expectations of the subjects so that they do not use a register related to a direct Human-to-Human communication style.

12.4 Future works

In this thesis, the results in (i) Chapter 5: "Speech Rate Difference", (ii) Chapter 6: "The Three Talk Types", (iii) Chapter 7: "Speech Rate Differences Between Datasets" and (iv) Chapter 11: "User Survey Study", are conclusive in my opinion. However the same cannot be said about (i) Chapter 8: "Dialogue Acts Differences Between Datasets", (ii) Chapter 9: "ASR Performance & Speech Rate" and (iii) Chapter 10: "ASR Performance & Perceived Emotions Analysis". The study in these three chapters look at the data from a broad prospective. Just the overview, but they need to be looked into in more detail.

- Dialogue acts differences between computer-mediated dialogue and non-mediated dialogue data.

With Chapter 8, the utterance frequency and the word count of the utterances are looked into, and then the dialogue acts are compared. But this comparison excluded the dialogue acts of *Off-Talk* and it didn't compare the segmentation of sub-tasks within the overall task of replicating the whole route. Comparing the dialogue act sequence within each subtask with the *Off-Talk* dialogue acts included would provide a clearer indication of how the two communication types differ. What is intentionally communicated to the interlocutor, and what is uttered but not communicated to the interlocutor. Currently the hypothesis is that the 'cost' of communication (Newlands et al., 2003) increased and reduced the frequency and quantity of short and simple agreement/acknowledgement utterances. This has yet to be verified in detail to provide a definite answer for the computer-mediated dialogue data collected here, for this study. At the current state, the answer is provided, but it remains weak with areas remaining to strength the claim that utterances of the computer-mediated dialogue data are communicated to the interlocutor when the weight of the information is higher.

- ASR Performance & Speech Rate.

With Chapter 9, the association between the ASR Performance and the speech rate of the subjects was studied, but even though there were significant differences of speech rates between role, gender and language used by the subject, this was not reflected in the ASR performance of the computer

system. What is lacking in this analysis is a deeper look. As mentioned in § 6.5, not all speech rate reductions were caused by the ASR errors. Some subjects reduced their speech rate while the ASR output was correct.

- ASR Performance & Perceived Emotions.

With Chapter 10, the first step needed to continue with this studies is to annotate the sub-tasks within the map task. What I mean by this is to identify when an instruction to the next item on the map starts and finishes. This will provide definite brackets that identify where the sub-tasks are accomplished or not, and would allow to identify which type of perceived emotions are neighbouring communication elements in the context of each sub-tasks. An investigation of the communications items in each sub-task context would also be interesting to determine which communication items (utterances) are involved in each sub-task and if an ordering patterns can be found.

12.5 Conclusion

This chapter discussed the extent to which the results found in previous chapters answered the three main research questions that triggered the present thesis work, and gave an insight at possible questions that should be explored in future works. The following and last chapter provides a general conclusion.

Chapter 13

Conclusions

In this thesis, I have looked at the user of a Speech-to-Speech Machine Translation (S2S-MT) system. By looking at the self-adaptation of the user to the S2S-MT system within the dialogue to their interlocutor, and not the individual software components (ASR, MT, and TTS), I was able to identify patterns of communication that will help facilitate computer mediated communication if applied to the design of S2S-MT systems.

13.1 Introduction

To understand how a user will self-adapt to computer mediated interlingual task oriented communication, the ILMT-s2s corpus[1] was created in a way so that it could be compared to a corpus that has been widely used in the analysis of direct Human-to-Human (H2H) and remote collaboration in a monolingual setting, the HCRC Edinburgh Map Task corpus. The ILMT-s2s corpus is the first contribution I have made to the community. At the time of its creation, to the best of my knowledge, there was no other corpus that was directly comparable to the direct Human-to-Human communication dialogues of the non-mediated dialogue dataset (the HCRC Edinburgh Map Task corpus), using a S2S-MT system as the tool to mediate communication. From this dataset, I have been fortunate enough to be able to make ten easily identifiable contributions to the community, in the way of nine conference papers that I am the lead author, and therefore hold the responsibility of the contents, and one as second author — (i) Starting with the introduction to the idea of collecting the interlingual computer mediated map task (ILMT-s2s) corpus (Hayakawa, Campbell, & Luz, 2014), (ii) a report of the initial analysis of subject alignment to the ILMT-s2s System and its association with the ASR performance (Hayakawa, Cerrato, et al., 2015), (iii) a report of the initial perceived emotion analysis and its association with the ASR performance, with results of automatically identifying the three perceived emotion from the biosignal data of the subjects (Hayakawa, Haider, et al., 2015), (iv) a report of the initial analysis of the usability of the ILMT-s2s System in the setting with and without the

[1] International Standard Language Resource Number (ISLRN): 100-610-774-625-0

video channel (Cerrato et al., 2016), (v) a presentation of the contents and collection method of the computer-mediated dialogue dataset — the ILMT-s2s corpus (Hayakawa, Luz, Cerrato, & Campbell, 2016), (vi) a report of the three talk types and their characteristics found in the computer-mediated dialogue data, with results of automatically identifying these three talk types from the biosignal data of the subjects (Hayakawa, Haider, et al., 2016), (vii) a report of the comparison of direct H2H communication dialogues and interlingual computer-mediated dialogues on how the turns differ in dialogues acts and word count between the two different communication scenarios (Hayakawa, Luz, & Campbell, 2016), (viii) a report of the comparison of direct H2H communication dialogues and interlingual computer-mediated dialogues on how the speech rate differes between the two different communication scenarios (Hayakawa et al., 2017b), (ix) a report on the effect of including the video channel in interlingual computer mediated communicaiton (Hayakawa et al., 2017a), and (x) finally a detailed report of the method used to calculate the speech rate for the short utterances (*Mdn*: 4, *M*: 5, *SD*: 6 word utterances) that typify the utterance length (word count) of the data within the computer-mediated dialogue dataset (Hayakawa et al., 2018).

13.2 Thesis Contribution and Findings

The three main questions to be answered about the computer-mediated dialogue data were, (i) how are the subjects adapting to the interlingual computer mediated communication style, (ii) how does the communication style compare with the direct Human-to-Human communication style of the non–mediated dialogue dataset, and (iii) why would the subjects of the computer-mediated dialogue data collection be adapting, if, they are at all adapting to the interlingual computer mediated interaction.

Detailed explanations to the answer of these questions can be found between Chapter 4 and Chapter 12 of this thesis. However a simple summary is as follows.

■ How are the subjects adapting to the interlingual computer mediated communication style?

From the data of the ILMT-s2s corpus it has been identified that the subjects used three different basic speech rates. One while talking to the system that was the slowest, a second when talking to themselves, and the third when talking directly to a fellow human which was the fastest. When the subjects of the computer-mediated dialogue dataset talked to their interlocutor through the S2S-MT system (the ILMT-s2s System) Automatic Speech Recognition (ASR) performance did affect the speech rate, but more prominent was the gender difference with female subjects speaking slower than male subjects and the difference in the role of the subject with the Information Giver talking slower than the Information Follower.

■ How does the communication style compare with the direct Human-to-Human communication style of the non-mediated dialogue dataset?

The comparison of the interlingual computer mediated communication of the ILMT-s2s corpus and the direct Human-to-Human communication of the non-mediated dialogue dataset indicated that

when the subject of both datasets were talking directly to a fellow human, their speech rate was very similar. Also the fact that the Information Giver spoke slower than the Information Follower was another similarity between the datasets, but no speech rate difference was identified between the gender of the subjects in the non-mediated dialogue dataset, indicating that the identification of female subjects speaking slower than the male subjects might be an effect of the computer mediated communication. Furthermore, the dialogue acts of the two datasets showed clear difference. The communication style used for the interlingual computer-mediated dialogues was closer to high latency video-conferencing communication with a reduction in communicated backchannels, shorter utterances for the Information Giver and longer utterances from the Information Follower, indicating that the cost (required effort)[2] of communication had risen and used with more consideration.

- Why could the subjects of the computer-mediated dialogue data collection be adapting, if, they are at all adapting to the interlingual computer mediated interaction?

The subjects of the computer-mediated dialogue dataset did adapt to the ILMT-s2s System. From the comparison with the non-mediated dialogue dataset it is clear that there is a difference between the two datasets, but there is no definitive answer to the question of why. However there are indications that the ASR performance is not the only reason that the subject's spoke more slowly to the ILMT-s2s System. Of the subjects who had a perfect ASR result for their first utterances, the onset of reduced speech rate occurred before the first ASR error, so both the difficulty of the task on hand (the map task) and the ASR performance can be attributed as the cause of slower speech. From the user survey results, there was a clear indication that the subjects using the w/ Video setting, that were able to see their interlocutor, and had the role of Information Follower, following the instructions, didn't perceive the ILMT-s2s System as highly as subjects using the w/o Video setting, who couldn't see their interlocutor, or had the role of Information Giver, providing the instructions. However, for the role of the subject there was no significant difference in the ASR performance, and even though there was a significant difference for the setting, the effect size was small. Since there is no clear answer to the "why", maybe the importance should be to eliminate the problem without knowing why it was caused. The problem that the female subjects used a slower speech rate than the male subjects: an effort that did not reward the female subjects with a better ASR performance. The problem that the subjects clearly prefer the communication modality where they cannot see the interlocutor: manage the expectations of the subjects so that they do not use a register related to a direct Human-to-Human communication style.

[2]To quote Newlands et al. (2003, p. 344), "The use of precise instruction giving in the CMC context also seems to be a sensible way of adapting to the restrictive nature of this text-based communicative context. It may take less collaborative effort to ground precise instructions in this written context than the rather underspecified messages often found in spoken interactions; this would reduce the number of turns (or cycles of presentation and acceptance phases) required to establish mutual understanding. In contrast, in spoken interactions the use of highly specified instructions might be counter-productive, as they would require several repetitions before the information could be assimilated correctly. Very precise oral instructions could also be deemed to violate Grice's Cooperative Principle and the associated conversational maxims of quantity, quality, relation and manner (Grice, 1975)."

13.3 Looking back at the ILMT-s2s corpus

Looking back at the data collection and analysis that I performed on the data, evidently I have a few things that I would like to have done differently. There is a notice that was on nearly every manager's desk while I was working in industry, which read "現場現物確認 (Verify at the actual location with the actual item)". I understood this as a reminder that theory may be reasonable and logical, but the actual actions of humans and the consequences of the actual environment may cause side effects that could alter the result of the theory, indicating the importance to look at the bigger picture. One could say that my need to revert to this verification method is proof that my theory is not robust enough. All I can say is that, working with humans, this verification method has helped me find a better version of the actual truth, instead of the presumed truth, and prevented the implementation of ineffective countermeasures. Why I raise this point now, is that I did not follow this advice during the comparison of the dialogue act data from the non-mediated dialogue data and the computer-mediated dialogue data in Chapter 8 and the perceived emotion analysis in Chapter 10. I stated in section § 4.4.6: "20 Frame/Window Normalization" the importance to visualise the data that is available to see the big picture first. But for these two studies, the big picture was only verified during the writing of this thesis and not when the theory of the analysis was planned. This doesn't mean that the theory is wrong, just that other observations could have been noticed in advance. To be more precise, for the dialogue act comparison, it is the overlook of the *Off-Talk* dialogue acts which show that the dialogue act of "*Acknowledgement*" is used frequently by the subject, but just not transmitted to the interlocutor as *On-Talk*, and for the perceived emotion analysis, it is the overlook of the ASR performance of the TTS output utterances from the interlocutor. By looking at the ASR performance of the TTS output utterances, it may have provided a better understanding of how the ASR performance of what the subject heard was affecting the subject, especially for the perceived emotion of "*Surprised*" was used by the Portuguese subjects. Both of these items were noticeable once the data as a whole was visualised as illustrated in Figure 3.6 for the dialogue act comparison, and Figure 4.25, and Figure 10.10 for the perceived emotions analysis.

Looking at the data of the ILMT-s2s corpus there remain areas that could be studied. As mentioned in § 4.4.3: "Annotation", the annotation data of the computer-mediated dialogue dataset was not limited to the perceived emotions, dialogue acts and three talk types, that were listed in Figure 1.3 of § 1.5, but also included the facial expressions and head movement annotation labels. The biosignal data too also included eye gaze tracking and pupil dilation data. Furthermore, none of the studies in this thesis used the actual wording of the dialogue utterances as the data source or even the MT accuracy. A study from O'Malley, Langton, Anderson, Doherty-Sneddon, & Bruce (1996b, p. 177) suggested that "when speakers are not physically co-present, they are less confident in general that they have mutual understanding [. . .], and therefore over-compensate by increasing the level of both verbal and non-verbal information". The increase in non-verbal information was not studied for this thesis and

still remains an area that can be studied from the facial expressions and head movement annotation labels. Lexical alignment to MT that was reported by Schneider & Luz (2011) to be "detrimental to the interaction" has not been investigated either. Does the repetition of incorrect MT output actually occur in the task oriented dialogues of the computer-mediated dialogue dataset and also, how do the subjects recover from the translation errors caused by the erroneous alignment? Furthermore, a study of cognitive load measured from the pupillary responses (Paas, Tuovinen, Tabbers, & Van Gerven, 2003) in comparison with the manually labelled perceived emotions and the ASR performance would be an interesting topic to look into as a means of understanding if the perceived emotion labels of "*Amused*", "*Frustrated*", and "*Surprised*" relate to the cognitive load expressed through pupillary responses. Though more complex, one could also look into the quality of the MT output and its effect on the subjects. An evaluation of MT for all the utterances would be required as a pre-process, where a gold-standard translation for all utterances is created and compared with the MT output of the system. But also a point of interest of this evaluation would be the availability of the important words within the translation output. This is not limited to the MT, but also to the ASR performance and the overall dialogue communication strategy, but the interest would be to see the minimum requirement for the instruction of the task to be understood. Is there a group of specific instructive words that need to be transmitted or translated correctly for adequate communication within the map task scenario or not. This may also clarify how the subject recovers from ASR errors, to what extent will they repeat their utterances until they think that the ASR has captured enough words for the utterance to be deemed as sufficient to continue. These are just some of the questions that can be studied from the data of my corpus and provide a contribution, even if it may only be a small contribution, to the academic community.

13.4 Conclusions

Machine mediated technological advancement will not be reversed, due to the possibility to enhance human life (Baranyi & Csapó, 2012; Baranyi, Csapó, & Várlaki, 2014). However, talking to a computer interface is said to be similar for the subject as talking to a person who has hearing impairment, an effort that is more "exaggerated" than H2H communication due to the difficulty to understand its limitation (Branigan et al., 2010). It is therefore critical that computer mediated communication systems have the capacity to adapt to variable human communication styles, by increasing the varying differences that we have (Esposito, Esposito, & Vogel, 2015) so that the transitional adaptation is as smooth as possible. An adaptation period of 8 months was required for highly motivated, error critical professionals to appreciate an image/video channel (Kane & Luz, 2006a). Now, how long will it take someone who is scared of change and technology? I sincerely hope that the contributions mentioned in this thesis will provide design cues that will help develop systems that will help the people who need it, and not just the people that are in the know.

Appendix A

ILMT-s2s corpus

A.1 ILMT-s2s corpus Data Summary

Table A.1: A table of the media files that are included in each dialogue of the corpus including the number of system crashes of the ILMT-s2s System and the size of the live video screen window.

Table A.2: A table detailing all fifteen dialogues of the ILMT-s2s corpus with the duration, count of *On-Talk* utterances, and count of *On-Talk* words.

Table A.3: A table detailing all sixteen dialogues of the HCRC Edinburgh Map Task corpus with the duration, utterance count and word count.

A simple summary[1] of the media files in the ILMT-s2s corpus collection is provided below in Table A.1.

Dialogue	Vid. Stream	Front Vid.	Head Vid.	Back Vid.	Eye Track.	Audio	Crashes	(IG : IF)
m1 1 w/ V	Small	✓	✓ IG	✗	✓ IF	✓	0	(0 : 0)
m1 2 w/ V	Small	✓	✓ IF	✓ IF	✓ IG	✓	2	(1 : 1)
m1 3 w/ V	Big	✓	✗	✓ IG	✓ IF	✓	6	(4 : 2)
m1 4	–	✓	✓ IG	✓ IG	✓ IF	✓	0	(0 : 0)
m1 5	–	✓	✓ IF	✓ IF	✓ IG	✓	0	(0 : 0)
m1 6	–	✓	✗	✓ IG	✓ IF	✓	0	(0 : 0)
m1 7	–	✓	✓ IG	✓ IG	✓ IF	✓	0	(0 : 0)
m7 8 w/ V	Small	✓	✓ IF	✓ IF	✓ IG	✓	3	(1 : 2)
m7 9 w/ V	Big	✓	✓ IF	✗	✓ IG	✓	0	(0 : 0)
m7 10 w/ V	Big	✓	✓ IG	✗	✓ IF	✓	4	(2 : 2)
m7 11 w/ V	Big	✓	✗	✓ IF	✓ IG	✓	0	(0 : 0)
m7 12 w/ V	Big	✓	✓ IF	✓ IF	✓ IG	✓	0	(0 : 0)
m7 13	–	✓	✓ IF	✓ IF	✓ IG	✓	0	(0 : 0)
m7 14	–	✓	✓ IG	✓ IG	✓ IF	✓	0	(0 : 0)
m7 15	–	✓	✓ IF	✓ IF	✓ IG	✓	0	(0 : 0)
Total							15	(8 : 7)

Table A.1: *ILMT-s2s corpus data summary — Media Files*

[1]The indication "m1", "m7", "w/ V", "v/ Vc" and "w/ EC" under "Dialogue" in Table A.1, Table A.2, and Table A.3 denote, "Map 01", "Map 07", "Recorded with live video streaming between subjects", "System crashed during recorded with live video streaming between subjects" and "With eye contact" respectively.

A simple summary[2] of the dialogues spoken to the ILMT-s2s System (not including *Off-Talk*) is provided below in Table A.2.

Dialogue	Duration	*On-Talk*					
		Utt. Count	(IG :	IF [% Diff.])	Word Count	(IG :	IF [% Diff.])
m1 1 w/ V	00:23:02	88	(55 :	33 [−40%])	717	(450 :	267 [−40%])
m1 2 w/ Vc	00:48:57	217	(100 :	117 [+17%])	1,522	(593 :	929 [+57%])
m1 3 w/ Vc	01:14:20	209	(106 :	103 [−3%])	818	(422 :	396 [−6%])
m1 4	00:22:31	132	(59 :	73 [+24%])	530	(265 :	265 [±0%])
m1 5	00:40:20	262	(201 :	61 [−70%])	1,349	(915 :	434 [−53%])
m1 6	00:44:07	171	(93 :	78 [−16%])	1,310	(744 :	566 [−24%])
m1 7	00:37:55	183	(94 :	89 [−5%])	1,369	(888 :	481 [−46%])
m7 8 w/ Vc	00:59:29	174	(121 :	53 [−56%])	1,437	(1,016 :	421 [−59%])
m7 9 w/ V	00:29:13	202	(98 :	104 [+6%])	1,011	(522 :	489 [−6%])
m7 10 w/ Vc	00:36:07	168	(76 :	92 [+21%])	548	(254 :	294 [+16%])
m7 11 w/ V	00:43:17	157	(101 :	56 [−45%])	942	(481 :	461 [−4%])
m7 12 w/ V	00:48:19	176	(100 :	76 [−24%])	1,283	(551 :	732 [+33%])
m7 13	00:26:13	98	(49 :	49 [±0%])	652	(345 :	307 [−11%])
m7 14	00:25:59	126	(69 :	57 [−17%])	473	(294 :	179 [−39%])
m7 15	00:19:38	76	(34 :	42 [+24%])	416	(235 :	181 [−23%])
Total	09:39:27	2,439	(1,356 :	1,083 [−20%])	14,377	(7,975 :	6,402 [−20%])

Table A.2: *ILMT-s2s corpus data summary — Dialogues*

As a reference, a simple summary[3] of the HCRC Edinburgh Map Task corpus is provided below in Table A.3.

Dialogue	Duration	Utt. Count	(IG :	IF [% Diff.])	Word Count	(IG :	IF [% Diff.])
m1 1 w/ EC	00:06:14	140	(77 :	63 [−18%])	1,020	(744 :	276 [−63%])
m1 2 w/ EC	00:03:55	72	(40 :	32 [−20%])	587	(463 :	124 [−73%])
m1 3	00:05:15	143	(77 :	66 [−14%])	825	(601 :	224 [−63%])
m1 4	00:03:06	91	(45 :	46 [+2%])	536	(426 :	110 [−71%])
m1 5 w/ EC	00:07:52	151	(81 :	70 [−14%])	925	(788 :	137 [−83%])
m1 6 w/ EC	00:19:00	478	(257 :	221 [−14%])	3,004	(1,999 :	1,005 [−50%])
m1 7	00:07:42	175	(97 :	78 [−20%])	1,586	(1,188 :	398 [−66%])
m1 8	00:02:43	84	(45 :	39 [−13%])	595	(518 :	77 [−85%])
m7 9 w/ EC	00:08:14	199	(114 :	85 [−25%])	1,326	(1,007 :	319 [−68%])
m7 10 w/ EC	00:07:18	151	(78 :	73 [−6%])	1,294	(849 :	445 [−48%])
m7 11	00:14:41	449	(239 :	210 [−12%])	2,916	(1,597 :	1,319 [−17%])
m7 12	00:10:32	361	(196 :	165 [−16%])	2,071	(1,165 :	906 [−22%])
m7 13 w/ EC	00:05:13	171	(94 :	77 [−18%])	870	(657 :	213 [−68%])
m7 14 w/ EC	00:05:39	193	(109 :	84 [−23%])	948	(708 :	240 [−66%])
m7 15	00:04:57	103	(61 :	42 [−31%])	834	(758 :	76 [−90%])
m7 16	00:10:01	265	(146 :	119 [−18%])	1,751	(1,238 :	513 [−59%])
Total	02:02:22	3,226	(1,756 :	1,470 [−16%])	21,088	(14,706 :	6,382 [−57%])

Table A.3: *HCRC Edinburgh Map Task corpus data summary — Dialogues*

[2] A Note of Caution: The *On-Talk* Word Count comparison is over two different languages (English and Portuguese) and the language characteristics have not been taken into account. Therefore, it is not as simple to say that IG is explaining more within each utterance when compared with the IF until it is split by language.

[3] It is interesting to note the difference between IG and IF word count of the HCRC corpus and the ILMT-s2s corpus.

A.2 ILMT-s2s corpus Data Processing Workflows

Details of Step 1: A Final Cut Pro X Library was created for all fifteen dialogues as a location to edit the data and also to save all the original media files from each dialogue. Within each dialogue Library, three Project folders were created. One Project folder for each subject and one Project folder the media files to be placed once the video and audio of each subject was synchronised. These three folders were created to simplify the selection of the media files while editing without the need to renaming the original files.

Details of Step 2: After each recording session, the media files were imported into the Final Cut Pro X Library/Project. The 29.97 fps MPEG4-AVC/H.264 AVCHD format video recorded on the SONY HDR-XR500 and SONY HDR-CX370, and the MPEG-4 AVC/H.264 format video from the Panasonic HX-A100 were imported directly from the camcorder. The video from the SMI Eye Tracking Glasses that was formatted in 30 fps MPEG-4 ASP format using the Xvid codec was converted into a 29.97 fps Apple ProRes 422 using Compressor from Apple Inc.,[4] to prevent possible future compatibility problems with Xvid, and then imported into the appropriate Library/Project folder. The remaining media files (audio files) were copied from their respective recording devices and then imported into the Library/Project.

Details of Step 3: The audio files recorded while the subject was talking to the ILMT-s2s System using the "Push-to-talk" button, are synchronised with the video recorded on the SONY HDR-XR500. The audio from the SONY HDR-XR500 video was converted to monaural audio so that the recordings from the ILMT-s2s System could be attached as the second channel of a stereo audio file. This was to make it possible, for example, to create an output file with the audio from the SONY HDR-XR500 on the right channel and the audio from the ILMT-s2s System on the left channel of the stereo audio file.

Details of Step 4: With the SONY HDR-XR500 video and ILMT-s2s System utterance audio recordings synchronised, the file was then synchronised with the remaining media files. For subjects w/o Bio, this meant just the video from the Panasonic HX-A100, while for subjects w/ Bio, this meant the video from the SMI Eye Tracking Glasses and the audio from the Mind Media Nexus-4. Unfortunately, a timing problem, where the timeline of the media files differs slightly and causes the data to slowly go off sync, was noticed on some files. When this was noticed, the affected files were adjusted in relation to the main SONY HDR-XR500 video file. Details about how the affected file was adjusted can be found in the XML file exported in step 7.

Details of Step 5: Once step 4 was completed, the synchronised subject recording's file was moved from the subject Project folder to the dialogue Project folder to be synchronised with their interlocutor's media file. Since the subjects were in different rooms, there were not many common sounds so the Final Cut Pro X synchronisation function could not be used for this task, neither

[4] Refer to https://www.apple.com/final-cut-pro/compressor/ for details (accessed on 2017/10/15).

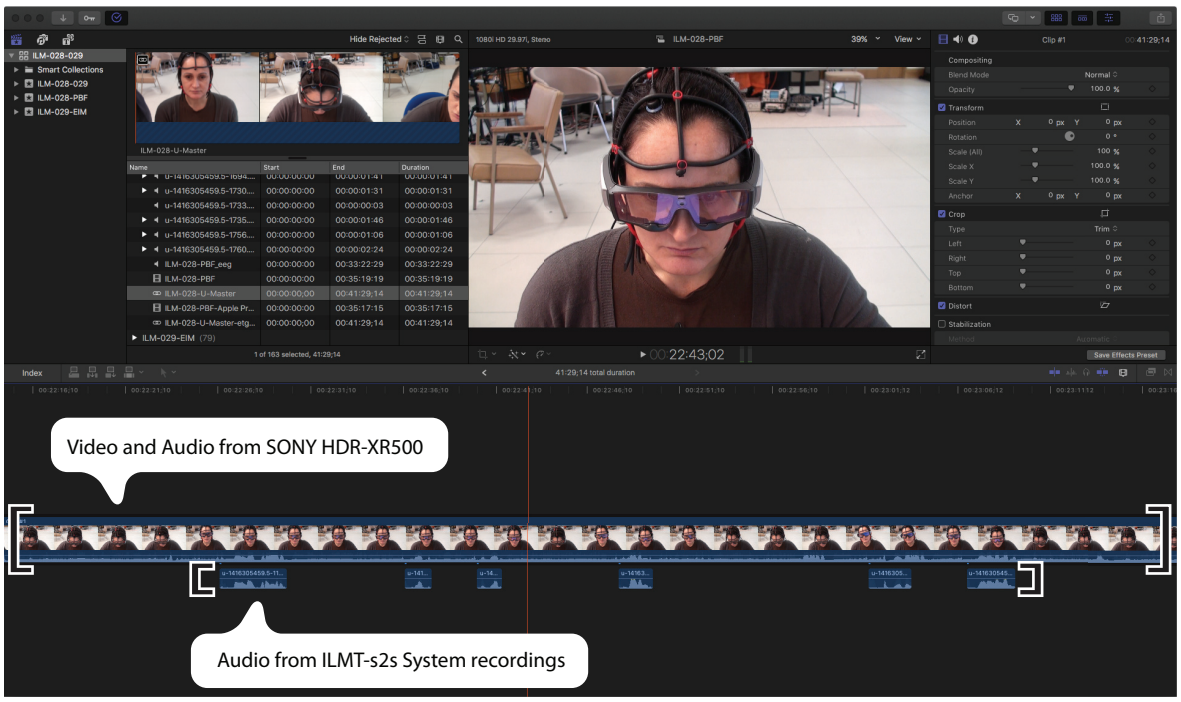


Figure A.1: *Image of audio synchronisation*

could ILMT-s2s System test utterances, since the exact delay of the system was unknown. Luckily, though the subjects were in separate rooms, a door with a strong spring was located nearly half way between the subject and also since the floor was raised, footsteps of the technician were also evident. For this reason, the sound of the door closing, or the foot steps of the technician was used to synchronise the media files of the two subjects and verified for start and end of the overlapping parts of the media files (if neither were usable, then external sounds such as police car sirens or some other kind of common identifier was used).

Details of Step 6: With the media files of both subjects synchronised, the synchronised dialogue file was edited (cut) so that only the portion of the recordings where the subjects are performing the map task (from START to FINISH, illustrated in Figure A.2 as the "Final cut") remained as the official data of the corpus and recordings of the subjects filling in the survey or waiting for the technician to attend to their needs deleted. With the wanted portion defined, the Project file could be used as a gold master that all export variants were made from.

Using this dialogue gold master, various configurations of the same portion of the recording were exported (Table 4.8) to create the media files of the ILMT-s2s corpus.

Details of Step 7: The "Export XML..." feature of Final Cut Pro X was used to retrieve a detailed log of the changes that were applied to the media files and also to calculate all the off-set times for the media files, if needed. For example, the offset time and duration of all the ILMT-s2s System audio files was used to automatically retrieve the segmentation details of "*On-Talk*", or the offset time of the audio recordings from the Mind Media Nexus-4 to calculate which parts of the biosignal data should be deleted, etc.

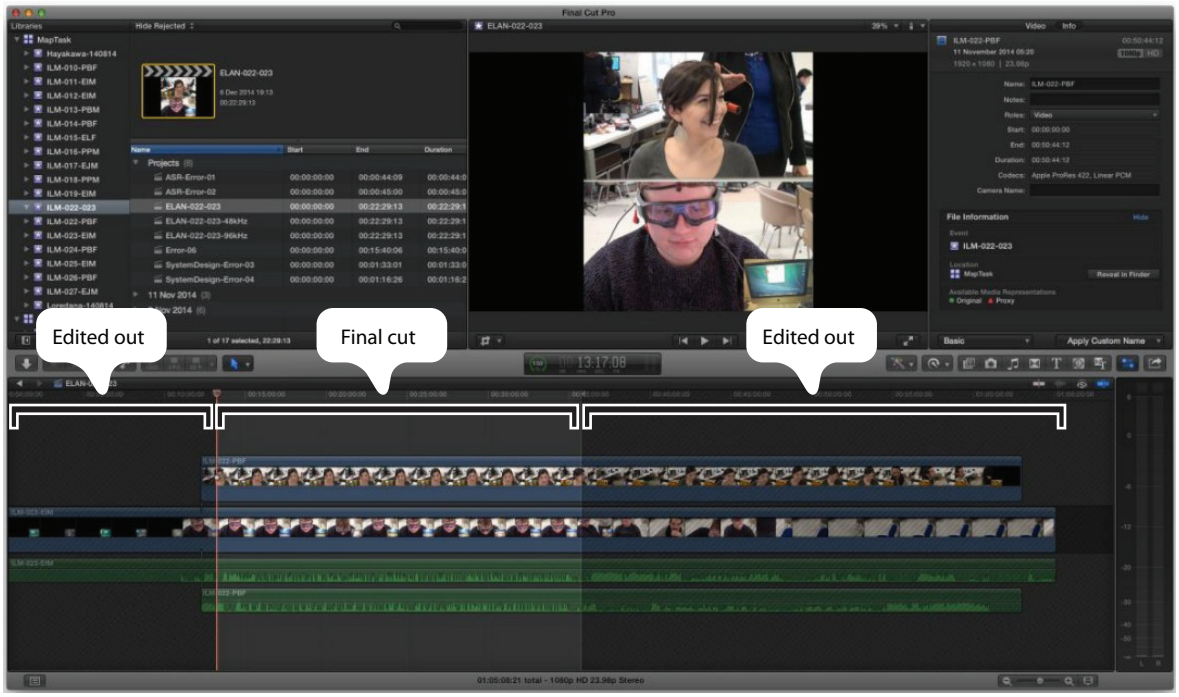


Figure A.2: *Portion of data that was edited out and portion that remained*

The ILMT-s2s System's audio utterance recordings were first synchronised with the video recordings from the SONY HDR-XR500 camcorder. The audio from the camcorder was converted into monaural audio, so the resulting video had one channel with the audio from the camcorder and the other channel with the audio from the ILMT-s2s System. This made it possible to combine and see the audio that was spoken to the other subject and the audio of the whole session in the same audio file, but still keep them separate. This new video file was then synchronised with the video and audio files from the other devices. Once all files from each subject were synchronised, the files from each subject pair were synchronised with each other so that files from both sides of the conversation started as they had done in reality as illustrated in the example of Figure A.2. The synchronised dialogues were then cut between the points where the subjects start and end the map task to remove the portion of the recording before the subjects start the map task and the portion after the subjects complete the map task, as displayed in the lighter grey section of Figure A.2.

A.3 Speech Rate & ASR Precision of All Dialogues

The following pages contain the representation[5] of the dialogues in the 20 sample window framework.

Figure A.3: Grouping by role, spoken language, video setting and crashes during video usage.

Figure A.4: Grouping of gender divided by role and spoken language.

Figure A.5: Grouping of gender divided by video setting and crashes during video usage.

Figure A.6: Grouping of roles divided by gender and spoken language.

Figure A.7: Grouping of roles divided by video setting and crashes during video usage.

Figure A.8: Grouping of spoken languages divided by role and gender.

Figure A.9: Grouping of spoken languages divided by video setting and crashes during video usage.

Figure A.10: Grouping of video setting divided by role and spoken language.

Figure A.11: Grouping of video setting divided by gender and crashes during video usage.

Figure A.12: Grouping of dialogues with systems crashes divided by role and spoken language.

Figure A.13: Grouping of dialogues with systems crashes divided by gender.

Figure A.14: Individual dialogue figures — Dialogues 1 to 4.

Figure A.15: Individual dialogue figures — Dialogues 5 to 8.

Figure A.16: Individual dialogue figures — Dialogues 9 to 12.

Figure A.17: Individual dialogue figures — Dialogues 13 to 15.

[5]For the "Reverse F-Score", 0% indicates that there are no errors in the ASR result and the higher the value, the higher the errors in the ASR result text. For the "Speech Rate", 0% indicates that the subject utterance is the same speed as the TTS output utterance of 180 wpm and the lower the value the slower the subject's utterance is when compared to the TTS output. For the individual dialogues of Figure A.14, Figure A.15, Figure A.16 and Figure A.17 black text denotes male subjects and red text denotes female speakers.

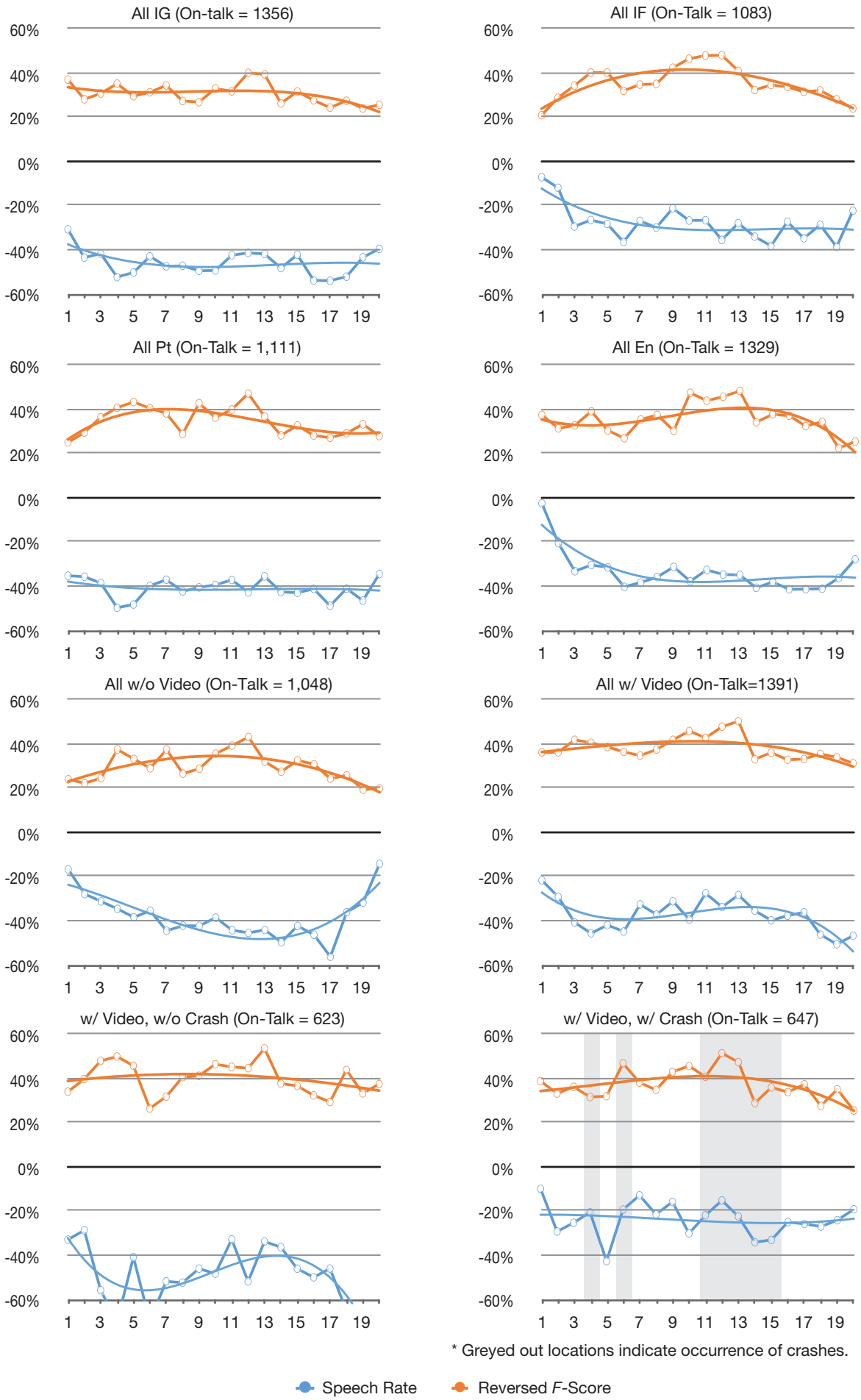


Figure A.3: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Different Grouping*

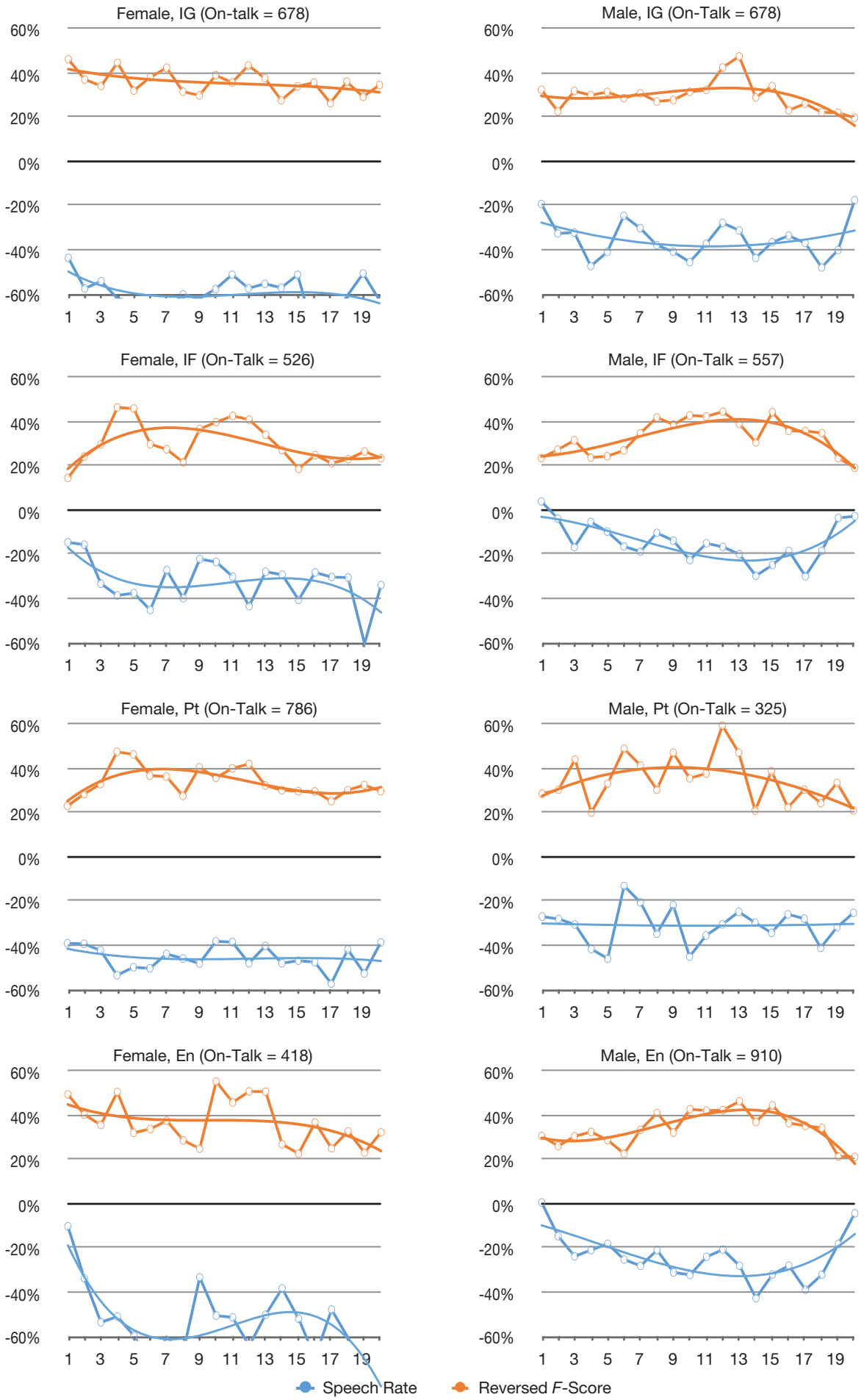


Figure A.4: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Gender 01*

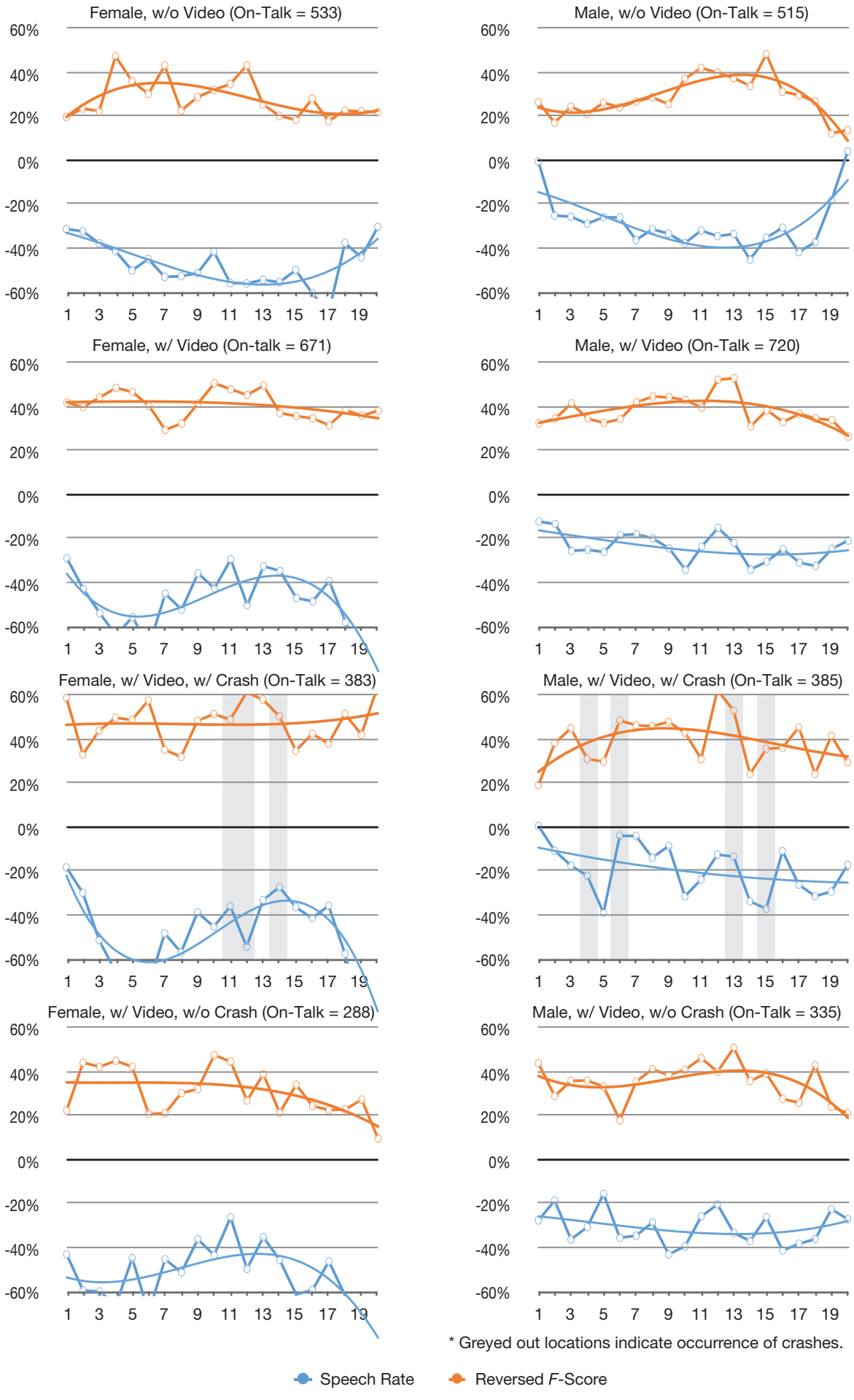


Figure A.5: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Gender 02*

Figure A.6: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Role 01*

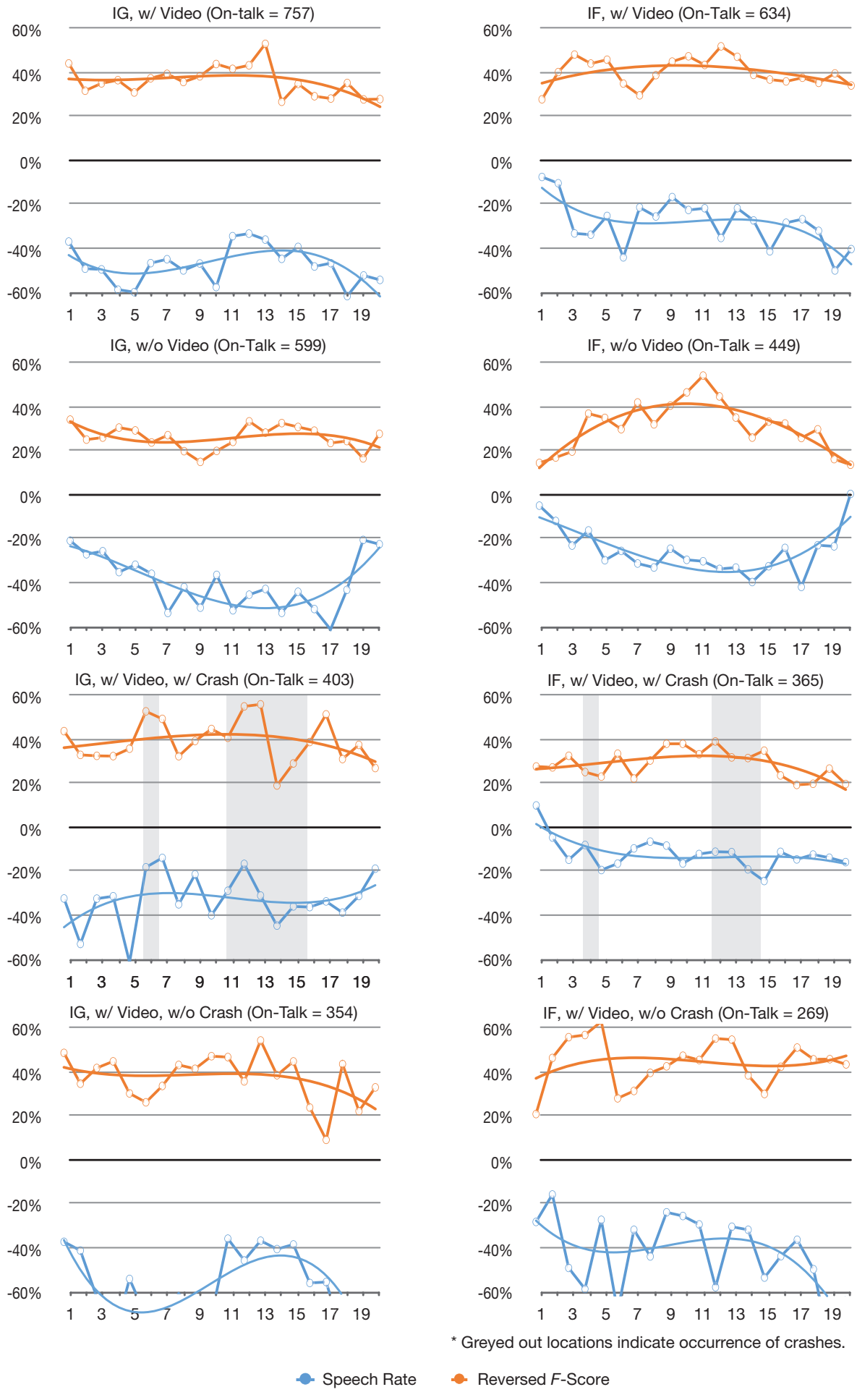


Figure A.7: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Role 02*

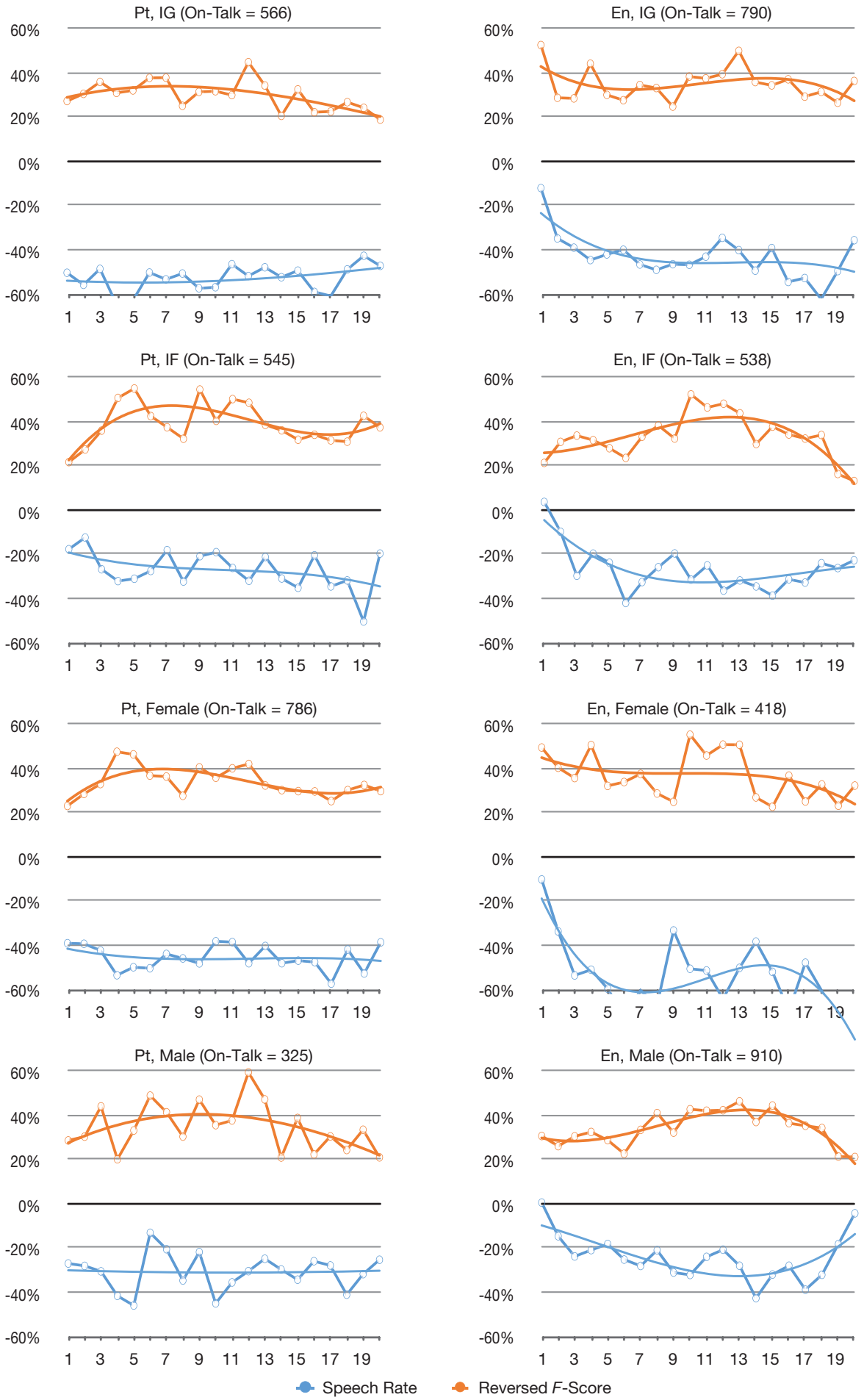


Figure A.8: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Spoken Language 01*

Figure A.9: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues by Spoken Language 02*

Figure A.10: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues without & with video 01*

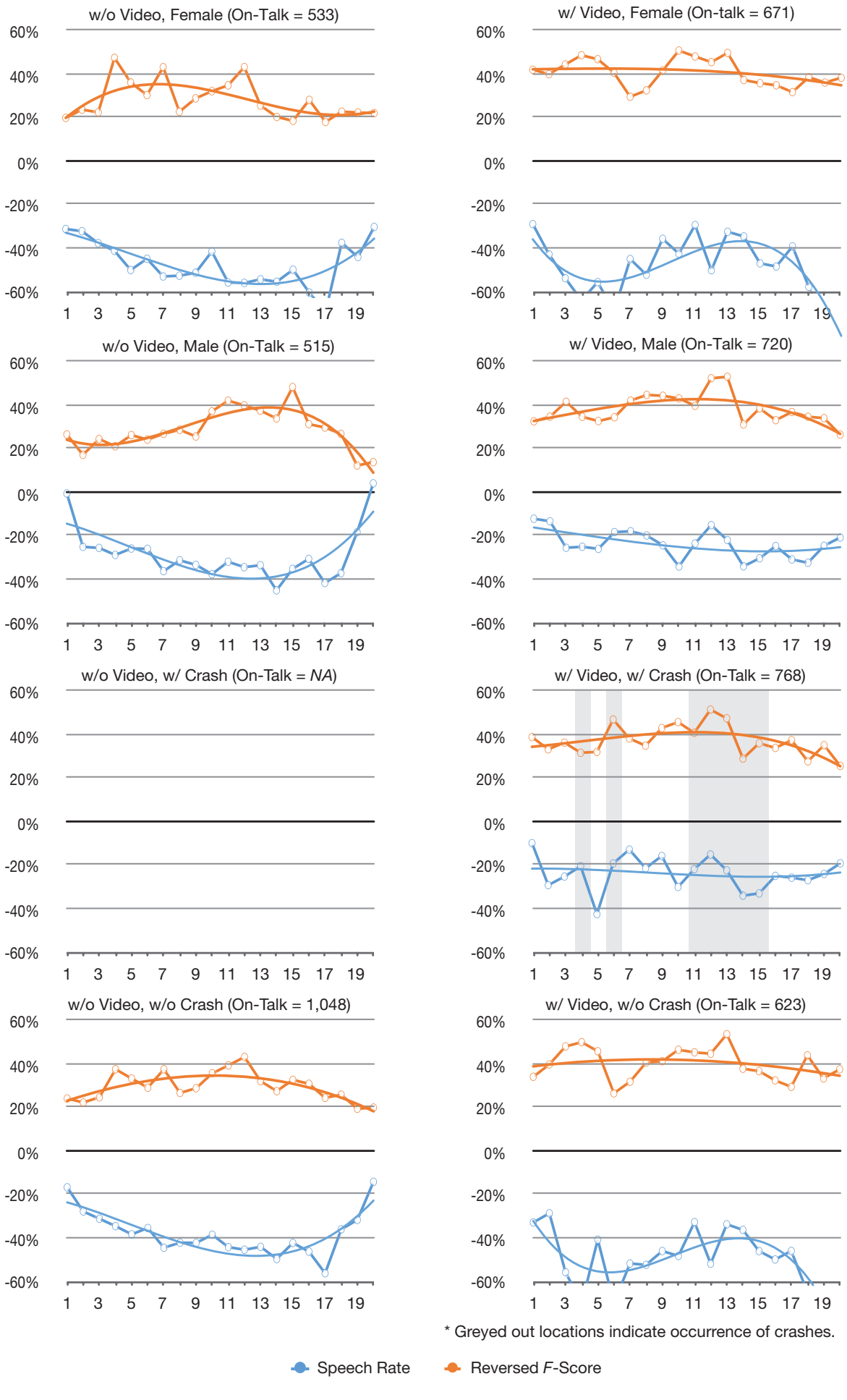


Figure A.11: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues without & with video 02*

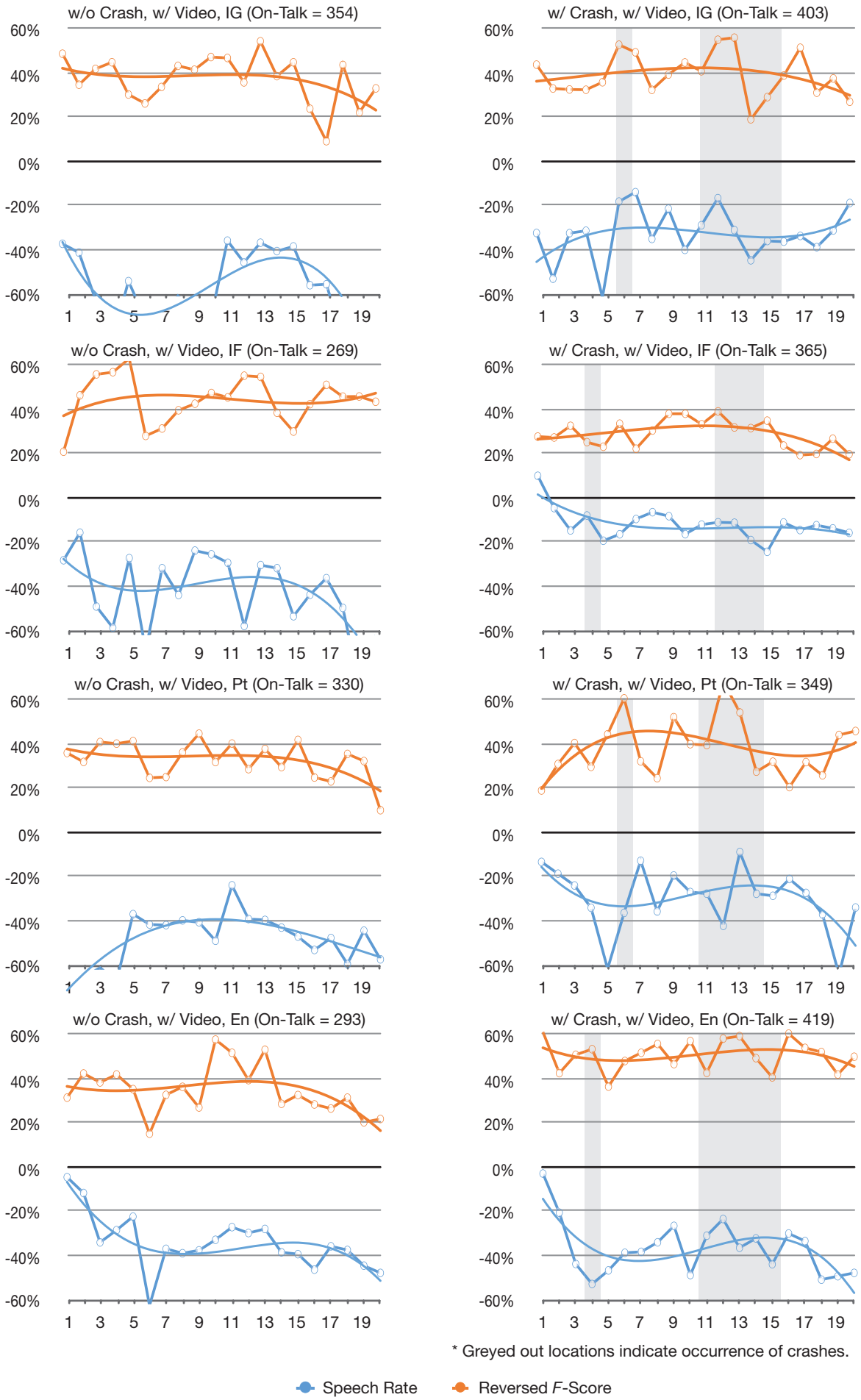


Figure A.12: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues w/ Video that crash 01*

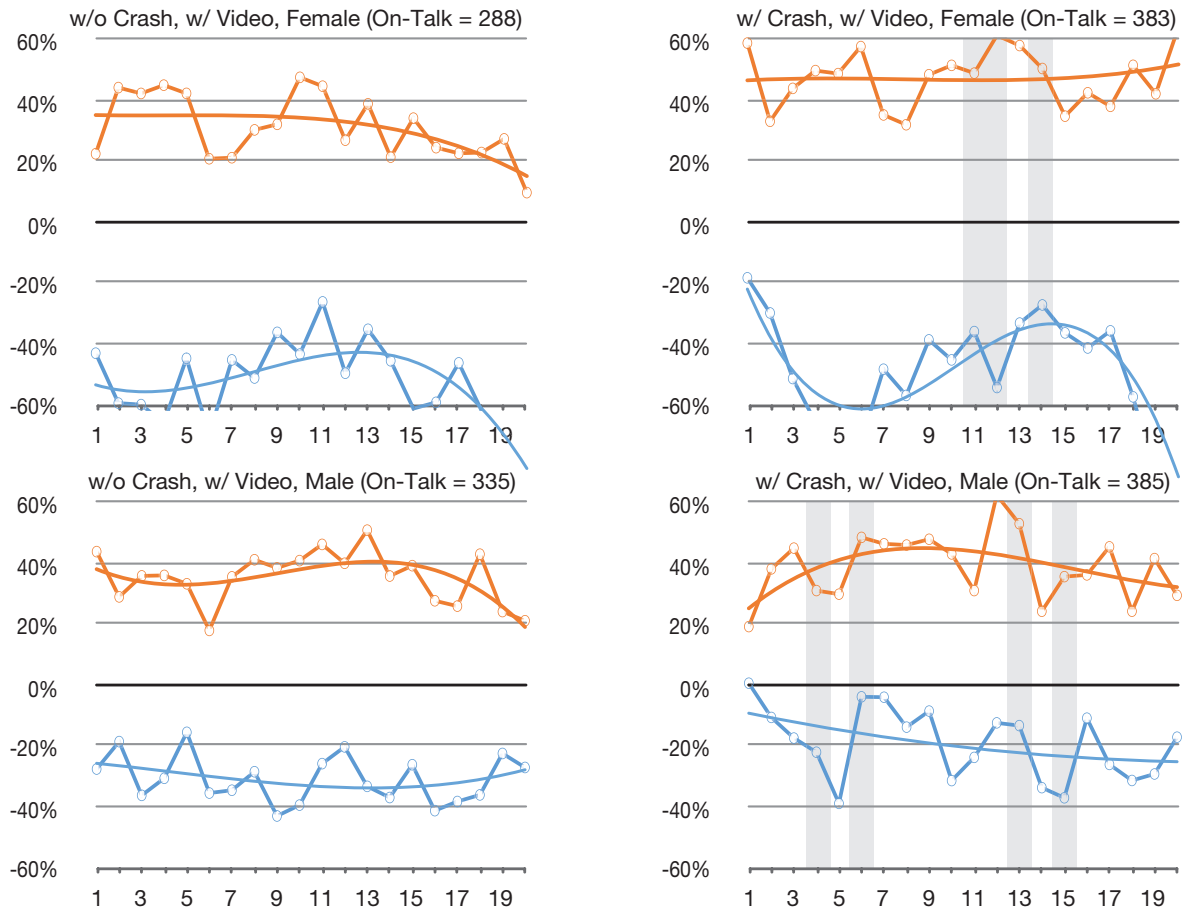


Figure A.13: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues w/ Video that crash 02*

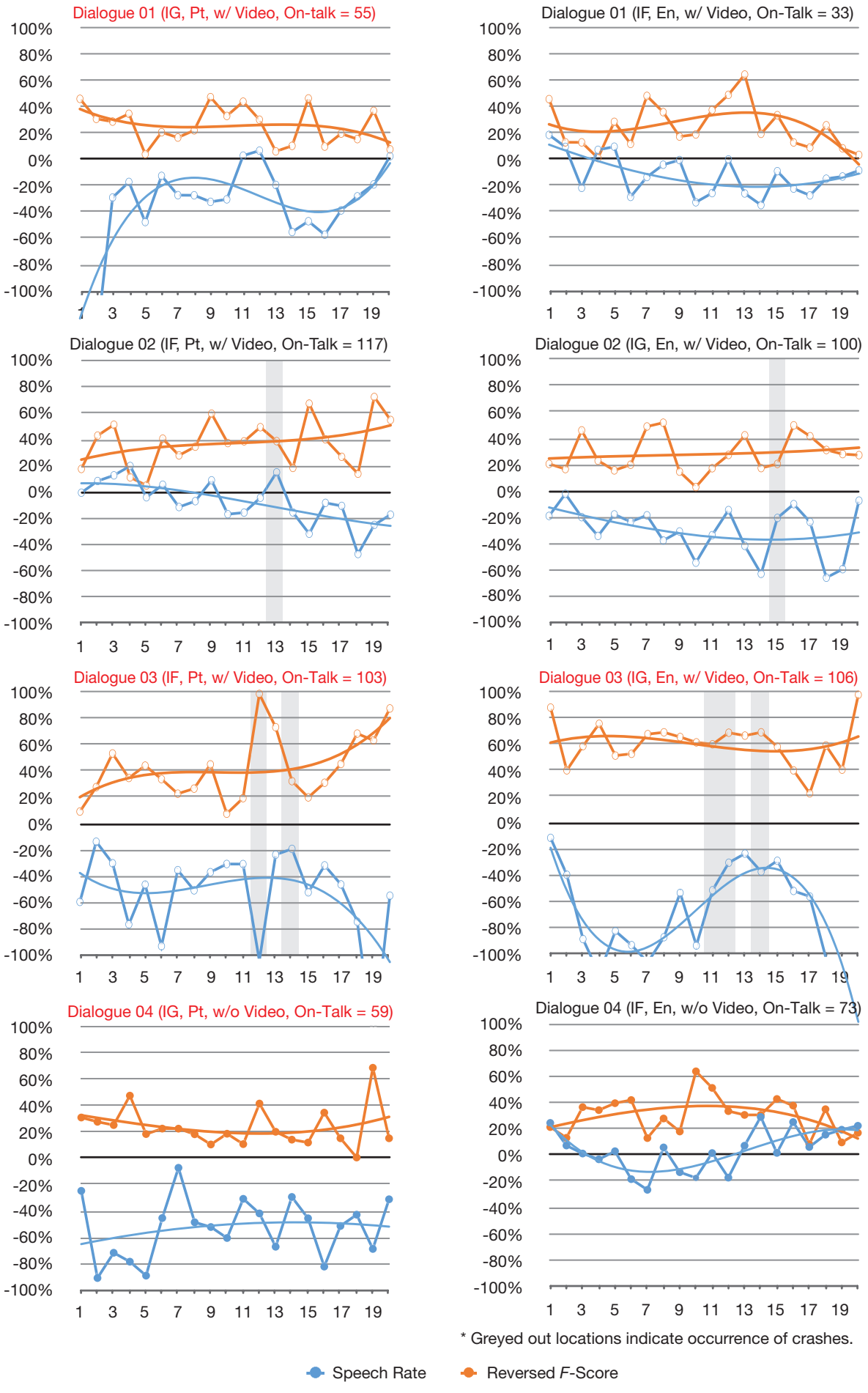


Figure A.14: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 1 to 4*

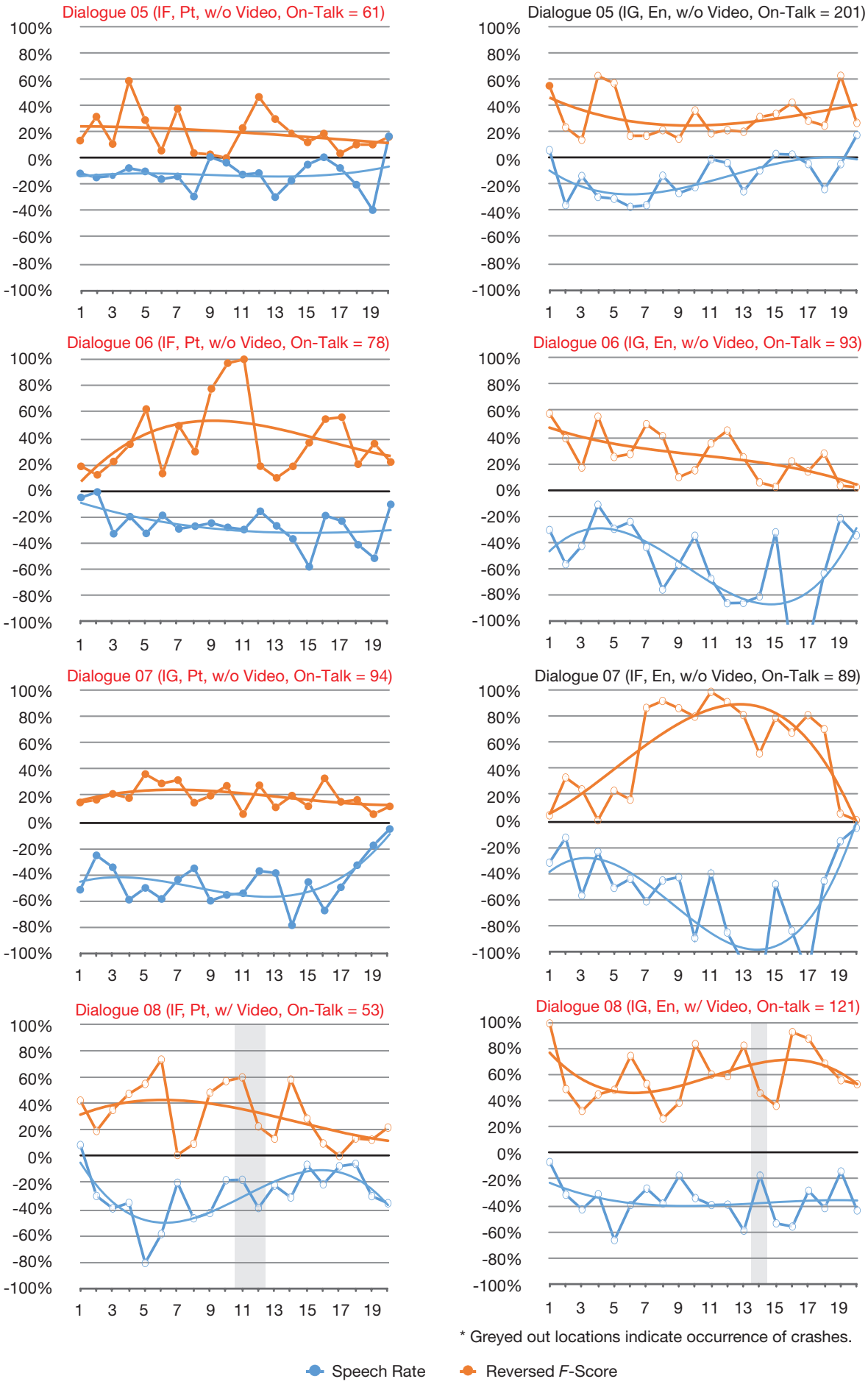


Figure A.15: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 5 to 8*

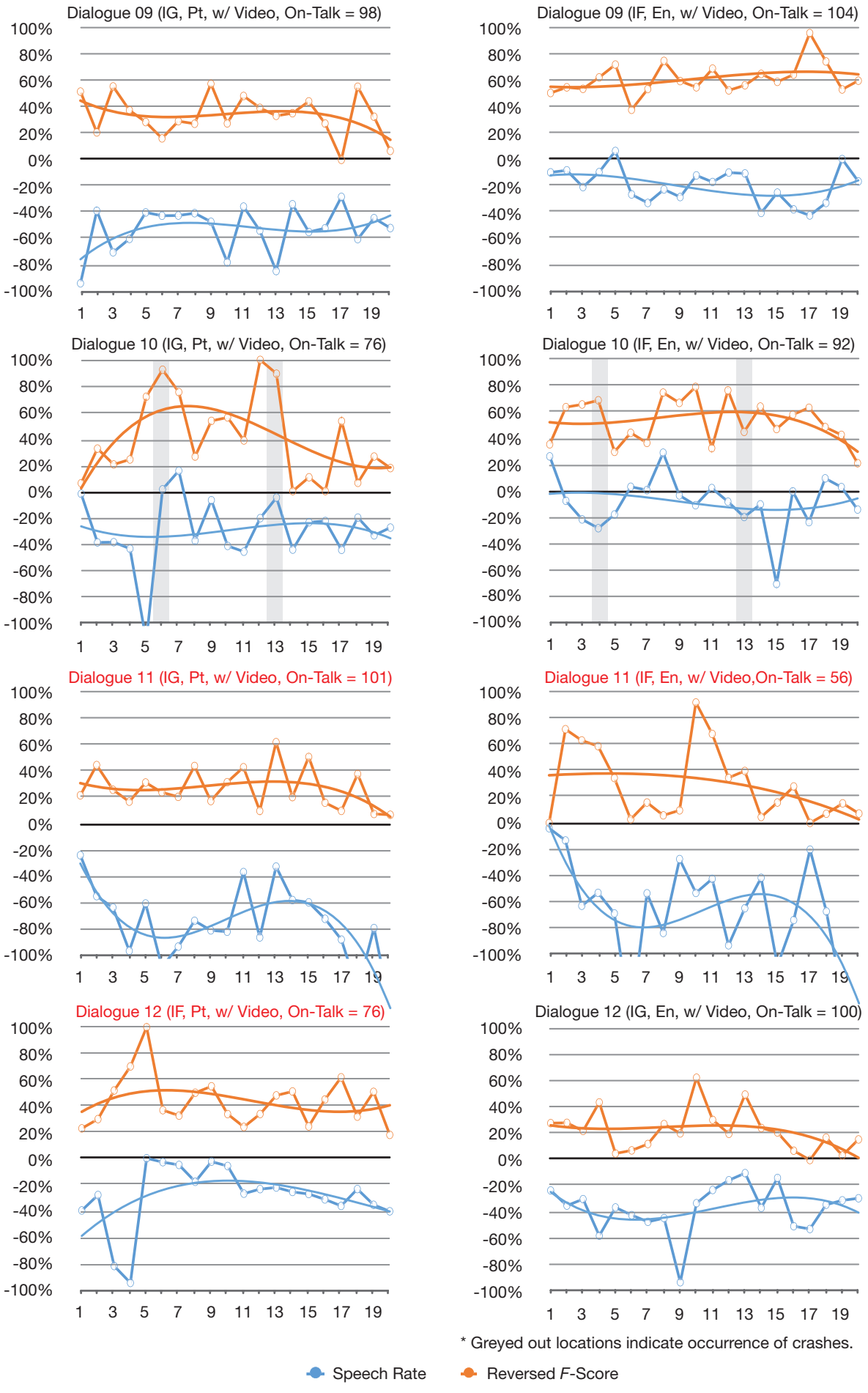


Figure A.16: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 9 to 12*

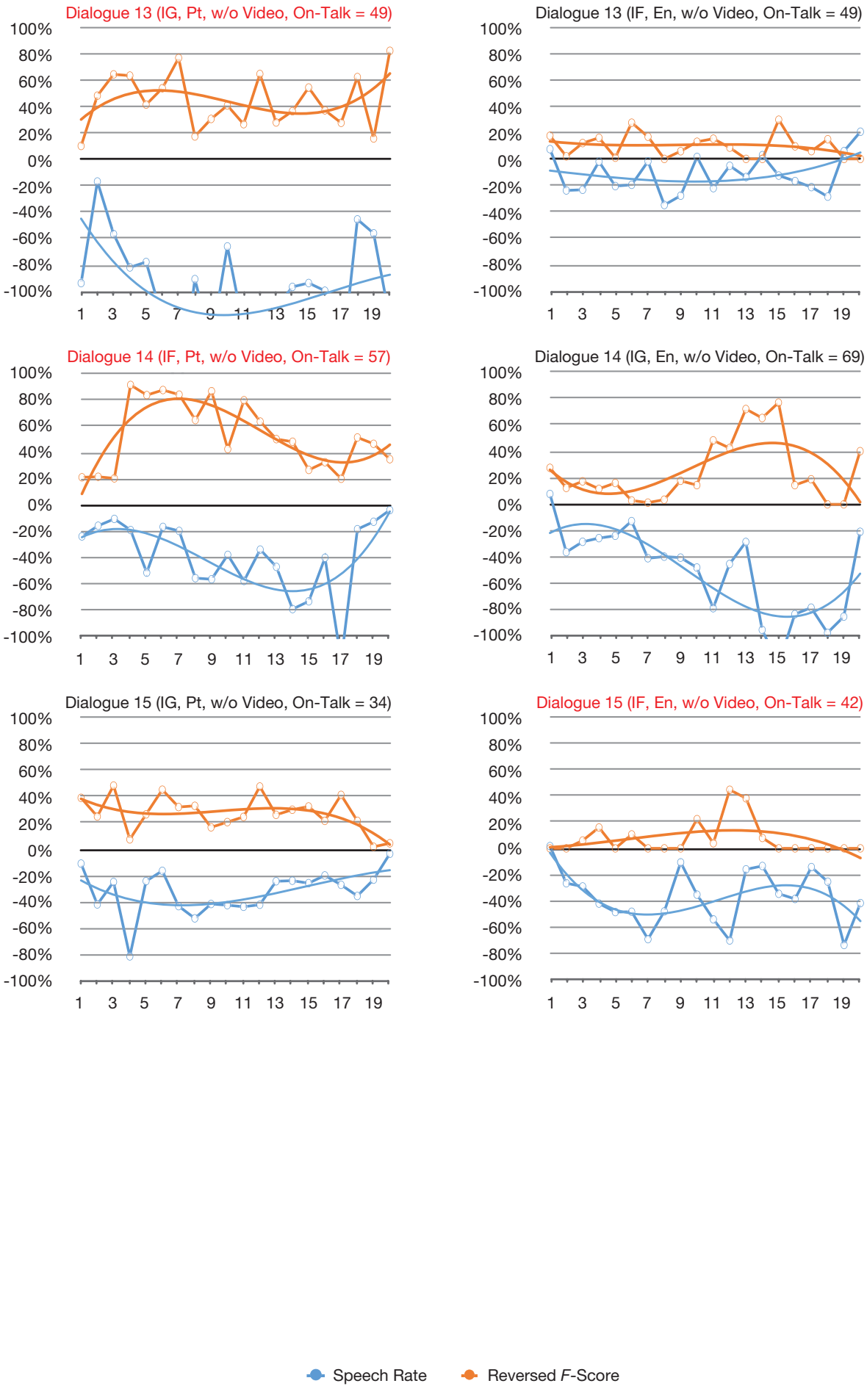


Figure A.17: *Plots of subject's Speech Rate and Reverse F-Score — Dialogues 13 to 15*

A.4 Speech Rate of The Three Talk Types

Table A.4: The results of a significance test and the effect size between *On-Talk* speech rates and *Off-Talk* speech rates for all subjects are listed — "Dialog", "Role", and "t.test" items that are highlighted in red are items with a significant difference whereas "Cohen's d" items highlighted red if the "t.test" results are significantly different and the effect size is "medium" or "large". Red "(%)" items are indications that there are more than 10% of *Off-Talk* utterances in the whole dialogue.

Table A.5: The results of an ANOVA significance test between *On-Talk* speech rates, *Off-Talk Self* speech rates, and *Off-Talk Other* speech rates of all subjects are listed — "Dialog", "Role", and "ANOVA" items that are highlighted in red are items with a significant difference. Red "(%)" items are indications that there are more than 10% of *Off-Talk Self* or *Off-Talk Other* utterances in the whole dialogue.

Table A.6: The results of Post-hoc comparisons and effect size between *On-Talk* speech rates, *Off-Talk Self* speech rates, and *Off-Talk Other* speech rates of all subjects are listed — "Dialog", "Role", and "t.test" items that are highlighted in red are items with a significant difference whereas "Cohen's d" items highlighted red if the "Post-hoc Tukey HSD test" results are significantly different and the effect size is "medium" or "large".

Table A.7: A summary and standard deviation of the speech rate of 1.) all *On-Talk* and *Off-Talk* utterances, 2.) only single word count *On-Talk* and *Off-Talk* utterances, and 3.) all multiple word *On-Talk* and *Off-Talk* utterances of all thirty subjects. The values of this table are used to create Figure 6.3.

Table A.8: A summary and standard deviation of the speech rate of 1.) all *On-Talk*, *Off-Talk Self* and *Off-Talk Other* utterances, 2.) only single word count *On-Talk*, *Off-Talk Self* and *Off-Talk Other* utterances, and 3.) all multiple word *On-Talk*, *Off-Talk Self* and *Off-Talk Other* utterances of all thirty subjects. The values of this table are used to create Figure 6.3.

Table A.9: The results of a significance test and effect size between *On-Talk* speech rates and *Off-Talk* speech rates of only single word utterances for all subjects — "Dialog", "Role", and "t.test" items that are highlighted in red are items with a significant difference whereas "Cohen's d" items highlighted red if the "t.test" results are significantly different and the effect size is "medium" or "large". Red "(%)" items are indications that there are more than 10% of *Off-Talk* utterances in the whole dialogue.

Table A.10: The results of a significance test and effect size between *On-Talk* speech rates and *Off-Talk* speech rates of all multiple word utterances for all subjects are listed — "Dialog", "Role", and "t.test" items that are highlighted in red are items with a significant difference whereas "Cohen's d" items highlighted red if the "t.test" results are significantly different and the effect

size is “medium” or “large”. Red “(%)” items are indications that there are more than 10% of *Off-Talk* utterances in the whole dialogue.

Table A.11: The results of an ANOVA significance test between *On-Talk* speech rates, *Off-Talk Self* speech rates, and *Off-Talk Other* speech rates of all multiple word utterances for all subjects are listed — “Dialog”, “Role”, and “ANOVA” items that are highlighted in red are items with a significant difference. Red “(%)” items are indications that there are more than 10% of *Off-Talk Self* or *Off-Talk Other* utterances in the whole dialogue.

Table A.12: The results of Post-hoc comparisons and effect size between *On-Talk* speech rates, *Off-Talk Self* speech rates, and *Off-Talk Other* speech rates of all multiple word utterances for all subjects are listed — “Dialog”, “Role”, and “t.test” items that are highlighted in red are items with a significant difference whereas “Cohen’s d” items highlighted red if the “Post-hoc Tukey HSD test” results are significantly different and the effect size is “medium” or “large”.

Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen’s d	ES
D.01	IG - Pt (♀)	55 : 6 (9.8%)	6 (9.8%)	– –	$p < 0.0106$	*	$d = 0.8577$	large
D.01	IF - En (♂)	33 : 25 (43.1%)	25 (43.1%)	– –	$p < 0.0001$	***	$d = 1.4665$	large
D.02	IG - En (♂)	100 : 59 (37.1%)	55 (34.6%)	4 (2.5%)	$p < 0.0209$	*	$d = 0.4566$	small
D.02	IF - Pt (♂)	117 : 45 (27.8%)	31 (19.1%)	14 (8.6%)	$p < 0.0548$	.	$d = 0.4975$	small
D.03	IG - En (♀)	106 : 81 (43.3%)	9 (4.8%)	72 (38.5%)	$p < 0.0000$	***	$d = 1.3018$	large
D.03	IF - Pt (♀)	103 : 116 (53.0%)	59 (26.9%)	57 (26.0%)	$p < 0.0974$	.	$d = 0.2194$	small
D.04	IG - Pt (♀)	59 : 33 (35.9%)	28 (30.4%)	5 (5.4%)	$p < 0.9658$	–	$d = 0.0113$	neg
D.04	IF - En (♂)	73 : 46 (38.7%)	40 (33.6%)	6 (5.0%)	$p < 0.1711$	–	$d = 0.3058$	small
D.05	IG - En (♂)	201 : 2 (1.0%)	2 (1.0%)	– –	$p < 0.0343$	*	$d = 2.9422$	large
D.05	IF - Pt (♀)	61 : 0 –	– –	– –	–	–	–	
D.06	IG - En (♀)	93 : 100 (51.8%)	91 (47.2%)	9 (4.7%)	$p < 0.0136$	*	$d = 0.3522$	small
D.06	IF - Pt (♀)	78 : 44 (36.1%)	44 (36.1%)	– –	$p < 0.0041$	**	$d = 0.7591$	med
D.07	IG - Pt (♀)	94 : 4 (4.1%)	– –	4 (4.1%)	$p < 0.1242$	–	$d = 0.6444$	med
D.07	IF - En (♂)	89 : 7 (7.3%)	6 (6.3%)	1 (1.0%)	$p < 0.0000$	***	$d = 1.3307$	large
D.08	IG - En (♀)	121 : 41 (25.3%)	11 (6.8%)	30 (18.5%)	$p < 0.0000$	***	$d = 1.0733$	large
D.08	IF - Pt (♀)	53 : 87 (62.1%)	30 (21.4%)	57 (40.7%)	$p < 0.2999$	–	$d = 0.1540$	neg
D.09	IG - Pt (♂)	98 : 5 (4.9%)	5 (4.9%)	– –	$p < 0.1685$	–	$d = 2.3886$	large
D.09	IF - En (♂)	104 : 24 (18.8%)	24 (18.8%)	– –	$p < 0.1914$	–	$d = 0.3951$	small
D.10	IG - Pt (♂)	76 : 52 (40.6%)	21 (16.4%)	31 (24.2%)	$p < 0.3571$	–	$d = 0.1932$	neg
D.10	IF - En (♂)	92 : 78 (45.9%)	46 (27.1%)	32 (18.8%)	$p < 0.6776$	–	$d = 0.0677$	neg
D.11	IG - Pt (♀)	101 : 2 (1.9%)	2 (1.9%)	– –	$p < 0.2964$	–	$d = 4.8509$	large
D.11	IF - En (♀)	56 : 73 (56.6%)	59 (45.7%)	14 (10.9%)	$p < 0.0000$	***	$d = 1.0413$	large
D.12	IG - En (♂)	100 : 54 (35.1%)	54 (35.1%)	– –	$p < 0.3042$	–	$d = 0.2263$	small
D.12	IF - Pt (♀)	76 : 12 (13.6%)	10 (11.4%)	2 (2.3%)	$p < 0.0605$	.	$d = 1.5212$	large
D.13	IG - Pt (♀)	49 : 106 (68.4%)	106 (68.4%)	– –	$p < 0.0000$	***	$d = 0.7862$	med
D.13	IF - En (♂)	49 : 2 (3.9%)	2 (3.9%)	– –	$p < 0.0414$	*	$d = 0.8022$	large
D.14	IG - En (♂)	69 : 59 (46.1%)	59 (46.1%)	– –	$p < 0.0000$	***	$d = 0.9758$	large
D.14	IF - Pt (♀)	57 : 10 (14.9%)	10 (14.9%)	– –	$p < 0.1606$	–	$d = 1.2068$	small
D.15	IG - Pt (♂)	34 : 15 (30.6%)	13 (26.5%)	2 (4.1%)	$p < 0.9102$		$d = 0.0495$	neg
D.15	IF - En (♀)	42 : 1 (2.3%)	– –	1 (2.3%)	–	–	–	–

Table A.4: *Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (all utterances)*

Dialog	Role - Lang	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA	sig.
D.01	IG - Pt (♀)	55	6 (9.8%)	– –	$p < 0.0507$	.
D.01	IF - En (♂)	33	25 (43.1%)	– –	$p < 0.0000$	***
D.02	IG - En (♂)	100	55 (34.6%)	4 (2.5%)	$p < 0.0000$	***
D.02	IF - Pt (♂)	117	31 (19.1%)	14 (8.6%)	$p < 0.0166$	*
D.03	IG - En (♀)	106	9 (4.8%)	72 (38.5%)	$p < 0.0000$	***
D.03	IF - Pt (♀)	103	59 (26.9%)	57 (26.0%)	$p < 0.0001$	***
D.04	IG - Pt (♀)	59	28 (30.4%)	5 (5.4%)	$p < 0.0013$	***
D.04	IF - En (♂)	73	40 (33.6%)	6 (5.0%)	$p < 0.0633$	.
D.05	IG - En (♂)	201	2 (1.0%)	– –	$p < 0.0001$	***
D.05	IF - Pt (♀)	61	– –	– –	–	
D.06	IG - En (♀)	93	91 (47.2%)	9 (4.7%)	$p < 0.0168$	*
D.06	IF - Pt (♀)	78	44 (36.1%)	– –	$p < 0.0001$	***
D.07	IG - Pt (♀)	94	– –	4 (4.1%)	$p < 0.2100$	–
D.07	IF - En (♂)	89	6 (6.3%)	1 (1.0%)	$p < 0.0044$	**
D.08	IG - En (♀)	121	11 (6.8%)	30 (18.5%)	$p < 0.0000$	***
D.08	IF - Pt (♀)	53	30 (21.4%)	57 (40.7%)	$p < 0.0003$	***
D.09	IG - Pt (♂)	98	5 (4.9%)	– –	$p < 0.0000$	***
D.09	IF - En (♂)	104	24 (18.8%)	– –	$p < 0.0835$	.
D.10	IG - Pt (♂)	76	21 (16.4%)	31 (24.2%)	$p < 0.0000$	***
D.10	IF - En (♂)	92	46 (27.1%)	32 (18.8%)	$p < 0.0392$	*
D.11	IG - Pt (♀)	101	2 (1.9%)	– –	$p < 0.0000$	***
D.11	IF - En (♀)	56	59 (45.7%)	14 (10.9%)	$p < 0.0000$	***
D.12	IG - En (♂)	100	54 (35.1%)	– –	$p < 0.1820$	–
D.12	IF - Pt (♀)	76	10 (11.4%)	2 (2.3%)	$p < 0.0000$	***
D.13	IG - Pt (♀)	49	106 (68.4%)	– –	$p < 0.0000$	***
D.13	IF - En (♂)	49	2 (3.9%)	– –	$p < 0.2720$	–
D.14	IG - En (♂)	69	59 (46.1%)	– –	$p < 0.0000$	***
D.14	IF - Pt (♀)	57	10 (14.9%)	– –	$p < 0.0008$	***
D.15	IG - Pt (♂)	34	13 (26.5%)	2 (4.1%)	$p < 0.1810$	–
D.15	IF - En (♀)	42	– –	1 (2.3%)	$p < 0.7910$	–

Table A.5: *ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)*

Dial. - Role	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
	On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
D.01 - IG	0.0507 .	–	–	$d = 0.8577$, large	–	–
D.01 - IF	0.0000 ***	–	–	$d = 1.4665$, large	–	–
D.02 - IG	0.0011 **	0.0204 *	0.0005 ***	$d = 0.6061$, med	$d = 2.0031$, large	$d = 1.4491$, large
D.02 - IF	0.0172 *	0.4174	0.8057	$d = 0.5919$, med	$d = 0.4938$, small	$d = 0.1208$, neg
D.03 - IG	0.0023 **	0.0000 ***	0.9279	$d = 1.1309$, large	$d = 1.3156$, large	$d = 0.1394$, neg
D.03 - IF	0.0010 **	0.6279	0.0002 ***	$d = 0.5781$, med	$d = 0.1864$, neg	$d = 0.6314$, med
D.04 - IG	0.4161	0.0036 **	0.0008 ***	$d = 0.2849$, small	$d = 2.2399$, large	$d = 1.3315$, large
D.04 - IF	0.0633 .	0.1011	0.6921	$d = 0.4049$, small	$d = 0.5454$, med	$d = 0.5316$, med
D.05 - IG	0.0001 ***	–	–	$d = 2.9422$, large	–	–
D.05 - IF	–	–	–	–	–	–
D.06 - IG	0.0985 .	0.0453 *	0.2813	$d = 0.3004$, small	$d = 1.3270$, large	$d = 0.4282$, small
D.06 - IF	0.0001 ***	–	–	$d = 0.7591$, med	–	–
D.07 - IG	–	0.2100	–	–	$d = 0.6444$, med	–
D.07 - IF	0.0091 **	0.2427	0.9424	$d = 1.2737$, large	–	–
D.08 - IG	0.7356	0.0000 ***	0.0026 *	$d = 0.2217$, small	$d = 1.6068$, large	$d = 1.0144$, large
D.08 - IF	0.0029 **	0.6949	0.0002 ***	$d = 0.7055$, med	$d = 0.2117$, small	$d = 0.7752$, med
D.09 - IG	0.0000 ***	–	–	$d = 2.3886$, large	–	–
D.09 - IF	0.0835 .	–	–	$d = 0.3951$, small	–	–
D.10 - IG	0.0000 ***	0.0247 *	0.0000 ***	$d = 1.2639$, large	$d = 0.9367$, large	$d = 1.4095$, large
D.10 - IF	0.6139	0.1132	0.0328 *	$d = 1.2639$, large	$d = 0.7130$, med	$d = 1.4095$, large
D.11 - IG	0.0000 ***	–	–	$d = 4.8509$, large	–	–
D.11 - IF	0.0000 ***	0.0000 ***	0.3604	$d = 0.9345$, large	$d = 1.2344$, large	$d = 0.5009$, med
D.12 - IG	0.1822	–	–	$d = 0.2263$, small	–	–
D.12 - IF	0.0000 ***	0.9051	0.0120 *	$d = 1.9544$, large	$d = 0.6287$, med	$d = 0.8757$, large
D.13 - IG	0.0000 ***	–	–	$d = 0.7862$, med	–	–
D.13 - IF	0.2716	–	–	$d = 0.8022$, large	–	–
D.14 - IG	0.0000 ***	–	–	$d = 0.9758$, large	–	–
D.14 - IF	0.0008 ***	–	–	$d = 1.2068$, large	–	–
D.15 - IG	0.7420	0.2445	0.1571	$d = 0.2384$, small	$d = 2.1540$, large	$d = 0.8586$, large
D.15 - IF	–	0.7912	–	–	–	–

Table A.6: *Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (all utterances)*

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
On-Talk — All utterances	−383.90	−58.94	−28.31	−36.79	−4.82	58.28	47.15
On-Talk — 1 word utterances	−328.00	−43.48	−13.92	−28.44	7.98	51.90	59.30
On-Talk — 2 ≤ word utterances	−383.90	−60.50	−31.10	−38.17	−7.32	58.28	44.69
Off-Talk — All utterances	−1,539.00	−57.66	−8.14	−40.24	15.73	83.30	110.51
Off-Talk — 1 word utterances	−1,539.00	−114.60	−30.07	−73.29	11.59	83.30	145.65
Off-Talk — 2 ≤ word utterances	−443.50	−27.79	0.65	−10.81	18.43	71.48	48.75

Table A.7: *Summary and standard deviation of 2 talk type all word, single word, and multi word utterance speech rates of all thirty subjects.*

	Min.	1st Qu.	*Mdn*	*M*	3rd Qu.	Max.	*SD*
On-Talk — All utterances	−383.90	−58.94	−28.31	−36.79	−4.82	58.28	47.15
On-Talk — 1 word utterances	−328.00	−43.48	−13.92	−28.44	7.98	51.90	59.30
On-Talk — 2 ≤ word utterances	−383.90	−60.50	−31.10	−38.17	−7.32	58.28	44.69
Off-Talk Self — All utterances	−1539.00	−74.21	−18.66	−53.23	11.51	83.30	123.93
Off-Talk Self — 1 word utterances	−1539.00	−126.90	−39.68	−83.80	6.46	83.30	155.17
Off-Talk Self — 2 ≤ word utterances	−443.50	−38.42	−6.33	−18.17	13.70	65.48	55.59
Off-Talk Other— All utterances	−433.40	−20.60	5.01	−7.93	21.84	74.29	54.23
Off-Talk Other — 1 word utterances	−433.40	−68.77	1.42	−28.82	21.62	74.29	82.11
Off-Talk Other — 2 ≤ word utterances	−139.40	−13.85	7.35	1.62	21.95	71.48	30.52

Table A.8: *Summary and standard deviation of 3 talk type all word, single word, and multi word utterance speech rates of all thirty subjects.*

Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen's d	ES
D.01	IG - Pt (♀)	8 : 3 (27.3%)	3 (27.3%)	– –	$p < 0.3419$	–	$d = 0.1354$	neg
D.01	IF - En (♂)	1 : 23 (95.8%)	23 (95.8%)	– –	–	–	–	
D.02	IG - En (♂)	21 : 48 (69.6%)	46 (66.7%)	2 (2.9%)	$p < 0.0000$	***	$d = 0.8793$	large
D.02	IF - Pt (♂)	9 : 22 (71.0%)	17 (54.8%)	5 (16.1%)	$p < 0.2144$	–	$d = 0.3273$	small
D.03	IG - En (♀)	20 : 33 (62.3%)	5 (9.4%)	28 (52.8%)	$p < 0.0668$	.	$d = 0.5307$	med
D.03	IF - Pt (♀)	18 : 53 (74.6%)	38 (53.5%)	15 (21.1%)	$p < 0.3606$	–	$d = 0.2080$	small
D.04	IG - Pt (♀)	11 : 20 (64.5%)	20 (64.5%)	–	$p < 0.3287$	–	$d = 0.3063$	small
D.04	IF - En (♂)	23 : 23 (50.0%)	21 (45.7%)	2 (4.3%)	$p < 0.0056$	**	$d = 0.8892$	large
D.05	IG - En (♂)	38 : 2 (5.0%)	2 (5.0%)	– –	$p < 0.0044$	**	$d = 2.9604$	large
D.05	IF - Pt (♀)	3 : 0 –	– –	– –	–	–	–	
D.06	IG - En (♀)	4 : 33 (89.2%)	32 (86.5%)	1 (2.7%)	$p < 0.0143$	*	$d = 0.5237$	med
D.06	IF - Pt (♀)	14 : 18 (56.3%)	18 (56.3%)	– –	$p < 0.0058$	**	$d = 0.9882$	large
D.07	IG - Pt (♀)	1 : 1 (50.0%)	– –	1 (50.0%)	–	–	–	
D.07	IF - En (♂)	9 : 0 –	– –	– –	–	–	–	
D.08	IG - En (♀)	17 : 18 (51.4%)	10 (28.6%)	8 (22.9%)	$p < 0.2962$	–	$d = 0.3553$	small
D.08	IF - Pt (♀)	2 : 35 (94.6%)	16 (43.2%)	19 (51.4%)	$p < 0.4488$	–	$d = 0.4045$	small
D.09	IG - Pt (♂)	4 : 4 (50.0%)	4 (50.0%)	– –	$p < 0.2269$	–	$d = 1.0086$	large
D.09	IF - En (♂)	9 : 11 (55.0%)	11 (55.0%)	– –	$p < 0.0022$	**	$d = 1.5692$	large
D.10	IG - Pt (♂)	20 : 25 (55.6%)	17 (37.8%)	8 (17.8%)	$p < 0.0460$	*	$d = 0.5716$	med
D.10	IF - En (♂)	23 : 28 (54.9%)	17 (33.3%)	11 (21.6%)	$p < 0.1205$	–	$d = 0.4128$	small
D.11	IG - Pt (♀)	14 : 2 (12.5%)	2 (12.5%)	– –	$p < 0.2586$	–	$d = 3.3609$	large
D.11	IF - En (♀)	8 : 22 (73.3%)	17 (56.7%)	5 (16.7%)	$p < 0.2328$	–	$d = 0.6630$	med
D.12	IG - En (♂)	14 : 42 (75.0%)	42 (75.0%)	– –	$p < 0.5028$	–	$d = 0.1354$	neg
D.12	IF - Pt (♀)	8 : 9 (52.9%)	8 (47.1%)	1 (5.9%)	$p < 0.1227$	–	$d = 0.7786$	med
D.13	IG - Pt (♀)	6 : 36 (85.7%)	36 (85.7%)	– –	$p < 0.8116$	–	$d = 0.0981$	neg
D.13	IF - En (♂)	3 : 0 –	– –	– –	–	–	–	
D.14	IG - En (♂)	10 : 30 (75.0%)	30 (75.0%)	– –	$p < 0.0060$	**	$d = 1.4063$	large
D.14	IF - Pt (♀)	20 : 8 (28.6%)	8 (28.6%)	– –	$p < 0.1581$	–	$d = 1.0358$	large
D.15	IG - Pt (♂)	1 : 11 (91.7%)	10 (83.3%)	1 (8.3%)	–	–	–	
D.15	IF - En (♀)	7 : 0 –	– –	– –	–	–	–	
–	–	346 :560 (52.4%)	453 (42.6%	107 (9.8%)	–	–	–	

Table A.9: *Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (only 1 word utterances)*

Dialog	Role - Lang	On T. : Off T. (%)	Self (%)	Other (%)	t.test	sig.	Cohen's d	ES
D.01	IG - Pt (♀)	47 : 3 (6.0%)	3 (6.0%)	– –	$p < 0.0856$	.	$d = 0.7786$	med
D.01	IF - En (♂)	32 : 2 (5.9%)	2 (5.9%)	– –	$p < 0.4909$	–	$d = 0.6810$	med
D.02	IG - En (♂)	79 : 11 (12.2%)	9 (10.0%)	2 (2.2%)	$p < 0.6601$	–	$d = 0.2149$	small
D.02	IF - Pt (♂)	108 : 23 (17.6%)	14 (10.7%)	9 (6.9%)	$p < 0.2682$	–	$d = 0.3221$	small
D.03	IG - En (♀)	86 : 48 (35.8%)	4 (3.0%)	44 (32.8%)	$p < 0.0000$	***	$d = 1.8069$	large
D.03	IF - Pt (♀)	85 : 63 (42.6%)	21 (14.2%)	42 (28.4%)	$p < 0.0169$	*	$d = 0.3998$	small
D.04	IG - Pt (♀)	48 : 13 (21.3%)	8 (13.1%)	5 (8.2%)	$p < 0.0002$	***	$d = 1.3152$	large
D.04	IF - En (♂)	50 : 23 (31.5%)	19 (26.0%)	4 (5.5%)	$p < 0.0542$	.	$d = 0.5551$	med
D.05	IG - En (♂)	163 : 0 –	– –	– –	–	–	–	
D.05	IF - Pt (♀)	58 : 0 –	– –	– –	–	–	–	
D.06	IG - En (♀)	89 : 67 (43.0%)	59 (37.8%)	8 (5.1%)	$p < 0.0000$	***	$d = 0.9390$	large
D.06	IF - Pt (♀)	64 : 26 (28.9%)	26 (28.9%)	– –	$p < 0.0852$	.	$d = 0.5054$	med
D.07	IG - Pt (♀)	93 : 3 (3.1%)	– –	3 (3.1%)	$p < 0.0178$	*	$d = 0.9134$	large
D.07	IF - En (♂)	80 : 7 (8.1%)	6 (6.9%)	1 (1.1%)	$p < 0.0000$	***	$d = 1.5463$	large
D.08	IG - En (♀)	104 : 23 (18.1%)	1 (0.8%)	22 (17.3%)	$p < 0.0000$	***	$d = 1.6702$	large
D.08	IF - Pt (♀)	51 : 52 (50.5%)	14 (13.6%)	38 (36.9%)	$p < 0.0094$	**	$d = 0.5226$	med
D.09	IG - Pt (♂)	94 : 1 (1.1%)	1 (1.1%)	– –	–	–	–	–
D.09	IF - En (♂)	95 : 13 (12.0%)	12 (12.0%)	– –	$p < 0.9090$	–	$d = 0.0528$	neg
D.10	IG - Pt (♂)	56 : 27 (32.5%)	4 (4.8%)	23 (27.7%)	$p < 0.2158$	–	$d = 0.3752$	small
D.10	IF - En (♂)	69 : 50 (42.0%)	29 (24.4%)	21 (17.6%)	$p < 0.0021$	**	$d = 0.6007$	med
D.11	IG - Pt (♀)	87 : 0 –	– –	– –	–	–	–	–
D.11	IF - En (♀)	48 : 51 (51.5%)	42 (42.4%)	9 (9.1%)	$p < 0.0000$	***	$d = 1.2100$	large
D.12	IG - En (♂)	86 : 12 (12.2%)	12 (12.2%)	– –	$p < 0.0004$	***	$d = 1.4567$	large
D.12	IF - Pt (♀)	68 : 3 (4.2%)	2 (2.8%)	1 (1.4%)	$p < 0.6478$	–	$d = 0.4365$	small
D.13	IG - Pt (♀)	43 : 70 (61.9%)	70 (61.9%)	– –	$p < 0.0000$	***	$d = 1.1616$	large
D.13	IF - En (♂)	46 : 2 (4.2%)	2 (4.2%)	– –	$p < 0.0571$	.	$d = 1.0330$	large
D.14	IG - En (♂)	59 : 29 (33.0%)	29 (33.0%)	– –	$p < 0.0001$	***	$d = 0.8651$	large
D.14	IF - Pt (♀)	37 : 2 (5.1%)	2 (5.1%)	– –	$p < 0.7728$	–	$d = 0.2997$	small
D.15	IG - Pt (♂)	33 : 4 (10.8%)	3 (8.1%)	1 (2.7%)	$p < 0.0022$	**	$d = 1.8327$	large
D.15	IF - En (♀)	35 : 1 (2.8%)	– –	1 (2.8%)	–	–	–	–

Table A.10: *Significance test and effect size between On-Talk speech rates and Off-Talk speech rates (w/o 1 word utterances)*

Dialog	Role - Lang	*On-Talk*	*Off-Talk Self* (%)	*Off-Talk Other* (%)	ANOVA	sig.
D.01	IG - Pt (♀)	47	3 (6.0%)	– –	$p < 0.1970$	–
D.01	IF - En (♂)	32	2 (5.9%)	– –	$p < 0.3570$	–
D.02	IG - En (♂)	79	9 (10.0%)	2 (2.2%)	$p < 0.0430$	*
D.02	IF - Pt (♂)	108	14 (10.7%)	9 (6.9%)	$p < 0.3280$	–
D.03	IG - En (♀)	86	4 (3.0%)	44 (32.8%)	$p < 0.0000$	***
D.03	IF - Pt (♀)	85	21 (14.2%)	42 (28.4%)	$p < 0.0109$	*
D.04	IG - Pt (♀)	48	8 (13.1%)	5 (8.2%)	$p < 0.0000$	***
D.04	IF - En (♂)	50	19 (26.0%)	4 (5.5%)	$p < 0.0356$	*
D.05	IG - En (♂)	163	– –	– –	–	–
D.05	IF - Pt (♀)	58	– –	– –	–	–
D.06	IG - En (♀)	89	59 (37.8%)	8 (5.1%)	$p < 0.0000$	***
D.06	IF - Pt (♀)	64	26 (28.9%)	– –	$p < 0.0325$	*
D.07	IG - Pt (♀)	93	– –	3 (3.1%)	$p < 0.1230$	–
D.07	IF - En (♂)	80	6 (6.9%)	1 (1.1%)	$p < 0.0009$	***
D.08	IG - En (♀)	104	1 (0.8%)	22 (17.3%)	$p < 0.0000$	***
D.08	IF - Pt (♀)	51	14 (13.6%)	38 (36.9%)	$p < 0.0318$	*
D.09	IG - Pt (♂)	94	1 (1.1%)	– –	$p < 0.7710$	–
D.09	IF - En (♂)	95	12 (12.0%)	– –	$p < 0.8590$	–
D.10	IG - Pt (♂)	56	4 (4.8%)	23 (27.7%)	$p <0.0000$	***
D.10	IF - En (♂)	69	29 (24.4%)	21 (17.6%)	$p < 0.0006$	***
D.11	IG - Pt (♀)	87	– –	– –	–	–
D.11	IF - En (♀)	48	42 (42.4%)	9 (9.1%)	$p < 0.0000$	***
D.12	IG - En (♂)	86	12 (12.2%)	– –	$p < 0.0000$	***
D.12	IF - Pt (♀)	68	2 (2.8%)	1 (1.4%)	$p < 0.7350$	–
D.13	IG - Pt (♀)	43	70 (61.9%)	– –	$p < 0.0000$	***
D.13	IF - En (♂)	46	2 (4.2%)	– –	$p < 0.1590$	–
D.14	IG - En (♂)	59	59 (33.0%)	– –	$p < 0.0003$	***
D.14	IF - Pt (♀)	37	2 (5.1%)	– –	$p < 0.6820$	–
D.15	IG - Pt (♂)	33	3 (8.1%)	1 (2.7%)	$p < 0.0065$	**
D.15	IF - En (♀)	35	– –	1 (2.8%)	$p < 0.6590$	–

Table A.11: *ANOVA significance test between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)*

Dial. - Role	Post-hoc Tukey HSD test ($p <$)			Cohen's d		
	On - Self	On - Other	Self - Other	On - Self, ES	On - Oth., ES	Self - Oth., ES
D.01 - IG	0.1973	–	–	$d = 0.7786$, med	–	–
D.01 - IF	0.3571	–	–	$d = 0.6810$, med	–	–
D.02 - IG	0.9289	0.0370 *	0.0415 *	$d = 0.1283$, neg	$d = 1.9285$, large	$d = 1.3227$, large
D.02 - IF	0.3195	0.8608	0.8523	$d = 0.4300$, small	$d = 0.1854$, neg	$d = 0.1745$, neg
D.03 - IG	0.0113 *	0.0000 ***	0.7976	$d = 1.3210$, large	$d = 1.8159$, large	$d = 0.5202$, med
D.03 - IF	0.9525	0.0085 **	0.1553	$d = 0.0651$, neg	$d = 0.6225$, med	$d = 0.5135$, med
D.04 - IG	0.0920 .	0.0000 ***	0.0346 *	$d = 0.7883$, med	$d = 2.2029$, large	$d = 2.4135$, large
D.04 - IF	0.2679	0.0575 .	0.3266	$d = 0.4371$, small	$d = 1.2586$, large	$d = 0.6697$, med
D.05 - IG	–	–	–	–	–	–
D.05 - IF	–	–	–	–	–	–
D.06 - IG	0.0000 ***	0.0138 *	0.9362	$d = 0.9069$, large	$d = 1.2415$, large	$d = 0.1124$, neg
D.06 - IF	0.0325 *	–	–	$d = 0.5054$, med	–	–
D.07 - IG	–	0.1228	–	–	$d = 0.9134$, large	–
D.07 - IF	0.0021 **	0.1612	0.9368	$d = 1.4851$, large	–	–
D.08 - IG	0.6534	0.0000 ***	0.7035	–	$d = 1.7034$, large	–
D.08 - IF	0.3362	0.0296 *	0.9129	$d = 0.4187$, small	$d = 0.5420$, med	$d = 0.1368$, neg
D.09 - IG	0.7709	–	–	–	–	–
D.09 - IF	0.8585	–	–	$d = 0.0528$, neg	–	–
D.10 - IG	0.0001 ***	0.0010 ***	0.0000 ***	$d = 2.0147$, large	$d = 1.3099$, large	$d = 2.4285$, large
D.10 - IF	0.2680	0.0004 ***	0.0740 .	$d = 0.3211$, small	$d = 1.1240$, large	$d = 0.5924$, med
D.11 - IG	–	–	–	–	–	–
D.11 - IF	0.0000 ***	0.0006 ***	0.8000	$d = 1.1427$, large	$d = 1.2100$, large	$d = 0.3152$, small
D.12 - IG	0.0000 ***	–	–	$d = 1.4567$, large	–	–
D.12 - IF	0.7264	–	–	$d = 0.5479$, med	–	–
D.13 - IG	0.0000 ***	–	–	$d = 1.1616$, large	–	–
D.13 - IF	0.1595	–	–	$d = 1.0330$, large	–	–
D.14 - IG	0.0003 ***	–	–	$d = 0.8651$, large	–	–
D.14 - IF	0.6821	–	–	$d = 0.2997$, small	–	–
D.15 - IG	0.0182 *	0.1309	0.9686	$d = 1.7384$, large	–	–
D.15 - IF	–	0.6587	–	–	–	–

Table A.12: *Post-hoc comparisons and effect size between On-Talk speech rates, Off-Talk Self speech rates, and Off-Talk Other speech rates (w/o 1 word utterance)*

A.5 ILMT-s2s User Survey Details and Replies

In § A.5.1: A table of the ILMT-s2s user survey fifteen statements and five questions is listed.

In § A.5.2: A figure of the ILMT-s2s user survey sheet included to present the layout of the 7 point Likert scare score and open text field for the statements and open text field of the questions.

In § A.5.3: A table of the ILMT-s2s user survey results of each subject is listed.

In § A.5.4: The subjects' open text field replies to the fifteen statements are listed in this section. When neither subject of the dialogue entered a reply the dialogue pair is not indicated, but if one of the subjects replied in the text field an empty comment (–) has been added to provide the details of the interlocutor. Spelling has not been corrected and remains as is.

In § A.5.5: The subjects' open text field replies to the five questions are listed in this section. When neither subject of the dialogue entered a reply the dialogue pair is not indicated, but if one of the subjects replied in the text field an empty comment (–) has been added to provide the details of the interlocutor. Spelling has not been corrected.

A.5.1 User Survey Statements and Questions

The list of the statements and questions asked to the subjects of the ILMT-s2s data collection are listed below in Table A.13.

	7 point Likert scale statements
1.	Overall, I am satisfied with how easy it was to use this system.
2.	It was simple to use this system.
3.	I could effectively complete the tasks using this system.
4.	I was able to complete the task quickly using this system.
5.	I was able to efficiently complete the task using this system.
6.	I felt comfortable using this system.
7.	It was easy to learn to use this system.
8.	I believe I could become productive quickly using this system.
9.	Whenever I made a mistake using the system, I could recover easily and quickly.
10.	The interface of the system was pleasant.
11.	I liked using the interface of this system.
12.	This system has all the functions and capabilities I expected it to have.
13.	I was satisfied with the voice of this system.
14.	I was satisfied with the output of this system.
15.	Overall, I am satisfied with this system.
	Open text field questions
1.	Please indicate why you changed the style of communication.
2.	Please indicate what made you give up clarifying the intention of the other participant.
3.	Please indicate all the things that irritated you.
4.	Please indicate all the things that pleased you.
5.	Please indicate what you felt was most difficult.

Table A.13: *List of survey statements and questions about the ILMT-s2s System*

A.5.2 ILMT-s2s User Survey Layout

TRINITY COLLEGE DUBLIN

QUESTIONNAIRE
for the
MULTILINGUAL MAP TASK CORPUS COLLECTION

RESEARCH PROJECT ID NUMBER: SCSS-000133/14
DOCUMENT ID NUMBER: *ILM-015-EEF*

Thank you for sparing the time to consider participating in this research.

1. ***Please be sure to read the below before answering this questionnaire.***

Each question hereafter is optional. Feel free to omit a response to any question; however we would be grateful if all questions are responded to.
Note:
- In the extremely unlikely event that illicit activity is reported we will be obliged to report it to appropriate authorities.
- Please do not name third parties in any open text field of this questionnaire. Any such reply will be anonymised.

2. ***About the system.***

1. Overall, I am satisfied with how easy it was to use this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

2. It was simple to use this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

3. I could effectively complete the tasks using this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

Ethics Application Guidelines – September 2011

Figure A.18: *Image of first page of the User Survey — 1 of 5 pages*

4. I was able to complete the task quickly using this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

5. I was able to efficiently complete the task using this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

6. I felt comfortable using this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

7. It was easy to learn to use this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

8. I believe I could become productive quickly using this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

Figure A.19: *Image of second page of the User Survey — 2 of 5 pages*

9. Whenever I made a mistake using the system, I could recover easily and quickly.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

10. The interface of the system was pleasant.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

11. I liked using the interface of this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

12. This system has all the functions and capabilities I expected it to have.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

13. I was satisfied with the voice of this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

Figure A.20: *Image of third page of the User Survey — 3 of 5 pages*

14. I was satisfied with the output of this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

15. Overall, I am satisfied with this system.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Comments:

3. ***About the actual dialogue.***

1. Please indicate why you changed the style of communication.

2. Please indicate what made you give up clarifying the intention of the other participant.

Figure A.21: *Image of fourth page of the User Survey — 4 of 5 pages*

3. Please indicate all the things that irritated you.

4. Please indicate all the things that pleased you.

5. Please indicate what you felt was most difficult.

Figure A.22: *Image of fifth page of the User Survey — 5 of 5 pages*

A.5.3 ILMT-s2s User Survey Replies

7 point Likert scale statements, evenly spaced from 1 = Strongly disagree to 7 = Strongly agree.

Dial.	S.01	S.02	S.03	S.04	S.05	S.06	S.07	S.08	S.09	S.10	S.11	S.12	S.13	S.14	S.15
01-IG	6	6	6	6	6	6	7	4	4	7	7	5	7	7	7
01-IF	3	4	4	3	2	3	6	2	4	6	4	4	6	2	3
02-IG	4	4	5	3	3	5	5	5	5	5	5	5	6	4	5
02-IF	7	7	7	3	5	5	6	3	2	6	7	3	7	3	3
03-IG	1	5	1	1	2	2	5	2	3	6	6	4	5	4	3
03-IF	1	6	1	1	1	3	6	3	2	4	5	3	3	4	2
04-IG	6	6	6	6	6	6	7	6	3	6	6	2	5	5	5
04-IF	4	6	4	5	4	6	5	4	5	6	6	4	6	4	5
05-IG	7	7	7	7	7	7	7	7	6	6	7	7	7	6	6
05-IF	3	7	5	4	6	4	7	4	5	7	7	5	7	5	4
06-IG	2	5	4	1	4	2	5	4	1	6	6	5	6	4	2
06-IF	2	3	2	1	1	2	7	2	3	5	6	5	3	5	3
07-IG	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
07-IF	2	7	2	1	1	7	7	1	6	5	6	5	6	5	2
08-IG	6	7	7	3	3	7	7	3	3	7	7	7	7	2	4
08-IF	2	2	4	4	4	3	3	2	3	3	3	4	6	4	2
09-IG	3	6	4	4	5	6	7	2	4	4	4	4	6	4	5
09-IF	2	7	1	4	1	1	1	1	1	7	7	7	7	1	2
10-IG	4	6	4	3	4	6	6	3	6	6	6	6	7	5	4
10-IF	3	4	5	4	5	5	6	3	4	5	4	4	6	3	4
11-IG	5	6	4	2	5	5	7	4	6	6	6	3	7	5	4
11-IF	2	5	3	1	1	2	5	1	5	6	6	1	6	2	2
12-IG	4	4	3	4	4	5	4	4	5	6	5	5	6	5	5
12-IF	4	3	3	4	3	5	6	2	3	3	3	2	6	4	4
13-IG	7	7	7	7	7	7	7	7	7	7	7	6	7	7	7
13-IF	6	6	4	6	6	6	6	6	4	7	7	7	7	7	7
14-IG	4	5	3	4	4	5	5	4	5	6	6	6	6	4	4
14-IF	7	7	6	6	6	7	7	7	7	7	7	7	7	7	7
15-IG	5	6	7	5	4	5	7	5	6	5	7	6	6	5	6
15-IF	5	5	4	7	6	5	5	5	6	6	5	4	2	5	5

Table A.14: *ILMT-s2s user Likert survey results*

A.5.4 The Fifteen 7 Point Likert Scale Statements' Open Text Field Comments

1. Overall, I am satisfied with how easy it was to use this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The translation was frequently incorrect. However, we managed to find ways to adapt.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): The interface of program is really easy to use.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): the system requires the user to take mental initiative and help the programme by selecting phrases it can process but is not unfriendly to use

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): I'm not sure if this was a flaw of the system, but it crashed several times and was often slow and too sensitive (i.e. if you clicked too fast it would send an error).

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Simple to click a button, but difficult to communicate using it.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): the system itself was not bad. It is easier though when I allow to see the english translation of what I said

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): There was some words that I could not understand, and the translation sometimes did not worked as it was supposed to work.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): very easy, interesting to see set up

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): The sistem is very simple but It was mostly hard, because the sentences I received wasnt clear and sometimes nonsense, with words which wasnt in the context.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): The system did not recognise over half the things I said

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): –

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Connection was slow and communication was difficult. There was a lot of interference when mike and speakers were both working

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): Apart from the technical difficulties, it was easy to use.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): The system didn't recognize a lot of the words and seemed to have preference for slang.

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The ASR was terrible. The interface was easy to use, but the ASR was just too bad.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): if I spoke slowly and keept it simple

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): Translation is a barrier to ease of use

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): –

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): It didn't recognize my English very well and the incorrect guesses it made were not even linguistically logical

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): The ASR was much better than the previous time I used his system.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Overall the system understood my voice but the responses were odd sometimes and I wondered if the translations were correct.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Some worlds the system translated in a wrong way, but was possible to understand the doubts from my collegue. I'm not sure if I can explain myself but I had to understand the sistem's problems and to change my words in order to be clear.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

2. It was simple to use this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The use of the system itself was manageable in terms of the interface.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): The possibility of see the people by video conference and listen the frases in our own lenguage, it is helpfull for the comunication.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): refer comments above

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): –

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Simple to click a button and speak

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): same as above. One problem was that sometimes while I was speaking, the microphone would pick the speech of what the other participant said as my speech and send it

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): It is really easy to work with it, and to understand what we need to do to make the system work.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): I think the sistem was easy because its only press a button and speak . But couldnt received the instructions correctly so I was trying to guess what she was saying.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): All you do is speak into a microphone

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Very easy to use.

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Only one button to use

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The interface was very easy to use and understand.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): Resonable interface

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): –

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): With the exception that it broke down a few times, it was relatively intuitive

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): The system is easy, but translation sometimes was wrong

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): Difficult at the beginning, but as time went on and I got used to it, it became easier

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): There is nothing difficult about it.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Yes, it was simple to use the system.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

3. I could effectively complete the tasks using this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): Although I feel we completed the task eventually, my confidence in the route would not have been extremely high.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): yes, I could complete the task.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): with a few mistakes and misinterpretations

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): Again since there were so many crashes it took more than an hour and I still wasn't halfway through the path. Furthermore, the translations were so difficult to understand at times that it was really hard getting the messages across.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Slow to convert to text, and very faulty. Very difficult to get correct words recognised – even when enunciating clearly.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I did not give whole instructions such as go right or left but we got the landmarks right

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I could complete the tasks, but it was not easy, due to the translation that sometimes was wrong.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): It would take a long time to complette tasks.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): Maybe, if it worked better???

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): –

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Tasks were completed with great difficulty

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The interface, yes, but the ASR was not recognizing my speech.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): I think we did well enough.

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): –

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): I Finished, but it really shouldn't have taken that long

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): I beleave that tecnology is a good idea, but maybe will be necessary to better the system

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): not certain if task was completed correctly

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): Yes I was able to communicate what I wanted. But I could not be sure. I was guessing a lot and hoping that the other would also work out what I ntended to say.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): We eventually got to the ship, although I am not sure if we got there on the same path.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Yes, absolutely. I think that wasn't the same map but was easy to complete the tasks.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): Depends on the task. No prosody made it dificult to identify questions.

4. I was able to complete the task quickly using this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Sometimes the system didn't understand a few words.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The task certainly took longer than it would if between two speakers of the same language

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): the system had some difficults to translate some sentence and I needed to say many times the same sentence.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): the system requires patience to and a relatively slow pace for coherent communication to prevail

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): It was very hard to get the message across. There were several things that were lost in translation. First, the microphone sometimes didn't pick up what I actually said or I needed to speak more clearly, because it would make up words or add things that weren't there. The program also didn't make up for accents all that well. I also noticed that it was really difficult to ask questions because all of the phrasing was like normal sentences, so it would be nice if there was some way for it to pick up on intonations (i.e. 'to the right.' is not the same as 'to the right?'). Similarly, sometimes I would ask a question in Portuguese, but it would translate it as a normal sentence with a different message than what was intended. All of these caused set-backs and made the process a very long one.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Took a while to communicate due to recognition errors

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): same as above

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I took more time than I should, because the translation sometimes was not correct

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): This task took what seems like forever

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): The system had problems with the voice recognition and translation thus increasing the time it took to complete.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): It didn't take too long.

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): I needed to talk slowly and translation was completely different

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): Yes, I didn't feel it morw difficult to H2H communication.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): After a start where I thought we would never get there, we eventually reached the ship.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Yes, I think so.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

5. I was able to efficiently complete the task using this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): We did complete the task eventually but I could not state that it was efficient. I feel if this was a real-life scenario, where my actual route depended upon this system, my feelings would lean stronger towards the negative.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): It was efficiently to complete the task, but it need some improvements.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): as it required my me to phrase sentences slowly it is slightly inefficient

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): Refer to question 4

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): same as 3

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I was able to complete the task, but I'm not sure how efficient I was.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): yes; was interesting to see some translations, and figuring out how to explain task

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): I completed it, but it was by no way efficient. A lot of guesswork was needed

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): No, it was not efficiant. It was difficult to get the right ASR.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): once we got used with the limitations

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): –

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): I'm not sure if was my voice or words, but I had some troubles with some translations. I change my mind/words sometimes.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

6. I felt comfortable using this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): Being able to see the other speakers face aided my comfort with the system as I was better able to discern her emotional state. However, the translation at times produced awkward responses.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): Yes, I felt.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): –

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): It was easy enough to use, but there were still several flaws.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): it is very entertaining

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): The system is ok, but not effective.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): It frustrated me a lot

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The interface, yes, but no the ASR.

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Probabily if I had to talk with a friend we will have more information for this conversation.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): Though I did realise that I was relying on the other's guess work.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Yes, felt like using Siri or Google Now.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Not so confortabel maybe because it was my first time using that system but I'm sure that I will be more confortable next time.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

7. It was easy to learn to use this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): Learning to use the system itself was easy from a technical standpoint. I did note how we needed to modify our manner of speaking throughout the conversation. We learned to adapt.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): Yes, it was really easy how I sad before the interface of program is easy to understand.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): once explained i had no major problems in using the system

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): –

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Easy to click the button - very straightforward to learn and explain to others.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): the interface is quite user friendly. I just wish I could also see the original of what the other participant said. But that's cheating I guess.

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): It is easy to work with the system, and easy to learn how to use it as well.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): It was a simple sraightforward process

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): Yet again, the interface, yes, but not the ASR.

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Yes, the system looks easy. In my opinion the problem was translation.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): intuitive

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Yes, it was. Quite natural.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

8. I believe I could become productive quickly using this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): This system is very useful, but a bit slow to process, I also didn't like that you have to push the button to talk, it would be easier just talking and that is it, then you can talk while doing something else.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): Repeated use of this system would certainly improve my efficiency with it as I would know to modify my own behaviour with it from the start. However, even after a long period I would still feel strongly that the system itself had many flaws.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): Due the priblems with some sentences that I sad and some sentences that I received with some words in Englsih, I believe tha some modifications are needs to improve and turn the system more productive.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): with a lot of practice and adapting to the system

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): If the flaws were to be fixed then I would be much more comfortable using the system.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Too many faults, and slow to communicate. Gestures were often resorted to since they were quicker and easier.

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Not so quickly because not everything is translated as it should be.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): It could take a time, but I think we did something in the map in the end.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): If it understood me more, i could maybe write things with it... maybe?

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): No, with this ASR, it cannot replace humans.

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Yes, I believe.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): uncertain, I imagine tasks where both sides had the same information in front of them would be easier though, rather than the contrasting maps we received.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): I think I could get to grips better with this system with practice, the voice recognition seemed relatively good but not sure how it was with translating my English.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): It depends for wich reason I have to use, but I would say yes.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

9. Whenever I made a mistake using the system, I could recover easily and quickly.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): More or less, it helped that I was seeing the other person face, so I knew by his facial expressions when some words weren't matching. But if I had to wait for them to say something was wrong would take a bit of time.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The source of mistakes in using the system originated primarily from the translation itself. We did recover but this took time. The delay in the translation became noticable.

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): Not necessarily. There were some words that no matter how many different ways I said them the microphone would pick them up as other things. And the same would often happen to the other person, too. Then if she said something I didn't understand it was really difficult to ask questions.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Often made repeat mistakes, but not very consistent.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): not that quickly as the program showed some errors. I think I tried to speak faster than it was translating

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I had to try again and talk in a slower pace to make the system understand sometimes.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): I had to repeat over and over and over...

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): When I made a mistake, it took a number of attempts to come up with a different way to say something so that the system would recognise what I was saying.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): No, it was very difficult to correct wrong ASR.

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): simply said not "the word the computer understood", but "the word I meant" eg. not mum but map

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): As the system translated my speech I was able to see how words with similar sounds were being mistranslated, but by quickly thinking of synonyms the problem was easily solved

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): I didn't make a mistake.

10. The interface of the system was pleasant.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Simple and effective.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The interface of the system itself did not negatively affect the experience. Perhaps a clearer indication of the other speaker's state would help. Additional effort was required to remember to check this before speaking

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): –

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Simple and straightforward.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): simple enough

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Clean and simple interface.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): The interface was simple and nice

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): Except for the EEG

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Yes, was normal. My doubt is about the exclamation marks, question marks, and so on...sometimes I couldn't understand if my collegue had doubts or not.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

11. I liked using the interface of this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): While I had no particular issue with the interface in use, issues with the translation skew my overall perception of the usability of the system.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): it was a very interesting medium of communication

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): It was simple and straightforward.

12. This system has all the functions and capabilities I expected it to have.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): As stated earlier, a clearer indication of the other speakers state might help. An ability to make a drawing visible to the other participant would also clearly help with this task. Afterwards I realized that we should have tried using hand gestures on camera to indicate the route. Perhaps it was the presence of the finger sensors which prevented me from considering this.

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): It would be nice if it could catch on intonation. And if you could see what it was the other person was saying written in the language that they said it in.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I wish I could see the original of the other participants speech

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): –

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): i fel the translations to be inaccurate

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): Text chat

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): I expected to be able to speak at my standard pace, but I couldn't. I had to speak in a slower way so the system could understand some words. Also, the system still doesn't recognise some words.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): its speech recognition is lacking

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): The translation was complicate

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): I didn't really come with any expectations, and was overall amazed by the simplicity of the system as well as by its speed. The translations were almost simultaneous.

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): –

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): I was wondering if is possible to add a "key word" in order to corret when the system translate in a wrong way my voice.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

13. I was satisfied with the voice of this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): I have no issues with the voice of the system. It was very clear.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): the voice clearly speaks the informatin it has processed

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): The pronunciations were a bit off sometimes

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Good voice, nice pace, and good accent.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Yes, the voice is agradable

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Good clear voice.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): Was clear and easy to understand.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

14. I was satisfied with the output of this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The output of the translation was often clearly flawed.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): mostly, yes. There were some bad translations but in the end we managed

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): –

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): The translation output seemed to be off.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): No, not the ASR output.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Voice recognition was good but not sure of the translation side.

15. Overall, I am satisfied with this system.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): Overall, we accomplished the task and therefore the system must be given some credit. However, if in a real-life scenario, I know I would not be happy with the performance.

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Have a lot of things to improve, but it is a good system.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): It was extremely annoying to use

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): –

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): Not with the quality of the ASR.

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): When the settings are for Portuguese (Portugal) it would be nice that no Portuguese words (Brazil) would appear.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): –

A.5.5 The Five Open Text Field Questions' Comments

1. Please indicate why you changed the style of communication.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): I was trying to talk normally, but I notice that if I paused more the system understood better.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): We did change the style of our communication. This was mainly to counteract the poor translation. We learned to speak more clearly, to repeat key phrases multiple times during a single translation and near the end, the other speaker read over the entire version of her map. If trying the process again I might attempt this from the start.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): how I sad before, the possibility of talk and see the other people and have our frase transleted to our own lenguage it will be really helpfull for global comunication.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): to suit the needs of both the system and and the other participant's native language

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): –

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Had to enunciate very carefully to ensure the landmarks were identified.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I'm not sure I get the statement. But I did start using shorter sentences as I saw the system would translate them better

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Spoke in shorter sentences to avoid bad speach recognition

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I tried to speak in a lower pace with a good accent, to make it easier for the machine to understand what I was saying.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): Sometimes I needed to enunciate and repeat words

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): I tried speak slowly because I though it would be better for the another person get what I was saying.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): I had to pronouce things in a different manner in order for the system to understand me

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Some words that I used, weren't recognized correctly, maybe because they were more colloquial rather than formal. So I chaged the choice of words to one I judge would be better recognized by the software.

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Shorter sentances and using key words (left, right, up down, around, north south etc) gave less room for translation or interpretation error by the system

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): try to find a correct information

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): So that the person was able to understand what I was saying. So it would be easier for the system to translate what I was saying. Had to think about how the system was going to translate something.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): The system didn't understand me sometimes.

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The system just would not understand what I was saying. I changed the words to be simpler, but it still didn't work.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): to make possible for the sistem to correctly translate I had to use simpler words and phrases

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): short statments for better translation

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): I had to change the style to fit the system: talk in a slower pace, try to articulate the words really well, and sometimes try to use the terms that the system used to translate a word (for example, I was using the word 'pinhal', as in the map, and then I realised the translation when the other participant spoke was 'pinheiral', so I started to use 'pinheiral', so the right word would appear for the other participant.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): I had to make my sentences shorter because the program couldn't recognize them

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Probabily I will keep to talk with other systems. This is a option, is a different and maybe a long the time will be good. But nowaday, for me, I prefer to use others systems because it is more easy than this one.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): In order to facilitate the passsing of instruction, we needed to slow down and take each step, as there were certain things which were present on her map and not on mine, as well as vice versa.

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): I believe that I overall spoke more slowly and loudly than I usually would. I slowed down to make sure my pronunciation was clear as most of the times there was a mistranslation, it was due to my pronunciation. I was able to recognize how or why the system understood one word rather than the other, and so I tried to adapt to the system. As for loudness, I am not sure why I spoke loudly. . . I guess we often think that turning the volume up will make things clearer- not a good assumption though.

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): Since the ASR wa better than before, I didn't need to change the communication style.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): For the voice recognition system to recognoise my words correctly.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): I needed to understant when the system translate in a wrong way my voice or order. I had to change my mind (change my way of speaking) in order to use the system and make me clear to my partner.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): Enunciating my words helped the ASR. Since queries weren't recognised, some alternative means of denoting questions was needed

2. Please indicate what made you give up clarifying the intention of the other participant.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): –

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): I would not say I gave up clarifying the intention of the other participant. We worked to understand each other throughout.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): The task was easy, and the other participant needed to repeat the sentence many times, but it works.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): –

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): I do not understand the question. I felt like the entire conversation was us trying to clarify each others' intentions.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): –

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): I did not know actually that I had to use commands like left, right, over the landmarks. I believe the goal was to get all the landmarks until the last one, in that order. So, in the end, we just listed the landmarks he had to tick.

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): somtimes easier to start again

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Sometimes the machine was understanding random things, instead of what I was trying to say.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): when the output was clear enough to guess the intentions

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): It got so frustrating at some points that i just decided to assume she knew what i was saying. . .

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Nothing in particuplar made me give up, we just had to rephrase it to better understand each other.

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): When the information given did not correspond to my map. When no landmark beacons were given or the translation made no sense.

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): her expression when i was trying to say something using many possibles words

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): Im not sure what this question means.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): Some of the times it looked like he didn't have the same map. Other times, the system simply didn't understand what I was saying.

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The ASR just wouldn't recognize my speech correctly.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): –

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): I didn't give up, i think.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): this program didn't insert question marks or other punctuaion so any question looked like I was just foolishly repeating myself

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): had been better if the other participant could be more interage. I think he was not to much agradable to speak. For this reason the conversation is boring.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): If the meaning was relatively clear from what I received, then there was essentially no need to check every word, as it was quicker just to get the gist of what was meant.

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): I don't think I gave up clarifying the participant's intention. In fact, I think the conversation flowed quite nicely as the other participant would clarify himself even before I had the time to ask him to repeat, which was nice. As we are able to read what the system has heard us speak, we are able to anticipate the mistakes and problems the other participant might encounter, hence being able to immediately correct our speech.

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): Since the ASR was better than before, I didn't need to.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): It was frustrating when our answers and questions were not coherent sometimes but we stuck with it.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

3. Please indicate all the things that irritated you.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Some of the misunderstanding of words were a bit annoying, I thought it was funny here, but if I was using the system for a professional reason would have been a problem.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The main irritation was the poor translation. The presence of the sensors did without doubt hinder my movement and made me more aware of the experimental scenario. Also a more intuitive indication of the other speaker's state would be useful. The camera did not often help in this regard, as the other participant would often look down at the map while speaking. I needed to look at multiple places on screen and this was demanding cognitively over time.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): Not irritated but the problems with some sentences that I tried to send and some sentences that I received with English words, make me to think more to understand the dialogue.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): some translations bear no resemblence to what the speaker has input

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): Look at question 4 of part 1.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Wearing the camera; having to repeat myself frequently; realising that only part of the message had been understood, so had to clarify things frequently. The delay between speaking and the translation made it difficult to know when to speak, since you were frequently risking talking when the translation comes through, or waiting for the technology for ages.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): some bad translations, such as baboons. But Also I noticed the same landmarks had their names in european portuguese, like camiao (caminhao).

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Having different maps

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Some words didn't came as I expected.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): didnt accept single words; had to repeat, but fine if i spoke more clearly

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): When person didnt speak anything.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): It didnt understand what I was saying and by the looks of things didnt understand the other participant either

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): Nothing about the system itself irritated me, it is very easy to use. Sometimes the words weren't translated correctly, so I had to ask my colleague to rephrase it again, or I changed the words that I used in order to be translated effectively to him.

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): The instructions were not clear and concise. The mic would often record the output from the speakers the system was quite slow

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): the system doesn't understand what i wanted to say

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): Having to repeat things over and over again, having trouble communicating, the other person not understanding me.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): The fact that some words consistently would not be translated at all. Also, the impossibility of having longer sentences. Sometimes, when the other participant spoke, it would be heard by the system and translate again.

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The fact that the ASR was constantly wrong. I have the feeling it was wrong 99% of the time. It's nearly a mirical that it was finished.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): getting succesive wrong translations for the same word

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): Lag time between input and translation

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): I had to change the style to fit the system: talk in a slower pace, try to articulate the words really well.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): the absurd translations, the lag, the crashing, the inability to effectively communicate, the lack of a significant flag that the other person is talking (sure there is a light up mic, but I think it would be more effective to put a notification in the text box like "so and so is talking"

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): The other participate because he talked a bit and he was not friendly. And second factor was the translation that in big part was completaly different. So, the comunication was totally wrong.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): When the system misunderstood what you were trying to say, it was rather annoying, as things had to be repeated in small blocks spoken clearly. which is quite different to ordinary conversation. It also quite irritated me that there were certain features present on only one of our maps, which made it difficult to explain the route.

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): I was not irritated at any points. I did laugh a few times though because I find that translating is something reasonably difficult and it was funny to see how pronunciation can impact so much on understanding.

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): This time, there wasn't much that irritated me.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Answers to my questions which were non sensical in the context.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): When the system translate the word "Pinhal" (kind of forest) to "Piaui" (Brazilian State). I know the difference, for me it was ok but if my collegue she doesn't know and even If she has two option on her map "Pinhal" and "Piaui" she can go to another place really far!

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): Not everything got translated, so there was some confusion about the landmarks

4. Please indicate all the things that pleased you.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): The interface and the simplicity of the app.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): Several awkward, funny translations were produced. This might actually have helped us identify better with each other, being able to see the other's expression. Just the fact that we did get a reasonable result, despite speaking different languages, I suppose was a positive.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): –

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): it was fairly easy to use and recoer from mistakes

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): It was really funny when the translations were goofy and wrong.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): It was good to see the status of the translation for the other participant, as it helped to determine when to speak.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): it was really funny and useful. Although the translation is not 100%, for a language that you have no knowledge at all, would be very helpful.

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): I think it the instructions could explain more clearltyt that we need to draw the path around places not to them

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): interesting system, have a potential and will be really useful

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): cool to see the progress, thinkin g about landmarks and interaction

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): read the sentences she was saying and what I said, because I could read it over again and try to identify.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): the interface was clear and precise

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): It is very fast and easy to use and the conversation was translated correctly most of the time.

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): Communicating through the system was interesting and fun.

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): i don't know it was really hard to find a right answer

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): Being able to see the other person made it easier to communicate. When the system recognised words entirely differently, it was very funny.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): To be able to communicate via this system.

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): The voice, the interface.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): It is in overall a good system, simple interface, fast enough but with room for improvments.

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): –

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): Fun experience overall. It's a very useful system if perfected, and it pleased me to try it.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): It was easy to use

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): The voice of translation and the idea about the new system

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): Was quite happy with the way the system understood what I said, if spoken clearly, and it seemed to translate what was said on the other side relatively well also, with the exception of some words like "barranco"

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): The system was simple to use and extremely quick. I also liked the fact that we could read what we said because that made it easier for me to understand the main difficulties and to adapt my pronunciation and choice of words to the system.

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): How well the ASR recognised my English.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Voice recognition and user friendly interface.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): The time, the sincronism, the layout really easy to use, the voice from the system.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

5. Please indicate what you felt was most difficult.

Dialogue 01 (IG, Pt, ♀, w/ Video, w/o Crash, w/o Bio): Hold the button to talk.

Dialogue 01 (IF, En, ♂, w/ Video, w/o Crash, w/ Bio): The main source of difficulty was the poor translation. The delay in the translation was also frustrating at times, particularly when I wanted to fix a mistake quickly. The need to look at multiple places on screen and to the table was irritating at times. Constantly checking to see the other user's state also became tiring. The camera often did not help in this regard as she would look down when speaking. An interface which grouped these elements more closely would help. I was also aware of the sensors, which restricted my

head and finger movements, and I think prevented me from trying means of communication I might have otherwise attempted.

Dialogue 02 (IF, Pt, ♂, w/ Video, w/ Crash, w/o Bio): To need to reapet the sentence many times.

Dialogue 02 (IG, En, ♂, w/ Video, w/ Crash, w/ Bio): inputing particularly long words was hard for the system to understand

Dialogue 03 (IF, Pt, ♀, w/ Video, w/ Crash, w/ Bio): Look at question 4 of part 1. In addition, homonyms were a huge issue. Sometimes the microphone would pick up your voice and the words would be interpreted as homonyms of what you actually intended to say. It's interesting because as a native speaker you can deduce what someone meant to say even if it comes out wrong, but the other person can't. There was no way for her to know that the word for temple ('templo') was very similar to the word for time ('tempo'), so whenever 'time' came up she looked immensely confused. It was funny but also frustrating.

Dialogue 03 (IG, En, ♀, w/ Video, w/ Crash, w/o Bio): Clarifying the directions - since"left", "right" and"down" were frequently mis-transcribed.

Dialogue 04 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): the most difficult thing was to wait for the other participant to talk back especially when I saw that the system had not got what I said correctly and I wanted to correct that. Also, when I was recording my voice and the system would read the other partcipant's output in portugues, so it would record that speech as mine

Dialogue 04 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): –

Dialogue 05 (IF, Pt, ♀, w/o Video, w/o Crash, w/o Bio): the system do not work with questions, commas and points.

Dialogue 05 (IG, En, ♂, w/o Video, w/o Crash, w/ Bio): understanding some of output and figuring it out

Dialogue 06 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): the most difficult thing was try to find the directions and things she said I didnt have on my map.

Dialogue 06 (IG, En, ♀, w/o Video, w/o Crash, w/o Bio): Trying to make the system understand I was saying white, instead of flight. . . etc

Dialogue 07 (IG, Pt, ♀, w/o Video, w/o Crash, w/o Bio): The maps weren't the same so it was confusing, because the other participant had different landmarks so I wasn't sure if he was going to the right place or not.

Dialogue 07 (IF, En, ♂, w/o Video, w/o Crash, w/ Bio): To try to reinterpret what had been said from the botched translation

Dialogue 08 (IF, Pt, ♀, w/ Video, w/ Crash, w/o Bio): to organize the words and try to find a meaning for it

Dialogue 08 (IG, En, ♀, w/ Video, w/ Crash, w/ Bio): Trying to find a way to say something that the system would understand.

Dialogue 09 (IG, Pt, ♂, w/ Video, w/o Crash, w/ Bio): To explain things in simple words.

Dialogue 09 (IF, En, ♂, w/ Video, w/o Crash, w/o Bio): Trying to adapt my speech so that the ASR could understand.

Dialogue 10 (IG, Pt, ♂, w/ Video, w/ Crash, w/o Bio): getting the right translation for same unusual words

Dialogue 10 (IF, En, ♂, w/ Video, w/ Crash, w/ Bio): Differences on the maps caused confusion

Dialogue 11 (IG, Pt, ♀, w/ Video, w/o Crash, w/ Bio): I had to change the style to fit the system: talk in a slower pace, try to articulate the words really well.

Dialogue 11 (IF, En, ♀, w/ Video, w/o Crash, w/o Bio): communicating

Dialogue 12 (IF, Pt, ♀, w/ Video, w/o Crash, w/o Bio): I was felling difficult in keep to talk with other participate.

Dialogue 12 (IG, En, ♂, w/ Video, w/o Crash, w/ Bio): Trying to understand where objects were placed on the opposite map in relation to the objects I could see before me.

Dialogue 13 (IG, Pt, ♀, w/o Video, w/o Crash, w/ Bio): Pronunciation/enunciation: (a) sometimes pronunciation got in the way as I would pronunce my 'e' as an 'i', making it a bit confusing for the system to translate; (b) some words have extremely similar sounds so the system wouldn't be able to hear the first or final letters, causing the words to be misunderstood. Also, sometimes as I was speaking the participant's translation would also be spoken, so the traslations mixed up both of our speeches. However, this was easy to fix once I realized the system shows exactly when the other person is speaking and listening.

Dialogue 13 (IF, En, ♂, w/o Video, w/o Crash, w/o Bio): This time, nothing much was difficult.

Dialogue 14 (IF, Pt, ♀, w/o Video, w/o Crash, w/ Bio): –

Dialogue 14 (IG, En, ♂, w/o Video, w/o Crash, w/o Bio): Not knowing what exactly the other person was looking at and making sense of some of the answers.

Dialogue 15 (IG, Pt, ♂, w/o Video, w/o Crash, w/ Bio): At the beggining to understand the system's problems and then for wich way I will be more clear. After that I think that was easy.

Dialogue 15 (IF, En, ♀, w/o Video, w/o Crash, w/o Bio): –

A.6 ILMT-s2s User Survey Results

In § A.6.1: Tables of the ILMT-s2s user survey statements median (Table A.15) and means (Table A.16) of all subjects as the "Total" but also the median and mean values of the various groupings — The role of the subject (IG or IF), the setting of the recording (with or without live video streaming), the languages spoken by the subject (English or Portuguese), the map used to elicit the dialogue (Map 01 or Map 07), the gender of the subject (male or female) and the dialogues with or without a crashes.

In § A.6.2: This section is a continuation of the data in § A.6.1 but represents the results in stacked bar charts and density plots. A list of the figures and descriptions of the figures can be found in § A.6.2.

In § A.6.3: This section is also a continuation of the data in § A.6.1 with a list of significance tests and effect size. A list of the tables and descriptions of the tables can be found in § A.6.3.

A.6.1 User Survey Statements' Median and Mean

7 point Likert scale statements, evenly spaced from 1 = Strongly disagree to 7 = Strongly agree

Statements	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	*Mdn*
Total	**4.0**	**6.0**	**4.0**	**4.0**	**4.0**	**5.0**	**6.0**	**4.0**	**4.5**	**6.0**	**6.0**	**5.0**	**6.0**	**4.5**	**4.0**	**5.0**
IG	5.0	6.0	5.0	4.0	4.0	6.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0	6.0
Diff.	2.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	2.0	2.0
IF	3.0	6.0	4.0	4.0	4.0	5.0	6.0	3.0	4.0	6.0	6.0	4.0	6.0	4.0	3.0	4.0
w/ Vid	3.5	5.5	4.0	3.0	3.5	5.0	6.0	3.0	4.0	6.0	5.5	4.0	6.0	4.0	4.0	4.0
Diff.	1.5	0.5	0.5	2.5	2.5	1.0	1.0	2.0	1.5	0.0	1.0	1.5	0.0	1.0	1.0	2.0
w/o Vid	5.0	6.0	4.5	5.5	6.0	6.0	7.0	5.0	5.5	6.0	6.5	5.5	6.0	5.0	5.0	6.0
V. w/o C. †	3.5	5.5	3.5	4.0	3.5	5.0	6.0	2.0	4.0	6.0	5.5	4.0	6.0	4.0	4.0	4.0
Eng	4.0	5.0	4.0	4.0	4.0	5.0	5.0	4.0	5.0	6.0	6.0	5.0	6.0	4.0	4.0	5.0
Diff.	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0
Por	5.0	6.0	5.0	4.0	5.0	5.0	7.0	4.0	4.0	6.0	6.0	5.0	7.0	5.0	4.0	6.0
w/ Bio	4.0	6.0	4.0	4.0	4.0	5.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0	5.0
Diff.	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
w/o Bio	4.0	6.0	4.0	4.0	4.0	5.0	6.0	4.0	4.0	6.0	6.0	5.0	6.0	4.0	4.0	5.0
Map 01	3.5	6.0	4.5	3.0	4.0	5.0	6.5	4.0	4.0	6.0	6.0	5.0	6.0	4.5	3.5	5.0
Diff.	0.5	0.0	0.5	1.0	0.0	0.0	0.5	0.5	1.0	0.0	0.0	0.5	0.0	0.0	0.5	0.0
Map 07	4.0	6.0	4.0	4.0	4.0	5.0	6.0	3.5	5.0	6.0	6.0	5.5	6.0	4.5	4.0	5.0
♂	4.0	6.0	4.0	4.0	4.0	5.0	6.0	3.5	5.0	6.0	6.0	5.0	6.0	4.0	4.5	5.0
Diff.	0.5	0.0	0.0	0.0	0.5	0.0	1.0	0.5	1.5	0.0	0.0	0.5	0.0	1.0	0.5	0.0
♀	4.5	6.0	4.0	4.0	4.5	5.0	7.0	4.0	3.5	6.0	6.0	4.5	6.0	5.0	4.0	5.0
w/ Crash	3.5	5.5	4.5	3.0	3.5	5.0	6.0	3.0	3.0	5.5	5.5	4.0	6.0	4.0	3.5	4.0
Diff.	0.5	0.5	0.5	1.0	1.0	0.0	1.0	1.0	2.0	0.5	0.5	1.0	0.0	1.0	1.5	1.0
w/o Crash	4.0	6.0	4.0	4.0	4.5	5.0	7.0	4.0	5.0	6.0	6.0	5.0	6.0	5.0	5.0	5.0

†: Dialogues with live video streaming that didn't crash.

Table A.15: *Median value of the thirty Subjects' Likert scale results*

7 point Likert scale statements, evenly spaced from 1 = Strongly disagree to 7 = Strongly agree

Statements	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	*M*
Total	**4.1**	**5.5**	**4.3**	**3.9**	**4.1**	**4.8**	**5.9**	**3.8**	**4.4**	**5.8**	**5.8**	**4.8**	**6.0**	**4.5**	**4.3**	**4.8**
IG	4.7	5.8	5.0	4.2	4.7	5.4	6.2	4.5	4.7	6.0	6.1	5.2	6.3	4.9	4.9	5.3
Diff.	1.2	0.5	1.3	0.6	1.3	1.1	0.7	1.4	0.7	0.5	0.6	0.9	0.7	0.9	1.3	0.9
IF	3.5	5.3	3.7	3.6	3.5	4.3	5.5	3.1	4.0	5.5	5.5	4.3	5.7	4.1	3.7	4.3
w/ Vid	3.6	5.1	3.9	3.1	3.4	4.3	5.4	2.8	3.8	5.4	5.3	4.2	6.1	3.7	3.7	4.3
Diff.	1.2	0.9	1.0	1.7	1.6	1.1	0.9	2.2	1.3	0.7	1.1	1.2	0.3	1.7	1.3	1.2
w/o Vid	4.8	6.0	4.9	4.8	4.9	5.4	6.4	4.9	5.1	6.1	6.4	5.4	5.9	5.4	5.0	5.4
V. w/o C. †	3.6	5.1	3.5	3.5	3.4	4.1	5.4	2.5	4.0	5.6	5.3	3.9	6.4	3.8	4.0	4.3
Eng	3.7	5.4	3.8	3.6	3.5	4.5	5.3	3.5	4.2	6.0	5.8	5.0	5.9	3.9	3.9	4.5
Diff.	0.9	0.3	1.1	0.6	1.1	0.6	1.2	0.6	0.3	0.5	0.1	0.5	0.1	1.3	0.7	0.5
Por	4.6	5.7	4.9	4.2	4.7	5.1	6.5	4.1	4.5	5.5	5.9	4.5	6.1	5.1	4.7	5.1
w/ Bio	4.2	5.6	4.5	3.7	3.9	5.3	6.3	3.9	4.9	5.6	5.7	5.0	5.9	4.5	4.5	4.9
Diff.	0.1	0.1	0.4	0.3	0.5	1.0	0.9	0.3	1.0	0.3	0.2	0.5	0.1	0.1	0.5	0.2
w/o Bio	4.1	5.5	4.1	4.1	4.3	4.3	5.4	3.6	3.9	5.9	5.9	4.5	6.1	4.5	4.1	4.7
Map 01	3.9	5.7	4.4	3.5	3.9	4.6	6.2	3.9	4.0	5.9	6.1	4.6	5.8	4.6	4.1	4.7
Diff.	0.4	0.3	0.0	0.8	0.3	0.4	0.7	0.2	0.7	0.2	0.4	0.4	0.4	0.3	0.4	0.1
Map 07	4.3	5.4	4.3	4.3	4.3	5.0	5.6	3.7	4.7	5.7	5.6	4.9	6.2	4.4	4.5	4.9
♂	4.1	5.6	4.3	4.0	3.9	5.1	5.6	5.6	4.5	5.7	5.8	5.2	6.4	4.1	4.4	4.8
Diff.	0.0	0.2	0.1	0.2	0.3	0.6	0.6	0.4	0.3	0.1	0.1	0.8	0.7	0.7	0.1	0.0
♀	4.1	5.4	4.4	3.8	4.3	4.6	6.1	3.9	4.3	5.8	5.9	4.4	5.7	4.8	4.3	4.8
w/ Crash	3.5	5.1	4.3	2.8	3.4	4.5	5.5	3.0	3.5	5.3	5.4	4.5	5.9	3.6	3.4	4.2
Diff.	0.9	0.6	0.1	1.6	1.0	0.5	0.5	1.0	1.2	0.7	0.6	0.4	0.2	1.2	1.3	0.8
w/o Crash	4.4	5.7	4.4	4.3	4.4	5.0	6.0	4.0	4.7	6.0	6.0	4.9	6.0	4.8	4.6	5.0

†: Dialogues with live video streaming that didn't crash.

Table A.16: *Mean value of the thirty Subjects' Likert scale results*

A.6.2 User Survey Stacked Bar Charts and Density Plots

The following pages contain the graphical representation of the Likert score results.

Figure A.23: Stacked bar chart and density plot of subject survey Likert scale results differentiated by "w/o Video" and "w/ Video" settings.

Figure A.24: Stacked bar chart and density plot of subject survey Likert scale results differentiated by the role of the subject, Information Giver or Information Follower.

Figure A.25: Stacked bar chart and density plot of subject survey Likert scale results differentiated by the language spoken by the subject, Portuguese or English.

Figure A.26: Stacked bar chart and density plot of subject survey Likert scale results differentiated by the gender of the subject, male or female.

Figure A.27: Stacked bar chart and density plot of subject survey Likert scale results differentiated by the map used by the subject, Map 01 or Map 07.

Figure A.28: Stacked bar chart of subject survey Likert scale results differentiated by the setting and role of the subject — "w/o Video – IG", "w/o Video – IF", "w/ Video – IG" and "w/ Video – IF".

Figure A.29: Density plot of subject survey Likert scale results differentiated by the setting and role of the subject — "w/o Video – IG", "w/o Video – IF", "w/ Video – IG" and "w/ Video – IF".

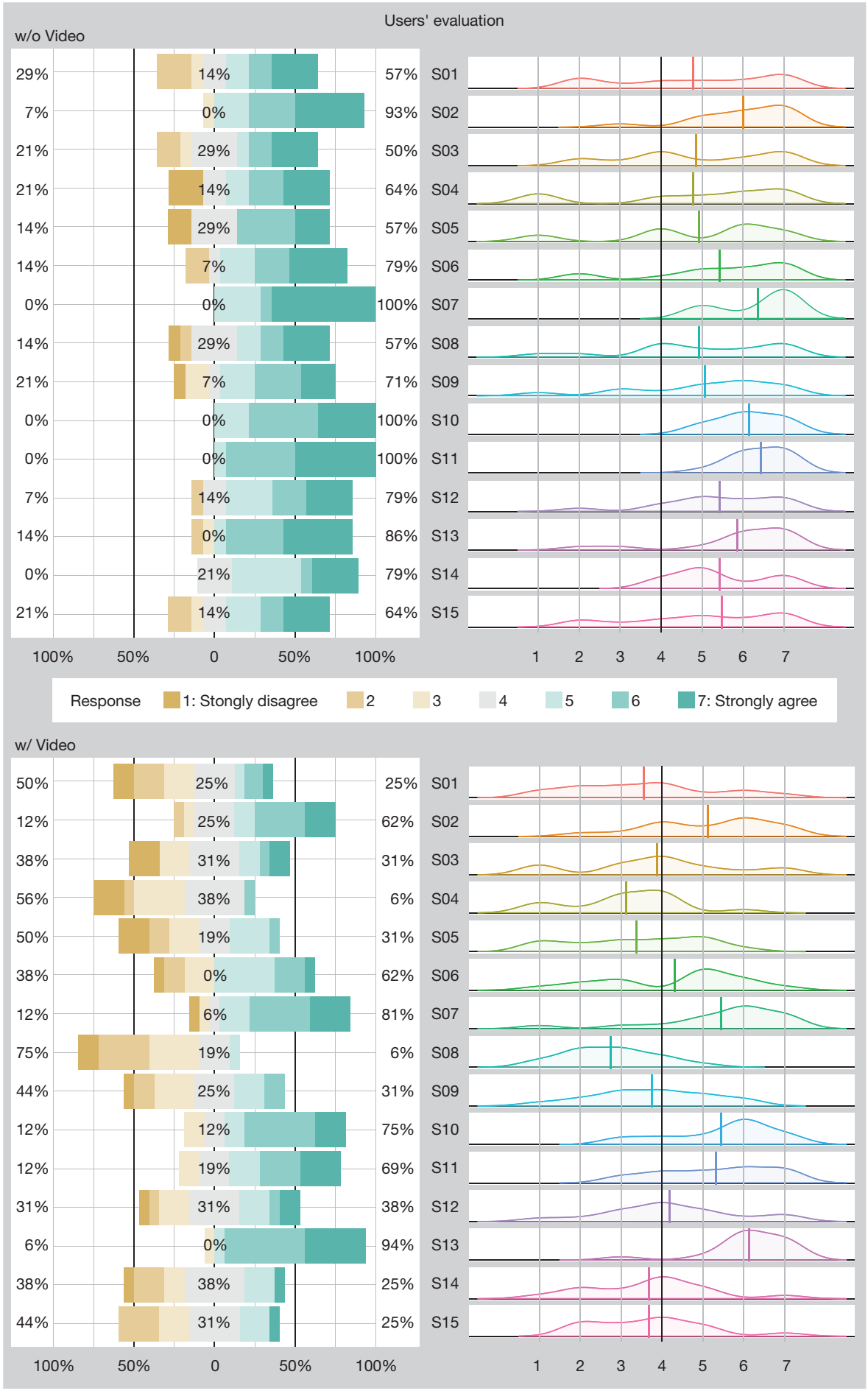


Figure A.23: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by video setting — without live video streaming and with live video streaming*

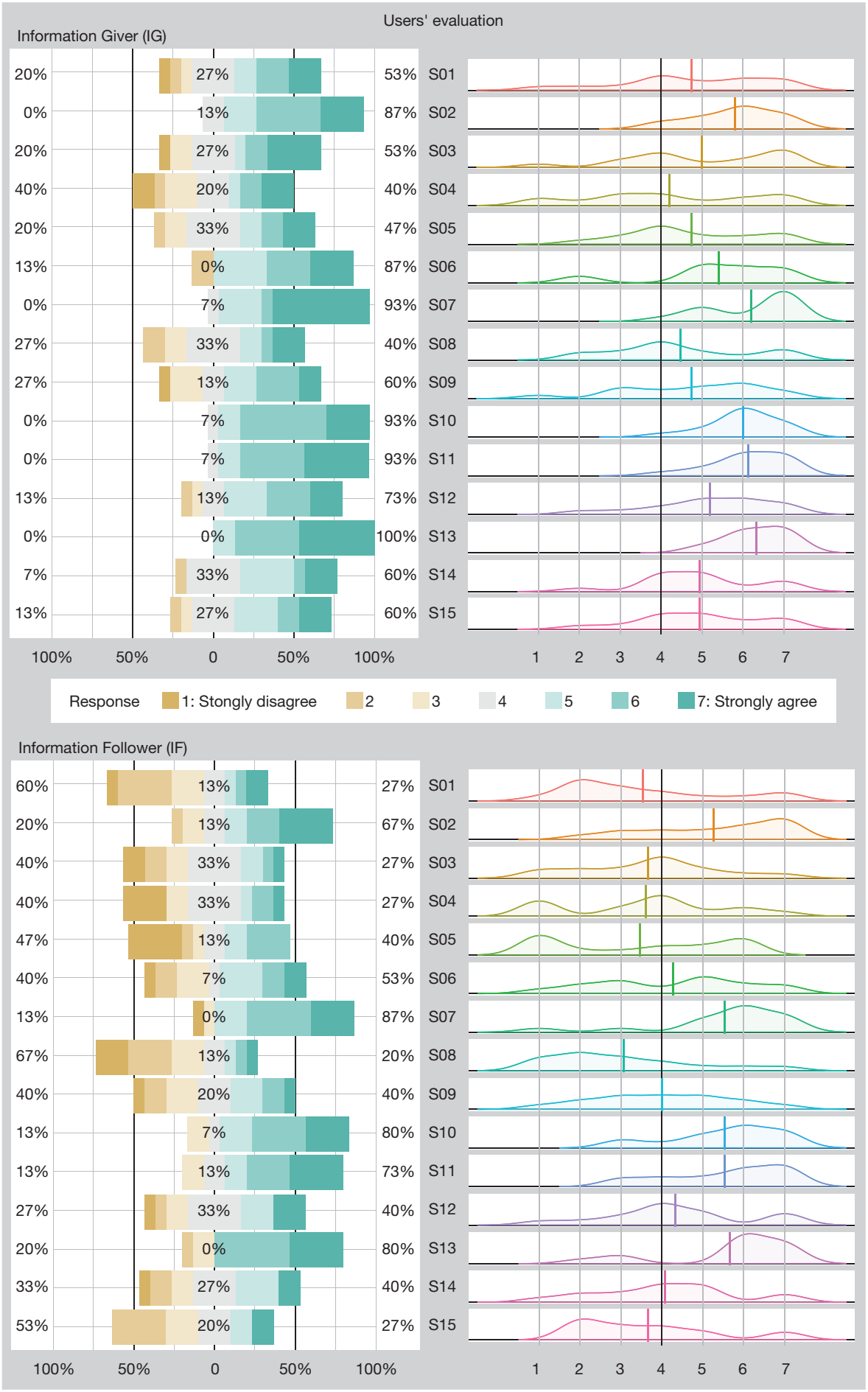


Figure A.24: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by subject role — between Information Giver (IG) and Information Follower (IF)*

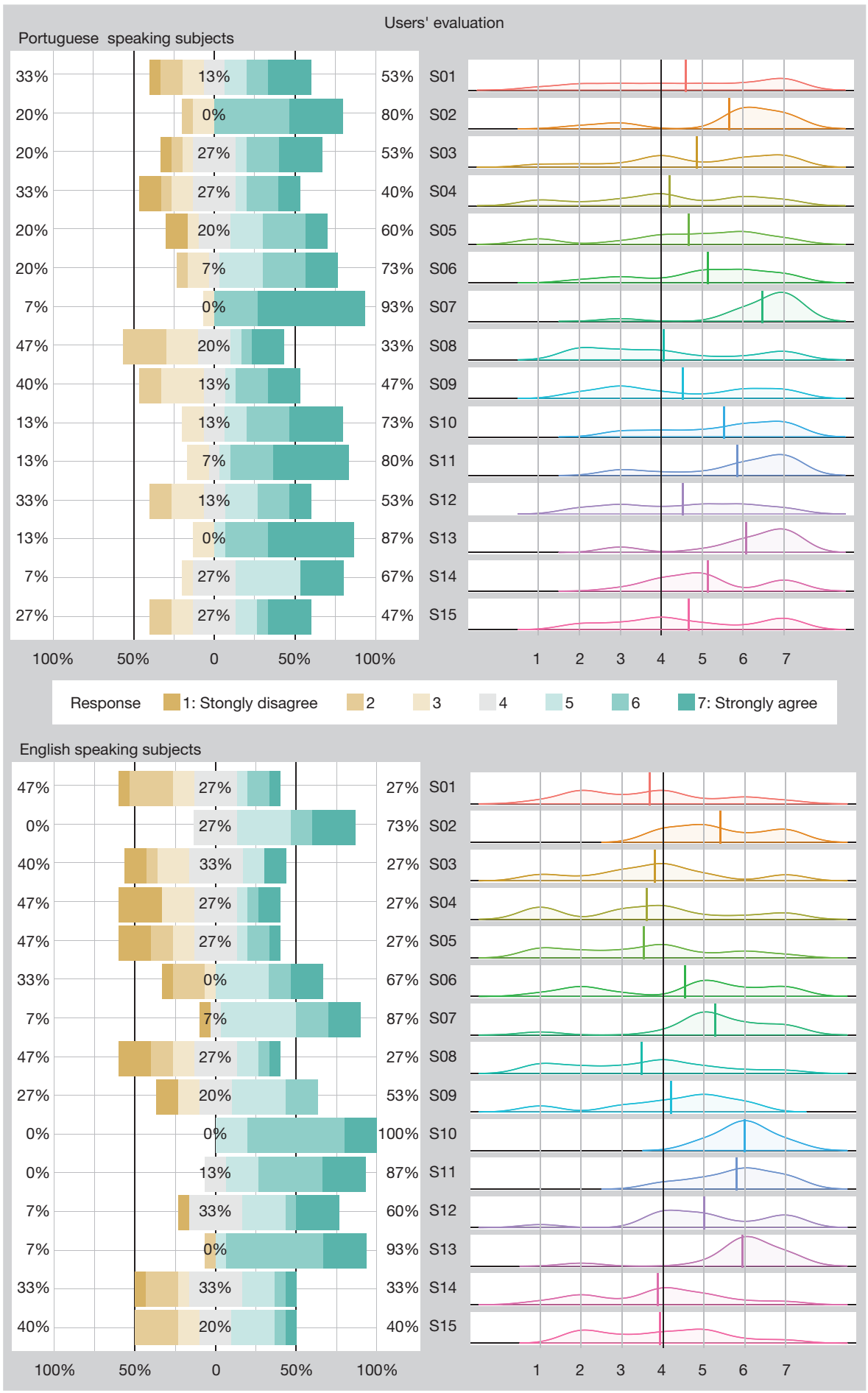


Figure A.25: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by subject role — between Portuguese speaking and English speaking subjects*

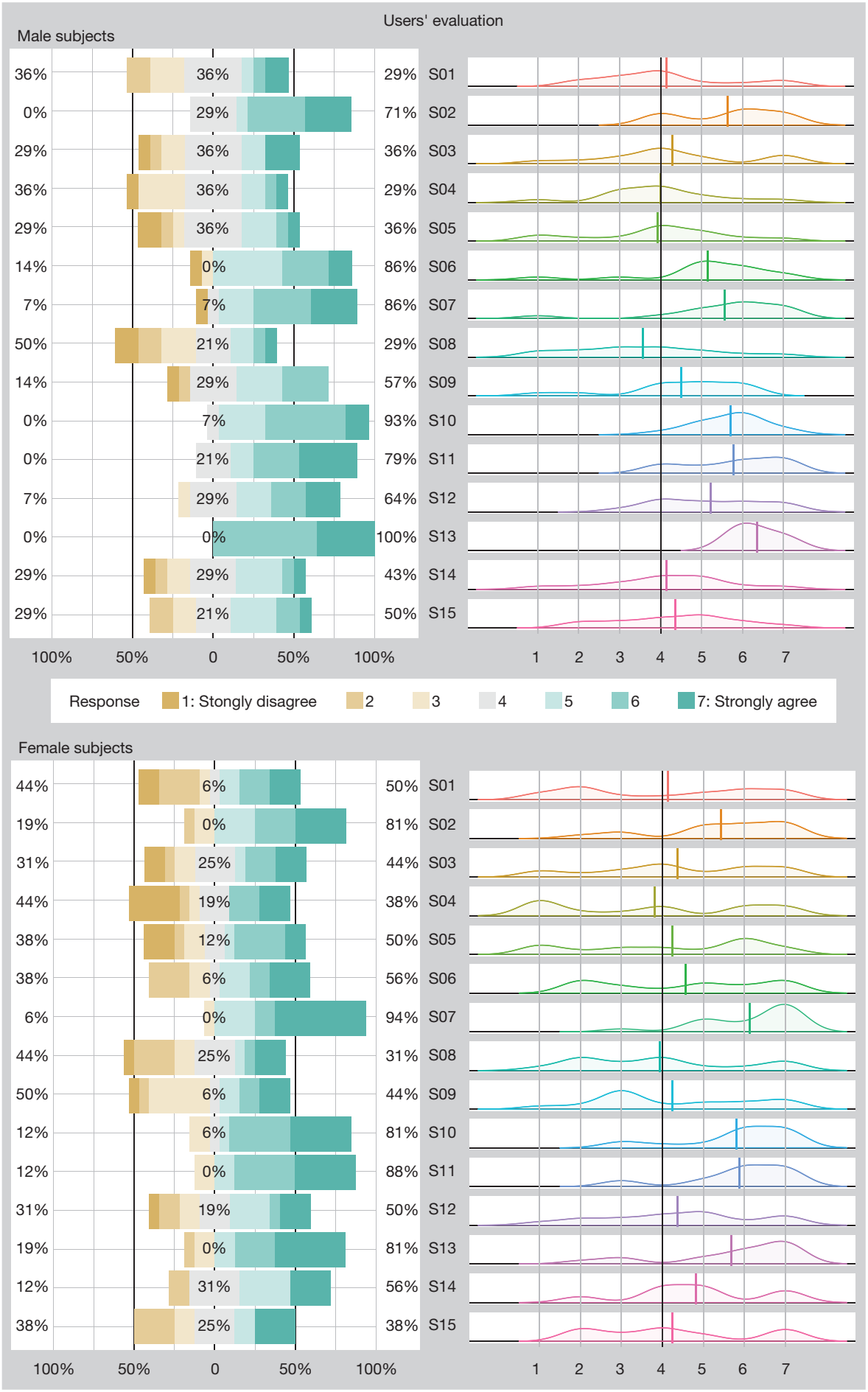


Figure A.26: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by Gender — male (♂) or female (♀)*

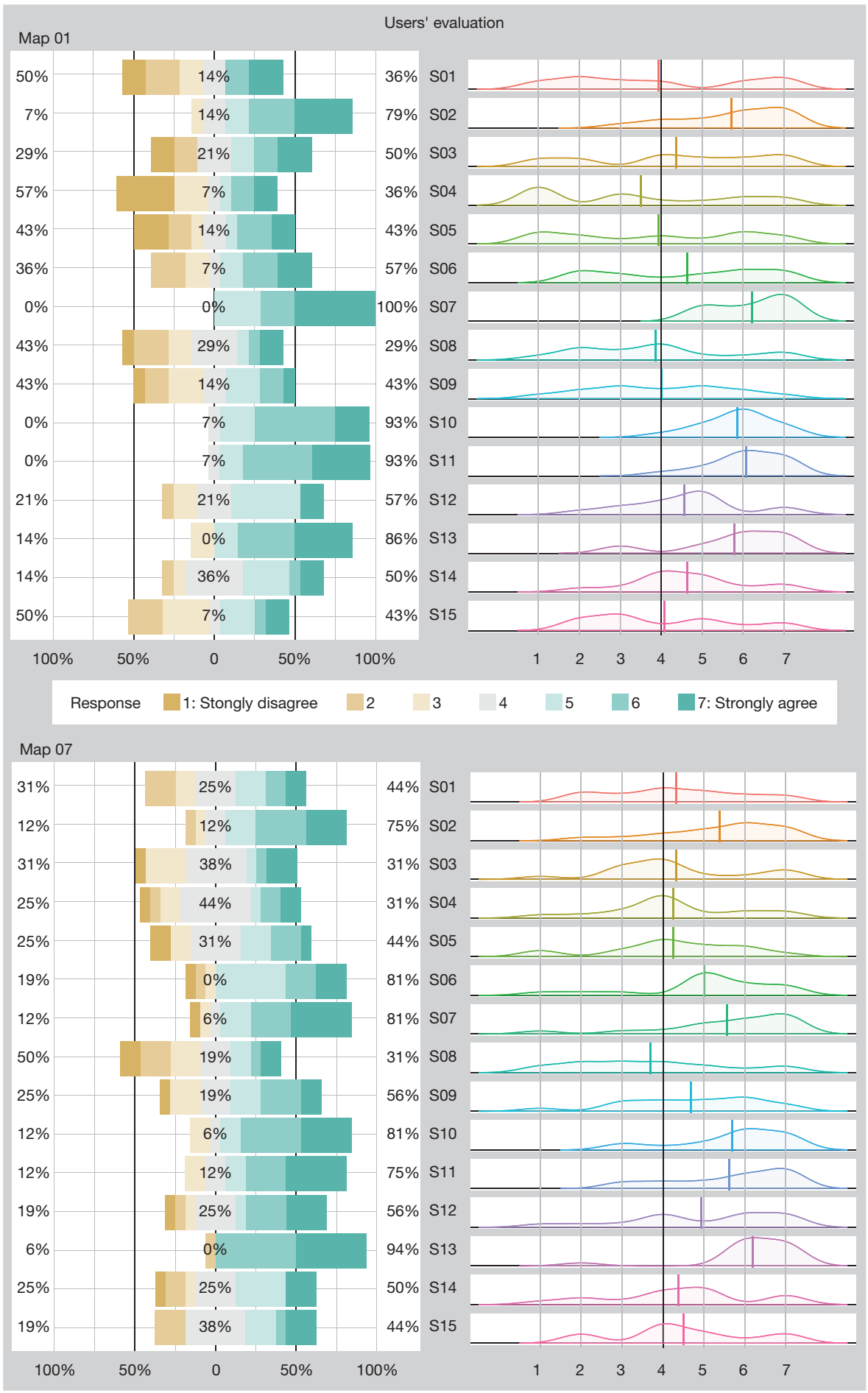


Figure A.27: *Stacked bar chart (left) and density plot (right) of subject survey Likert scale results divided by used map — map 01 or map 07*

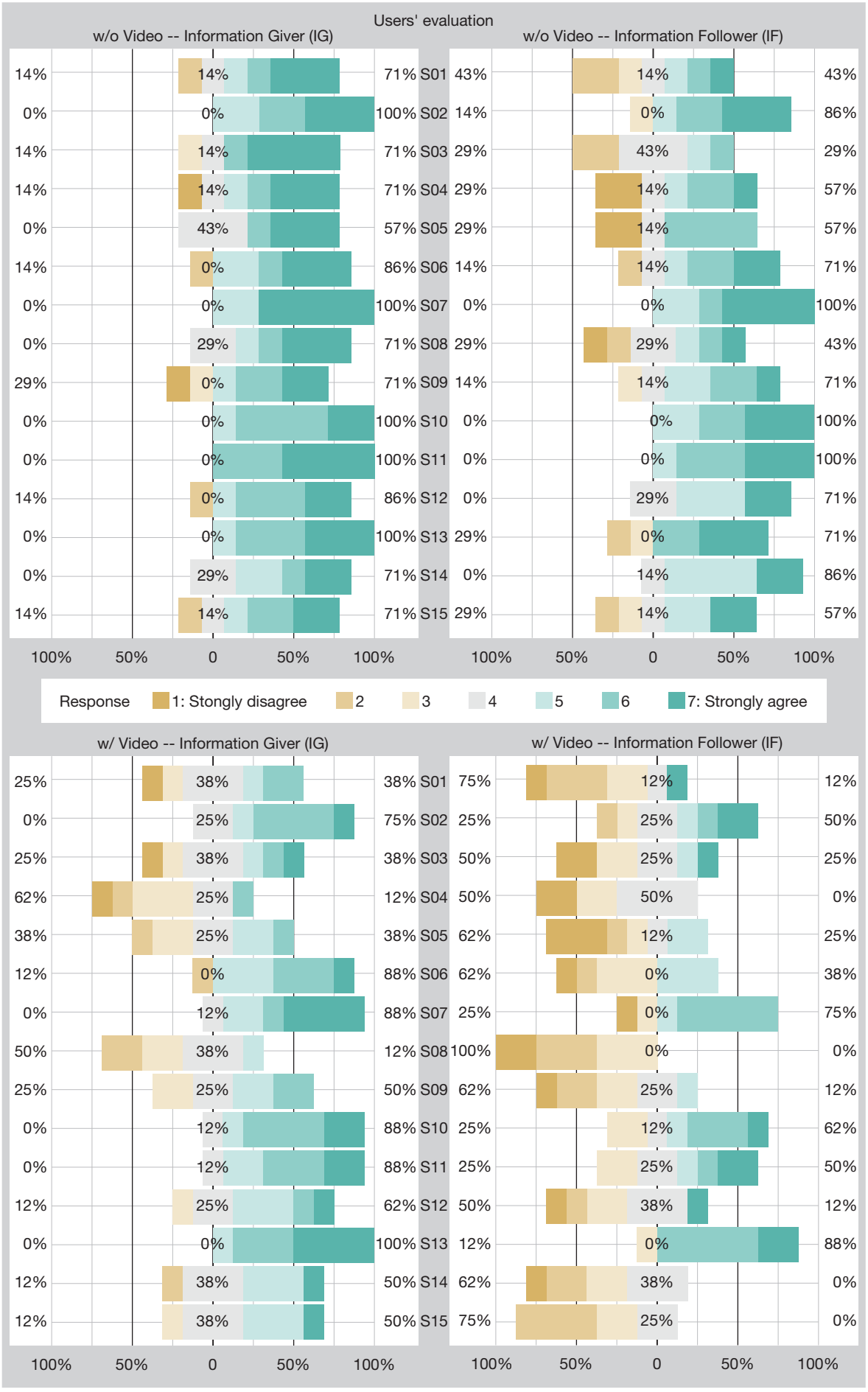


Figure A.28: *Stacked bar chart of subject survey Likert scale results differentiated by the setting and role of the subject — "w/o Video – IG", "w/o Video – IF", "w/ Video – IG" and "w/ Video – IF"*

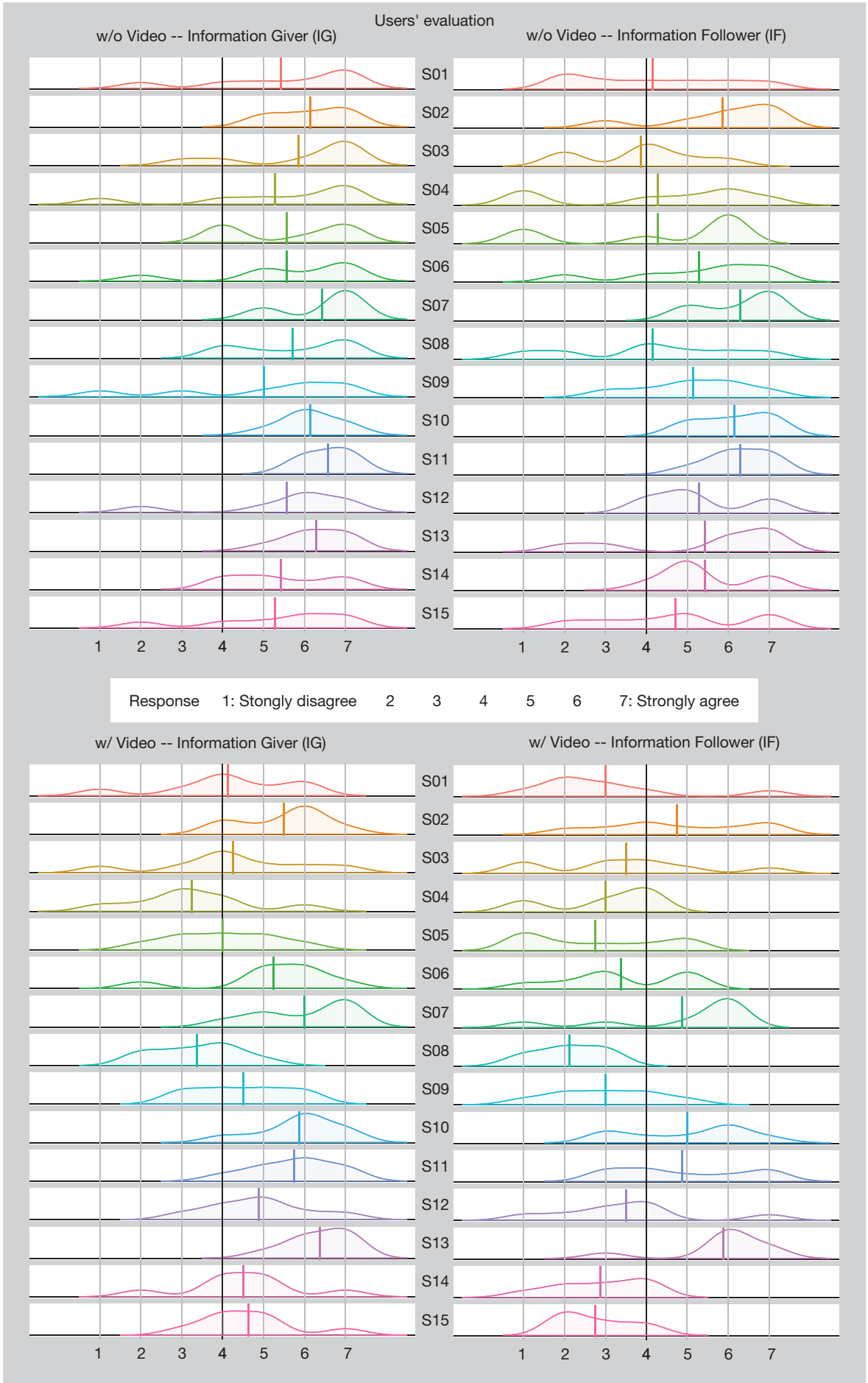


Figure A.29: *Density plot of subject survey Likert scale results differentiated by the setting and role of the subject — "w/o Video – IG", "w/o Video – IF", "w/ Video – IG" and "w/ Video – IF"*

A.6.3 User Survey Significance Test Results

The following pages contain the significance test and effect size results.

Table A.17: Median, mean, standard deviation, significance test and effect size of Likert scale results when the subjects are grouped by setting "w/o Video : w/ Video".

Table A.18: Median, mean, standard deviation, significance test and effect size of Likert scale results when the subjects are grouped by role "Information Giver : Information Follower".

Table A.19: Median, mean, standard deviation, significance test and effect size of Likert scale results when the subjects are grouped by the language spoken "Portuguese (Pt) : English (En)".

Table A.20: Median, mean, standard deviation, significance test and effect size of Likert scale results when the subjects are grouped by gender "male : female".

Table A.21: Median, mean, standard deviation, significance test and effect size of Likert scale results when the subjects are grouped by the map used "Map 01 : Map 07".

Table A.22: Median, mean, standard deviation, significance test and effect size of Likert scale results for settings "w/ Video – Information Follower : w/o Video – Information Giver" — Part 1 of 6.

Table A.23: Median, mean, standard deviation, significance test and effect size of Likert scale results for settings "w/ Video – Information Follower : w/o Video – Information Follower" — Part 2 of 6.

Table A.24: Median, mean, standard deviation, significance test and effect size of Likert scale results for settings "w/ Video – Information Follower : w/ Video – Information Giver" — Part 3 of 6.

Table A.25: Median, mean, standard deviation, significance test and effect size of Likert scale results for settings "w/ Video – Information Follower : w/ Video – Information Giver" — Part 4 of 6.

Table A.26: Median, mean, standard deviation, significance test and effect size of Likert scale results for settings "w/o Video – Information Giver : w/o Video – Information Follower" — Part 5 of 6.

Table A.27: Median, mean, standard deviation, significance test and effect size of Likert scale results for settings "w/o Video – Information Follower : w/ Video – Information Giver" — Part 6 of 6.

Table A.28: Significance test results (Mann-Whitney-Wilcoxon test, Kruskal–Wallis test, and t-test) and effect size (Cohen's d and Cliff's δ) of the comparison between settings "w/ Video" or "w/o Video".

Table A.29: Significance test results (Mann-Whitney-Wilcoxon test, Kruskal–Wallis test, and t-test) and effect size (Cohen's d and Cliff's δ) of the comparison between the role of "Information Giver" and "Information Follower".

Table A.30: Significance test results (Mann-Whitney-Wilcoxon test, Kruskal–Wallis test, and t-test) and effect size (Cohen's d and Cliff's δ) of the comparison between subject speaking "Portuguese (Pt)" and subjects speaking "English (En)".

Table A.31: Significance test results (Mann-Whitney-Wilcoxon test, Kruskal–Wallis test, and t-test) and effect size (Cohen's d and Cliff's δ) of the comparison between "male" and "female" subjects.

Table A.32: Significance test results (Mann-Whitney-Wilcoxon test, Kruskal–Wallis test, and t-test) and effect size (Cohen's d and Cliff's δ) of the comparison between subjects using "Map 01" and subjects using "Map 07".

w/o Video : w/ Video — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	5.0 : 3.5	4.79 : 3.56	1.97 : 1.79	$p < 0.0962$	.	$\delta = 0.3571$	medium
S.02	6.0 : 5.5	6.00 : 5.13	1.18 : 1.50	$p < 0.0944$	.	$\delta = 0.3527$	medium
S.03	4.5 : 4.0	4.86 : 3.88	1.83 : 1.89	$p < 0.1885$	–	$\delta = 0.2813$	small
S.04	5.5 : 3.0	4.79 : 3.13	2.29 : 1.36	$p < 0.0168$	*	$\delta = 0.4911$	large
S.05	6.0 : 3.5	4.93 : 3.38	2.02 : 1.36	$p < 0.0227$	*	$\delta = 0.4732$	medium
S.06	6.0 : 5.0	5.43 : 4.31	1.74 : 1.74	$p < 0.0706$	.	$\delta = 0.3839$	medium
S.07	7.0 : 6.0	6.36 : 5.44	0.93 : 1.63	$p < 0.0750$	.	$\delta = 0.3661$	medium
S.08	5.0 : 3.0	4.93 : 2.75	1.90 : 1.13	$p < 0.0018$	**	$\delta = 0.6384$	large
S.09	5.5 : 4.0	5.07 : 3.75	1.77 : 1.44	$p < 0.0268$	*	$\delta = 0.4598$	medium
S.10	6.0 : 6.0	6.14 : 5.44	0.77 : 1.31	$p < 0.1598$	–	$\delta = 0.2902$	small
S.11	6.5 : 5.5	6.43 : 5.31	0.65 : 1.40	$p < 0.0235$	*	$\delta = 0.4688$	medium
S.12	5.5 : 4.0	5.43 : 4.19	1.45 : 1.64	$p < 0.0325$	*	$\delta = 0.4554$	medium
S.13	6.0 : 6.0	5.86 : 6.13	1.56 : 1.02	$p < 0.9462$	–	$\delta = 0.0179$	negligible
S.14	5.0 : 4.0	5.43 : 3.69	1.16 : 1.49	$p < 0.0019$	**	$\delta = 0.6518$	large
S.15	5.0 : 4.0	5.00 : 3.69	1.80 : 1.40	$p < 0.0401$	*	$\delta = 0.4375$	medium

Table A.17: *Significance test and effect size between Likert scale results of "w/o Video : w/ Video"*

Information Giver : Information Follower — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	5.0 : 3.0	4.73 : 3.53	1.83 : 1.92	$p < 0.0849$	.	$\delta = 0.3689$	medium
S.02	6.0 : 6.0	5.80 : 5.27	1.01 : 1.71	$p < 0.5492$	–	$\delta = 0.1289$	negligible
S.03	5.0 : 4.0	5.00 : 3.67	1.89 : 1.72	$p < 0.0656$	.	$\delta = 0.3911$	medium
S.04	4.0 : 4.0	4.20 : 3.60	2.08 : 1.96	$p < 0.4859$	–	$\delta = 0.1511$	small
S.05	4.0 : 4.0	4.73 : 3.47	1.58 : 2.13	$p < 0.1245$	–	$\delta = 0.3289$	small
S.06	6.0 : 5.0	5.40 : 4.27	1.59 : 1.87	$p < 0.0895$	.	$\delta = 0.3600$	medium
S.07	7.0 : 6.0	6.20 : 5.53	1.08 : 1.64	$p < 0.2194$	–	$\delta = 0.2533$	small
S.08	4.0 : 3.0	4.47 : 3.07	1.68 : 1.83	$p < 0.0318$	*	$\delta = 0.4578$	medium
S.09	5.0 : 4.0	4.73 : 4.00	1.72 : 1.69	$p < 0.2302$	–	$\delta = 0.2578$	small
S.10	6.0 : 6.0	6.00 : 5.53	0.85 : 1.36	$p < 0.4174$	–	$\delta = 0.1689$	small
S.11	6.0 : 6.0	6.13 : 5.53	0.92 : 1.46	$p < 0.3176$	–	$\delta = 0.2089$	small
S.12	5.0 : 4.0	5.20 : 4.33	1.47 : 1.76	$p < 0.1392$	–	$\delta = 0.3156$	small
S.13	6.0 : 6.0	6.33 : 5.67	0.72 : 1.63	$p < 0.3817$	–	$\delta = 0.1778$	small
S.14	5.0 : 4.0	4.93 : 4.07	1.39 : 1.71	$p < 0.1460$	–	$\delta = 0.3067$	small
S.15	5.0 : 3.0	4.93 : 3.67	1.49 : 1.72	$p < 0.0348$	*	$\delta = 0.4489$	medium

Table A.18: *Significance test and effect size between Likert scale results of "Information Giver (IG) : Information Follower (IF)"*

Portuguese speaking subject (Pt) : English speaking subject (En) — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	5.0 : 4.0	4.60 : 3.67	2.06 : 1.76	$p < 0.2000$	–	$\delta = 0.2756$	small
S.02	6.0 : 5.0	5.67 : 5.40	1.63 : 1.18	$p < 0.3465$	–	$\delta = 0.2000$	small
S.03	5.0 : 4.0	4.87 : 3.80	1.92 : 1.78	$p < 0.1276$	–	$\delta = 0.3244$	small
S.04	4.0 : 4.0	4.20 : 3.60	1.97 : 2.06	$p < 0.4223$	–	$\delta = 0.1733$	small
S.05	5.0 : 4.0	4.67 : 3.53	1.88 : 1.92	$p < 0.1051$	–	$\delta = 0.3467$	medium
S.06	5.0 : 5.0	5.13 : 4.53	1.55 : 2.03	$p < 0.4448$	–	$\delta = 0.1644$	small
S.07	7.0 : 5.0	6.47 : 5.27	1.06 : 1.49	$p < 0.0044$	**	$\delta = 0.5822$	large
S.08	4.0 : 4.0	4.07 : 3.47	1.91 : 1.85	$p < 0.4743$	–	$\delta = 0.1556$	small
S.09	4.0 : 5.0	4.53 : 4.20	1.85 : 1.61	$p < 0.7048$	–	$\delta = 0.0844$	negligible
S.10	6.0 : 6.0	5.53 : 6.00	1.46 : 0.65	$p < 0.6142$	–	$\delta = 0.1067$	negligible
S.11	6.0 : 6.0	5.87 : 5.80	1.46 : 1.01	$p < 0.5286$	–	$\delta = 0.1333$	negligible
S.12	5.0 : 5.0	4.53 : 5.00	1.68 : 1.65	$p < 0.4345$	–	$\delta = 0.1689$	small
S.13	7.0 : 6.0	6.07 : 5.93	1.39 : 1.22	$p < 0.3578$	–	$\delta = 0.1867$	small
S.14	5.0 : 4.0	5.13 : 3.87	1.30 : 1.64	$p < 0.0381$	*	$\delta = 0.4356$	medium
S.15	4.0 : 4.0	4.67 : 3.93	1.80 : 1.58	$p < 0.3009$	–	$\delta = 0.2222$	small

Table A.19: *Significance test and effect size between Likert scale results of "Portuguese speaking subject (Pt) : English speaking subject (En)"*

male : female — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	4.0 : 4.5	4.14 : 4.13	1.61 : 2.25	$p < 0.9329$	–	$\delta = 0.0223$	negligible
S.02	6.0 : 6.0	5.64 : 5.44	1.22 : 1.59	$p < 0.8807$	–	$\delta = 0.0357$	negligible
S.03	4.0 : 4.0	4.29 : 4.38	1.82 : 2.03	$p < 0.9155$	–	$\delta = 0.0268$	negligible
S.04	4.0 : 4.0	4.00 : 3.81	1.47 : 2.43	$p < 0.8489$	–	$\delta = 0.0446$	negligible
S.05	4.0 : 4.5	3.93 : 4.25	1.73 : 2.18	$p < 0.5834$	–	$\delta = 0.1205$	negligible
S.06	5.0 : 5.0	5.14 : 4.56	1.56 : 2.00	$p < 0.5096$	–	$\delta = 0.1429$	negligible
S.07	6.0 : 7.0	5.57 : 6.13	1.60 : 1.20	$p < 0.2623$	–	$\delta = 0.2321$	small
S.08	3.5 : 4.0	3.57 : 3.94	1.79 : 1.98	$p < 0.6886$	–	$\delta = 0.0893$	negligible
S.09	5.0 : 3.5	4.50 : 4.25	1.51 : 1.91	$p < 0.6275$	–	$\delta = 0.1071$	negligible
S.10	6.0 : 6.0	5.71 : 5.81	0.83 : 1.38	$p < 0.3796$	–	$\delta = 0.1830$	small
S.11	6.0 : 6.0	5.79 : 5.88	1.19 : 1.31	$p < 0.7771$	–	$\delta = 0.0625$	negligible
S.12	5.0 : 4.5	5.21 : 4.38	1.31 : 1.86	$p < 0.2275$	–	$\delta = 0.2589$	small
S.13	6.0 : 6.0	6.36 : 5.69	0.50 : 1.66	$p < 0.5589$	–	$\delta = 0.1205$	negligible
S.14	4.0 : 5.0	4.14 : 4.81	1.56 : 1.60	$p < 0.2938$	–	$\delta = 0.2232$	small
S.15	4.5 : 4.0	4.36 : 4.25	1.50 : 1.91	$p < 0.7510$	–	$\delta = 0.0714$	negligible

Table A.20: *Significance test and effect size between Likert scale results of "Male : Female"*

Map 01 : Map 07 — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	3.5 : 4.0	3.93 : 4.31	2.27 : 1.66	$p < 0.5555$	–	$\delta = 0.1295$	negligible
S.02	6.0 : 6.0	5.71 : 5.38	1.33 : 1.50	$p < 0.5483$	–	$\delta = 0.1295$	negligible
S.03	4.5 : 4.0	4.36 : 4.31	2.17 : 1.70	$p < 0.7991$	–	$\delta = 0.0580$	negligible
S.04	3.0 : 4.0	3.50 : 4.25	2.35 : 1.65	$p < 0.2899$	–	$\delta = 0.2277$	small
S.05	4.0 : 4.0	3.93 : 4.25	2.27 : 1.69	$p < 0.7839$	–	$\delta = 0.0625$	negligible
S.06	5.0 : 5.0	4.64 : 5.00	1.95 : 1.71	$p < 0.7336$	–	$\delta = 0.0759$	negligible
S.07	6.5 : 6.0	6.21 : 5.56	0.89 : 1.71	$p < 0.3675$	–	$\delta = 0.1875$	small
S.08	4.0 : 3.5	3.86 : 3.69	1.88 : 1.92	$p < 0.8165$	–	$\delta = 0.0536$	negligible
S.09	4.0 : 5.0	4.00 : 4.69	1.75 : 1.66	$p < 0.2727$	–	$\delta = 0.2366$	small
S.10	6.0 : 6.0	5.86 : 5.69	0.86 : 1.35	$p < 0.9825$	–	$\delta = 0.0089$	negligible
S.11	6.0 : 6.0	6.07 : 5.63	0.92 : 1.45	$p < 0.5566$	–	$\delta = 0.1250$	negligible
S.12	5.0 : 5.5	4.57 : 4.94	1.40 : 1.88	$p < 0.4717$	–	$\delta = 0.1563$	small
S.13	6.0 : 6.0	5.79 : 6.19	1.37 : 1.22	$p < 0.3686$	–	$\delta = 0.1830$	small
S.14	4.5 : 4.5	4.64 : 4.38	1.39 : 1.78	$p < 0.7642$	–	$\delta = 0.0670$	negligible
S.15	3.5 : 4.0	4.07 : 4.50	1.77 : 1.67	$p < 0.4984$	–	$\delta = 0.1473$	small

Table A.21: *Significance test and effect size between Likert scale results of "map 01 : map 07"*

w/ Video – IF : w/o Video – IG — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	2.5 : 6.0	3.00 : 5.43	1.85 : 1.90	$p < 0.0447$	*	$\delta = 0.6250$	large
S.02	4.5 : 6.0	4.75 : 6.14	1.83 : 0.90	$p < 0.1374$	–	$\delta = 0.4643$	medium
S.03	3.5 : 7.0	3.50 : 5.86	2.00 : 1.68	$p < 0.0499$	*	$\delta = 0.6071$	large
S.04	3.5 : 6.0	3.00 : 5.29	1.31 : 2.21	$p < 0.0279$	*	$\delta = 0.6786$	large
S.05	2.5 : 6.0	2.75 : 5.57	1.75 : 1.51	$p < 0.0185$	*	$\delta = 0.7321$	large
S.06	3.0 : 6.0	3.38 : 5.57	1.51 : 1.81	$p < 0.0324$	*	$\delta = 0.6607$	large
S.07	6.0 : 7.0	4.88 : 6.43	1.89 : 0.98	$p < 0.0467$	*	$\delta = 0.6071$	large
S.08	2.0 : 6.0	2.13 : 5.71	0.83 : 1.38	$p < 0.0013$	**	$\delta = 1.0000$	large
S.09	3.0 : 6.0	3.00 : 5.00	1.31 : 2.24	$p < 0.0703$	.	$\delta = 0.5714$	large
S.10	5.5 : 6.0	5.00 : 6.14	1.51 : 0.69	$p < 0.1408$	–	$\delta = 0.4464$	medium
S.11	4.5 : 7.0	4.88 : 6.57	1.64 : 0.53	$p < 0.0530$	.	$\delta = 0.5893$	large
S.12	3.5 : 6.0	3.50 : 5.57	1.77 : 1.72	$p < 0.0608$	.	$\delta = 0.5893$	large
S.13	6.0 : 6.0	5.88 : 6.29	1.25 : 0.76	$p < 0.6080$	–	$\delta = 0.1607$	small
S.14	3.0 : 5.0	2.88 : 5.43	1.13 : 1.27	$p < 0.0038$	**	$\delta = 0.8929$	large
S.15	2.5 : 6.0	2.75 : 5.29	0.89 : 1.80	$p < 0.0150$	*	$\delta = 0.7500$	large

Table A.22: *Significance test and effect size between Likert scale results of "w/ Video – Information Follower (IF) : w/o Video – Information Giver (IG)" — Part 1 of 6*

w/ Video – Information Follower : w/o Video – Information Follower — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	2.5 : 4.0	3.00 : 4.14	1.85 : 1.95	$p < 0.2602$	–	$\delta = 0.3571$	medium
S.02	4.5 : 6.0	4.75 : 5.86	1.83 : 1.46	$p < 0.2598$	–	$\delta = 0.3571$	medium
S.03	3.5 : 4.0	3.50 : 3.86	2.00 : 1.46	$p < 0.6361$	–	$\delta = 0.1607$	small
S.04	3.5 : 5.0	3.00 : 4.29	1.31 : 2.43	$p < 0.1705$	–	$\delta = 0.3571$	medium
S.05	2.5 : 6.0	2.75 : 4.29	1.75 : 2.36	$p < 0.1527$	–	$\delta = 0.3750$	medium
S.06	3.0 : 6.0	3.38 : 5.29	1.51 : 1.80	$p < 0.0525$	.	$\delta = 0.5357$	large
S.07	6.0 : 7.0	4.88 : 6.29	1.89 : 0.95	$p < 0.0899$	.	$\delta = 0.4464$	medium
S.08	2.0 : 4.0	2.13 : 4.14	0.83 : 2.12	$p < 0.0596$	.	$\delta = 0.5000$	large
S.09	3.0 : 5.0	3.00 : 5.14	1.31 : 1.35	$p < 0.0163$	*	$\delta = 0.6964$	large
S.10	5.5 : 6.0	5.00 : 6.14	1.51 : 0.90	$p < 0.1515$	–	$\delta = 0.4464$	medium
S.11	4.5 : 6.0	4.88 : 6.29	1.64 : 0.76	$p < 0.1073$	–	$\delta = 0.5000$	large
S.12	3.5 : 5.0	3.50 : 5.29	1.77 : 1.25	$p < 0.0279$	*	$\delta = 0.6786$	large
S.13	6.0 : 6.0	5.88 : 5.43	1.25 : 2.07	$p < 1.0000$	–	$\delta = 0.0179$	negligible
S.14	3.0 : 5.0	2.88 : 5.43	1.13 : 1.13	$p < 0.0021$	**	$\delta = 0.8393$	large
S.15	2.5 : 5.0	2.75 : 4.71	0.89 : 1.89	$p < 0.0374$	*	$\delta = 0.5714$	large

Table A.23: *Significance test and effect size between Likert scale results of "w/ Video – Information Follower (IF) : w/o Video – Information Follower (IF)" — Part 2 of 6*

w/ Video – Information Follower : w/ Video – Information Giver — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	2.5 : 4.0	3.00 : 4.13	1.85 : 1.64	$p < 0.1501$	–	$\delta = 0.4531$	medium
S.02	4.5 : 6.0	4.75 : 5.50	1.83 : 1.07	$p < 0.4504$	–	$\delta = 0.2344$	small
S.03	3.5 : 4.0	3.50 : 4.25	2.00 : 1.83	$p < 0.4205$	–	$\delta = 0.2500$	small
S.04	3.5 : 3.0	3.00 : 3.25	1.31 : 1.49	$p < 1.0000$	–	$\delta = 0.0000$	negligible
S.05	2.5 : 4.0	2.75 : 4.00	1.75 : 1.31	$p < 0.1648$	–	$\delta = 0.4219$	medium
S.06	3.0 : 5.5	3.38 : 5.25	1.51 : 1.49	$p < 0.0260$	*	$\delta = 0.6563$	large
S.07	6.0 : 6.5	4.88 : 6.00	1.89 : 1.20	$p < 0.1908$	–	$\delta = 0.3906$	medium
S.08	2.0 : 3.5	2.13 : 3.38	0.83 : 1.06	$p < 0.0341$	*	$\delta = 0.6250$	large
S.09	3.0 : 4.5	3.00 : 4.50	1.31 : 1.20	$p < 0.0476$	*	$\delta = 0.5938$	large
S.10	5.5 : 6.0	5.00 : 5.88	1.31 : 0.99	$p < 0.2470$	–	$\delta = 0.3438$	medium
S.11	4.5 : 6.0	4.88 : 5.75	1.64 : 1.04	$p < 0.2831$	–	$\delta = 0.3281$	small
S.12	3.5 : 5.0	3.50 : 4.88	1.77 : 1.25	$p < 0.0681$	.	$\delta = 0.5469$	large
S.13	6.0 : 6.5	5.88 : 6.38	1.77 : 0.74	$p < 0.4183$	–	$\delta = 0.2344$	small
S.14	3.0 : 4.5	2.88 : 4.50	1.13 : 1.41	$p < 0.0226$	*	$\delta = 0.6719$	large
S.15	2.5 : 4.5	2.75 : 4.63	0.89 : 1.19	$p < 0.0058$	**	$\delta = 0.8125$	large

Table A.24: *Significance test and effect size between Likert scale results of "w/ Video – Information Follower (IF) : w/ Video – Information Giver (IG)" — Part 3 of 6*

w/o Video – Information Giver : w/ Video – Information Giver — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	6.0 : 4.0	5.43 : 4.13	1.90 : 1.64	$p < 0.1411$	–	$\delta = 0.4107$	medium
S.02	6.0 : 6.0	6.14 : 5.50	0.90 : 1.07	$p < 0.2751$	–	$\delta = 0.3393$	medium
S.03	7.0 : 4.0	5.86 : 4.25	1.68 : 1.83	$p < 0.1213$	–	$\delta = 0.4821$	large
S.04	6.0 : 3.0	5.29 : 3.25	2.21 : 1.49	$p < 0.0531$	.	$\delta = 0.5536$	large
S.05	6.0 : 4.0	5.57 : 4.00	1.51 : 1.31	$p < 0.0752$	.	$\delta = 0.5000$	large
S.06	6.0 : 5.5	5.57 : 5.25	1.81 : 1.49	$p < 0.5478$	–	$\delta = 0.1964$	small
S.07	7.0 : 6.5	6.43 : 6.00	0.98 : 1.20	$p < 0.4676$	–	$\delta = 0.2143$	small
S.08	6.0 : 3.5	5.71 : 3.38	1.38 : 1.06	$p < 0.0091$	**	$\delta = 0.7321$	large
S.09	6.0 : 4.5	5.00 : 4.50	2.24 : 1.20	$p < 0.3768$	–	$\delta = 0.2500$	small
S.10	6.0 : 6.0	6.14 : 5.88	0.69 : 0.99	$p < 0.7032$	–	$\delta = 0.1250$	negligible
S.11	7.0 : 6.0	6.57 : 5.75	0.53 : 1.04	$p < 0.1074$	–	$\delta = 0.4821$	large
S.12	6.0 : 5.0	5.57 : 4.88	1.72 : 1.25	$p < 0.2138$	–	$\delta = 0.3929$	medium
S.13	6.0 : 6.5	6.29 : 6.38	0.76 : 0.74	$p < 1.0000$	–	$\delta = 0.8494$	negligible
S.14	5.0 : 4.5	5.43 : 4.50	1.27 : 1.41	$p < 0.2521$	–	$\delta = 0.3571$	medium
S.15	6.0 : 4.5	5.29 : 4.63	1.80 : 1.19	$p < 0.3142$	–	$\delta = 0.3214$	small

Table A.25: *Significance test and effect size between Likert scale results of "w/o Video – Information Giver (IG) : w/ Video – Information Giver (IG)" — Part 4 of 6*

w/o Video – Information Giver : w/o Video – Information Follower — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	6.0 : 4.0	5.43 : 4.14	1.90 : 1.95	$p < 0.2412$	–	$\delta = 0.3878$	medium
S.02	6.0 : 6.0	6.14 : 5.86	0.90 : 1.46	$p < 0.9461$	–	$\delta = 0.0408$	negligible
S.03	7.0 : 4.0	5.86 : 3.86	1.68 : 1.46	$p < 0.0495$	*	$\delta = 0.6327$	large
S.04	6.0 : 5.0	5.29 : 4.29	2.21 : 2.43	$p < 0.3956$	–	$\delta = 0.2857$	small
S.05	6.0 : 6.0	5.57 : 4.29	1.51 : 2.36	$p < 0.2583$	–	$\delta = 0.3673$	medium
S.06	6.0 : 6.0	5.57 : 5.29	1.81 : 1.80	$p < 0.7413$	–	$\delta = 0.1224$	negligible
S.07	7.0 : 7.0	6.43 : 6.29	0.98 : 0.95	$p < 0.7624$	–	$\delta = 0.1020$	negligible
S.08	6.0 : 4.0	5.71 : 4.14	1.38 : 2.12	$p < 0.1690$	–	$\delta = 0.4490$	medium
S.09	6.0 : 5.0	5.00 : 5.14	2.24 : 1.35	$p < 0.8448$	–	$\delta = 0.0816$	negligible
S.10	6.0 : 6.0	6.14 : 6.14	0.69 : 0.90	$p < 1.0000$	–	$\delta = 0.0204$	negligible
S.11	7.0 : 6.0	6.57 : 6.29	0.53 : 0.76	$p < 0.5203$	–	$\delta = 0.2041$	small
S.12	6.0 : 5.0	5.57 : 5.29	1.72 : 1.25	$p < 0.4697$	–	$\delta = 0.2449$	small
S.13	6.0 : 6.0	6.29 : 5.43	0.76 : 2.07	$p < 0.7333$	–	$\delta = 0.1224$	negligible
S.14	5.0 : 5.0	5.43 : 5.43	1.27 : 1.13	$p < 1.0000$	–	$\delta = 0.0204$	negligible
S.15	6.0 : 5.0	5.29 : 4.71	1.80 : 1.89	$p < 0.6024$	–	$\delta = 0.1837$	small

Table A.26: *Significance test and effect size between Likert scale results of "w/o Video – Information Giver (IG) : w/o Video – Information Follower (IF)" — Part 5 of 6*

w/o Video – Information Follower : w/ Video – Information Giver — Survey result comparison							
Statement	*Mdn*	*M*	*SD*	wilcox.test	sig.	cliff.delta	effect size
S.01	4.0 : 4.0	4.14 : 4.13	1.95 : 1.64	$p < 1.0000$	–	$\delta = 0.0357$	negligible
S.02	6.0 : 6.0	5.86 : 5.50	1.46 : 1.07	$p < 0.4319$	–	$\delta = 0.2500$	small
S.03	4.0 : 4.0	3.86 : 4.25	1.46 : 1.83	$p < 0.7191$	–	$\delta = 0.1250$	negligible
S.04	5.0 : 3.0	4.29 : 3.25	2.43 : 1.49	$p < 0.3183$	–	$\delta = 0.3036$	small
S.05	6.0 : 4.0	4.29 : 4.00	2.36 : 1.31	$p < 0.5141$	–	$\delta = 0.2143$	small
S.06	6.0 : 5.5	5.29 : 5.25	1.80 : 1.49	$p < 0.8577$	–	$\delta = 0.0714$	negligible
S.07	7.0 : 6.5	6.29 : 6.00	0.95 : 1.20	$p < 0.7032$	–	$\delta = 0.1250$	negligible
S.08	4.0 : 3.5	4.14 : 3.38	2.12 : 1.06	$p < 0.4068$	–	$\delta = 0.2321$	small
S.09	5.0 : 4.5	5.14 : 4.50	1.35 : 1.20	$p < 0.3732$	–	$\delta = 0.2679$	small
S.10	6.0 : 6.0	6.14 : 5.88	0.90 : 0.99	$p < 0.6685$	–	$\delta = 0.1429$	negligible
S.11	6.0 : 6.0	6.29 : 5.75	0.76 : 1.04	$p < 0.3277$	–	$\delta = 0.3036$	small
S.12	5.0 : 5.0	5.29 : 4.88	1.25 : 1.25	$p < 0.6280$	–	$\delta = 0.1607$	small
S.13	6.0 : 6.5	5.43 : 6.38	2.07 : 0.74	$p < 0.5754$	–	$\delta = 0.1786$	small
S.14	5.0 : 4.5	5.43 : 4.50	1.13 : 1.41	$p < 0.1735$	–	$\delta = 0.4107$	medium
S.15	5.0 : 4.5	4.71 : 4.63	1.89 : 1.19	$p < 0.9523$	–	$\delta = 0.0357$	negligible

Table A.27: *Significance test and effect size between Likert scale results of "w/o Video – Information Follower (IF) : w/ Video – Information Giver (IG)" — Part 6 of 6*

Statement	wilcox.test	sig.	kruskai.test	sig.	t.test	sig.	cohen.d	ES	cliff.delta	ES
S.01	$p < 0.0962$	.	$p < 0.0921$	.	$p < 0.0878$	.	$d = 0.6529$	med	$\delta = 0.3571$	med
S.02	$p < 0.0944$	.	$p < 0.0903$	.	$p < 0.0847$	.	$d = 0.6436$	med	$\delta = 0.3527$	med
S.03	$p < 0.1885$	–	$p < 0.1815$	–	$p < 0.1606$	–	$d = 0.5264$	med	$\delta = 0.2813$	small
S.04	$p < 0.0168$	*	$p < 0.0158$	*	$p < 0.0277$	*	$d = 0.8964$	large	$\delta = 0.4911$	large
S.05	$p < 0.0227$	*	$p < 0.0215$	*	$p < 0.0301$	*	$d = 0.8539$	large	$\delta = 0.4732$	med
S.06	$p < 0.0706$	.	$p < 0.0673$	.	$p < 0.0910$	.	$d = 0.6411$	med	$\delta = 0.3839$	med
S.07	$p < 0.0750$	.	$p < 0.0715$	.	$p < 0.0659$	.	$d = 0.6804$	med	$\delta = 0.3661$	med
S.08	$p < 0.0018$	**	$p < 0.0017$	**	$p < 0.0012$	**	$d = 1.4198$	large	$\delta = 0.6384$	large
S.09	$p < 0.0268$	*	$p < 0.0254$	*	$p < 0.0356$	*	$d = 0.8245$	large	$\delta = 0.4598$	med
S.10	$p < 0.1598$	–	$p < 0.1534$	–	$p < 0.0812$	.	$d = 0.6434$	med	$\delta = 0.2902$	small
S.11	$p < 0.0235$	*	$p < 0.0222$	*	$p < 0.0092$	*	$d = 1.0002$	large	$\delta = 0.4688$	med
S.12	$p < 0.0325$	*	$p < 0.0308$	*	$p < 0.0365$	*	$d = 0.7972$	med	$\delta = 0.4554$	med
S.13	$p < 0.9462$	–	$p < 0.9284$	–	$p < 0.5900$	–	$d = 0.2057$	small	$\delta = 0.0179$	neg
S.14	$p < 0.0019$	**	$p < 0.0018$	**	$p < 0.0013$	**	$d = 1.2918$	large	$\delta = 0.6518$	large
S.15	$p < 0.0401$	*	$p < 0.0381$	*	$p < 0.0369$	*	$d = 0.8217$	large	$\delta = 0.4375$	med

Table A.28: *Significance test and effect size between Likert results from subjects w/o live video streaming and w/ live video streaming*

Statement	wilcox.test	sig.	kruskai.test	sig.	t.test	sig.	cohen.d	ES	cliff.delta	ES
S.01	$p < 0.0849$	.	$p < 0.0812$	.	$p < 0.0910$	.	$d = 0.6393$	med	$\delta = 0.3689$	med
S.02	$p < 0.5492$	–	$p < 0.5350$	–	$p < 0.3097$	–	$d = 0.3794$	small	$\delta = 0.1289$	neg
S.03	$p < 0.0656$	.	$p < 0.0626$	.	$p < 0.0529$	.	$d = 0.7382$	med	$\delta = 0.3911$	med
S.04	$p < 0.4859$	–	$p < 0.4727$	–	$p < 0.4223$	–	$d = 0.2974$	small	$\delta = 0.1511$	small
S.05	$p < 0.1245$	–	$p < 0.1194$	–	$p < 0.0761$	.	$d = 0.6748$	med	$\delta = 0.3289$	small
S.06	$p < 0.0895$	.	$p < 0.0856$	.	$p < 0.0852$	.	$d = 0.6523$	med	$\delta = 0.3600$	med
S.07	$p < 0.2194$	–	$p < 0.2112$	–	$p < 0.2015$	–	$d = 0.4795$	small	$\delta = 0.2533$	small
S.08	$p < 0.0318$	*	$p < 0.0302$	*	$p < 0.0379$	*	$d = 0.7958$	med	$\delta = 0.4578$	med
S.09	$p < 0.2302$	–	$p < 0.2222$	–	$p < 0.2474$	–	$d = 0.4313$	small	$\delta = 0.2578$	small
S.10	$p < 0.4174$	–	$p < 0.4049$	–	$p < 0.2694$	–	$d = 0.4131$	small	$\delta = 0.1689$	small
S.11	$p < 0.3176$	–	$p < 0.3071$	–	$p < 0.1898$	–	$d = 0.4930$	small	$\delta = 0.2089$	small
S.12	$p < 0.1392$	–	$p < 0.1337$	–	$p < 0.1550$	–	$d = 0.5341$	med	$\delta = 0.3156$	small
S.13	$p < 0.3817$	–	$p < 0.3696$	–	$p < 0.1644$	–	$d = 0.5278$	med	$\delta = 0.1778$	small
S.14	$p < 0.1460$	–	$p < 0.1402$	–	$p < 0.1391$	–	$d = 0.5567$	med	$\delta = 0.3067$	small
S.15	$p < 0.0348$	*	$p < 0.0330$	*	$p < 0.0397$	*	$d = 0.7884$	med	$\delta = 0.4489$	med

Table A.29: *Significance test and effect size between Likert results from subjects with the role of Information Giver (IG) and Information Follower (IF)*

Statement	wilcox.test	sig.	kruskai.test	sig.	t.test	sig.	cohen.d	ES	cliff.delta	ES
S.01	$p < 0.2000$	–	$p < 0.1927$	–	$p < 0.1935$	–	$d = 0.4868$	small	$\delta = 0.2756$	small
S.02	$p < 0.3465$	–	$p < 0.3357$	–	$p < 0.6130$	–	$d = 0.1870$	neg	$\delta = 0.2000$	small
S.03	$p < 0.1276$	–	$p < 0.1224$	–	$p < 0.1262$	–	$d = 0.5757$	med	$\delta = 0.3244$	small
S.04	$p < 0.4223$	–	$p < 0.4102$	–	$p < 0.4223$	–	$d = 0.2974$	small	$\delta = 0.1733$	small
S.05	$p < 0.1051$	–	$p < 0.1007$	–	$p < 0.1135$	–	$d = 0.5965$	med	$\delta = 0.3467$	med
S.06	$p < 0.4448$	–	$p < 0.4322$	–	$p < 0.3716$	–	$d = 0.3320$	small	$\delta = 0.1644$	small
S.07	$p < 0.0044$	**	$p < 0.0041$	**	$p < 0.0174$	**	$d = 0.9295$	large	$\delta = 0.5822$	large
S.08	$p < 0.4743$	–	$p < 0.4614$	–	$p < 0.3888$	–	$d = 0.3196$	small	$\delta = 0.1556$	small
S.09	$p < 0.7048$	–	$p < 0.6892$	–	$p < 0.6027$	–	$d = 0.1923$	neg	$\delta = 0.0844$	neg
S.10	$p < 0.6142$	–	$p < 0.5989$	–	$p < 0.2717$	.	$d = 0.4131$	small	$\delta = 0.1067$	neg
S.11	$p < 0.5286$	–	$p < 0.5145$	–	$p < 0.8855$	–	$d = 0.0531$	neg	$\delta = 0.1333$	neg
S.12	$p < 0.4345$	–	$p < 0.4222$	–	$p < 0.4495$	–	$d = 0.2801$	small	$\delta = 0.1689$	small
S.13	$p < 0.3578$	–	$p < 0.3462$	–	$p < 0.7821$	–	$d = 0.1020$	neg	$\delta = 0.1867$	small
S.14	$p < 0.0381$	*	$p < 0.0362$	*	$p < 0.0270$	*	$d = 0.8549$	large	$\delta = 0.4356$	med
S.15	$p < 0.3009$	–	$p < 0.2912$	–	$p < 0.2457$	–	$d = 0.4331$	small	$\delta = 0.2222$	small

Table A.30: *Significance test and effect size between Likert results from English speaking and Portuguese speaking subjects*

Statement	wilcox.test	sig.	kruskai.test	sig.	t.test	sig.	cohen.d	ES	cliff.delta	ES
S.01	$p < 0.9329$	–	$p < 0.9162$	–	$p < 0.9801$	–	$d = 0.0090$	neg	$\delta = 0.0223$	neg
S.02	$p < 0.8807$	–	$p < 0.8638$	–	$p < 0.6923$	–	$d = 0.1437$	neg	$\delta = 0.0357$	neg
S.03	$p < 0.9155$	–	$p < 0.8987$	–	$p < 0.8997$	–	$d = 0.0462$	neg	$\delta = 0.0268$	neg
S.04	$p < 0.8489$	–	$p < 0.8324$	–	$p < 0.7974$	–	$d = 0.0919$	neg	$\delta = 0.0446$	neg
S.05	$p < 0.5834$	–	$p < 0.5690$	–	$p < 0.6561$	–	$d = 0.1622$	neg	$\delta = 0.1205$	neg
S.06	$p < 0.5096$	–	$p < 0.4960$	–	$p < 0.3805$	–	$d = 0.3208$	small	$\delta = 0.1429$	neg
S.07	$p < 0.2623$	–	$p < 0.2530$	–	$p < 0.3011$	–	$d = 0.3943$	small	$\delta = 0.2321$	small
S.08	$p < 0.6886$	–	$p < 0.6731$	–	$p < 0.5988$	–	$d = 0.1934$	neg	$\delta = 0.0893$	neg
S.09	$p < 0.6275$	–	$p < 0.6127$	–	$p < 0.6925$	–	$d = 0.1439$	neg	$\delta = 0.1071$	neg
S.10	$p < 0.3796$	–	$p < 0.3678$	–	$p < 0.8121$	.	$d = 0.0851$	neg	$\delta = 0.1830$	small
S.11	$p < 0.7771$	–	$p < 0.7605$	–	$p < 0.8463$	–	$d = 0.0711$	neg	$\delta = 0.0625$	neg
S.12	$p < 0.2275$	–	$p < 0.2195$	–	$p < 0.1607$	–	$d = 0.5159$	med	$\delta = 0.2589$	small
S.13	$p < 0.5589$	–	$p < 0.5439$	–	$p < 0.1422$	–	$d = 0.5303$	med	$\delta = 0.1205$	neg
S.14	$p < 0.2938$	–	$p < 0.2840$	–	$p < 0.2568$	–	$d = 0.4231$	small	$\delta = 0.2232$	small
S.15	$p < 0.7510$	–	$p < 0.7350$	–	$p < 0.8650$	–	$d = 0.0618$	neg	$\delta = 0.0714$	neg

Table A.31: *Significance test and effect size between Likert results from male and female subjects*

Statement	wilcox.test	sig.	kruskai.test	sig.	t.test	sig.	cohen.d	ES	cliff.delta	ES
S.01	$p < 0.5555$	–	$p < 0.5414$	–	$p < 0.6064$	–	$d = 0.1952$	neg	$\delta = 0.1295$	neg
S.02	$p < 0.5483$	–	$p < 0.5341$	–	$p < 0.5162$	–	$d = 0.2386$	small	$\delta = 0.1295$	neg
S.03	$p < 0.7991$	–	$p < 0.7828$	–	$p < 0.9510$	–	$d = 0.0231$	neg	$\delta = 0.0580$	neg
S.04	$p < 0.2899$	–	$p < 0.2804$	–	$p < 0.3282$	–	$d = 0.3742$	small	$\delta = 0.2277$	small
S.05	$p < 0.7839$	–	$p < 0.7678$	–	$p < 0.6677$	–	$d = 0.1622$	neg	$\delta = 0.0625$	neg
S.06	$p < 0.7336$	–	$p < 0.7176$	–	$p < 0.6004$	–	$d = 0.1957$	neg	$\delta = 0.0759$	neg
S.07	$p < 0.3675$	–	$p < 0.3559$	–	$p < 0.1963$	–	$d = 0.4681$	small	$\delta = 0.1875$	small
S.08	$p < 0.8165$	–	$p < 0.8002$	–	$p < 0.8088$	–	$d = 0.0893$	neg	$\delta = 0.0536$	neg
S.09	$p < 0.2727$	–	$p < 0.2636$	–	$p < 0.2821$	–	$d = 0.4031$	small	$\delta = 0.2366$	small
S.10	$p < 0.9825$	–	$p < 0.9650$	–	$p < 0.6821$	–	$d = 0.1473$	neg	$\delta = 0.0089$	neg
S.11	$p < 0.5566$	–	$p < 0.5420$	–	$p < 0.3182$	–	$d = 0.3616$	small	$\delta = 0.1250$	neg
S.12	$p < 0.4717$	–	$p < 0.4587$	–	$p < 0.5470$	–	$d = 0.2188$	small	$\delta = 0.1563$	small
S.13	$p < 0.3686$	–	$p < 0.3567$	–	$p < 0.4070$	–	$d = 0.3108$	small	$\delta = 0.1830$	small
S.14	$p < 0.7642$	–	$p < 0.7479$	–	$p < 0.6484$	–	$d = 0.1659$	neg	$\delta = 0.0670$	neg
S.15	$p < 0.4984$	–	$p < 0.4850$	–	$p < 0.5037$	–	$d = 0.2490$	small	$\delta = 0.1473$	small

Table A.32: *Significance test and effect size between Likert results from subjects using map 01 and subjects using map 07*

Appendix B

The HCRC Edinburgh Map Task corpus

B.1 The Four Route Types of the HCRC Edinburgh Map Task corpus Maps

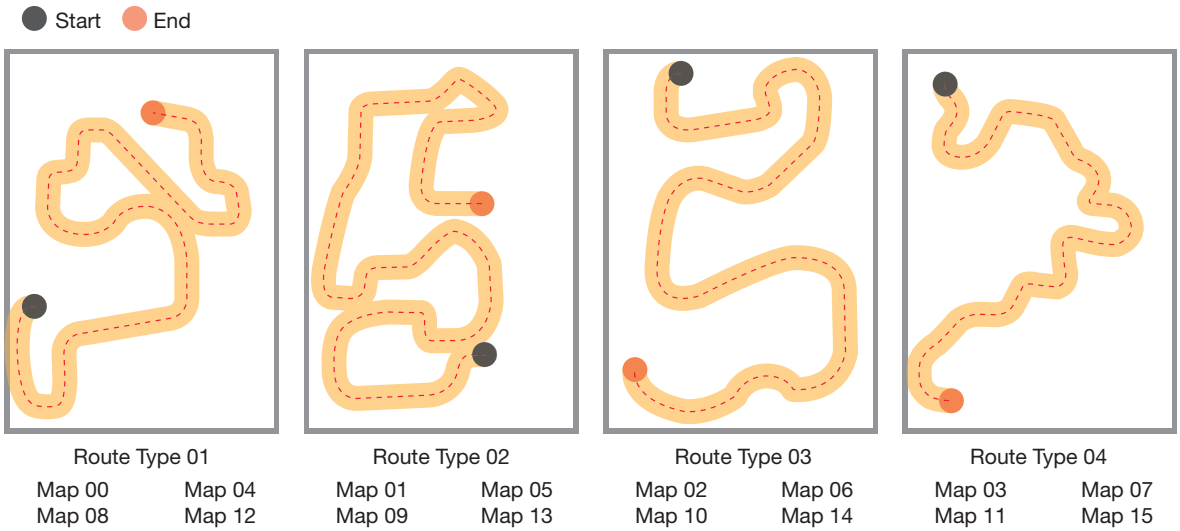


Figure B.1: *The 4 route patterns of the HCRC Edinburgh Map Task corpus maps*

B.2 The Map Differences Between Information Giver and the Information Follower

Route Type 01								
Items	Map 00		Map 04		Map 08		Map 12	
Role	IG	IF	IG	IF	IG	IF	IG	IF
START	✓	✓	✓	✓	✓	✓	✓	✓
Item 1	Common	Common	Common	Common	Common	Common	Common	Common
Item 2	**2 : 1**	—	**Abs/Pre**	—	**Abs/Pre**	—	—	**Abs/Pre**
Item 3	Common	Common	Common	Common	**Name**	**Name**	**Name**	**Name**
Item 4	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**
Item 5	Common	Common	Common	Common	Common	Common	Common	Common
Item 6	—	**Contrast**	**2 : 1**	—	**Contrast**	—	**2 : 1**	—
Item 7	Common	Common	Common	Common	Common	Common	**Contrast**	**Contrast**
Item 8	**Abs/Pre**	—	—	**Abs/Pre**	—	**Abs/Pre**	**Abs/Pre**	—
Item 9	Common	Common	Common	Common	**Name**	**Name**	Common	Common
Item 10	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	**Abs/Pre**	—
Item 11	**Abs/Pre**	—	**Abs/Pre**	—	**2 : 1**	—	—	**Abs/Pre**
Item 12	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**
FINISH	✓	—	✓	—	✓	—	✓	—
Item Other	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**
Item Other	Common	Common	Common	Common	Common	Common	Common	Common
Item Other	Common	Common	Common	Common	—	—	—	—
Abs/Pre	2	2	2	3	1	3	2	3
Name	0	0	0	0	2	2	1	1
2 : 1	2	1	2	1	2	1	2	1
Contrast	1	2	1	1	2	1	2	2
Locations	12	12	12	12	12	12	12	12
Landmarks	9	9	9	9	9	9	9	9
Same	6	6	6	6	4	4	5	5
Different	3	3	3	3	5	5	4	4
Empty	3	3	3	3	3	3	3	3

Abs/Pre: *Absent/Present* landmarks were found on one map but not the other

Name: *Name Change* landmarks were identical in form and location but had different labels on the two maps

2 : 1: *2 : 1* landmarks appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant irrelevant one

Contrast: *Contrast Feature* a pair of landmarks with similar names which might elicit contrastive pronunciations (Green bay and Crane Bay)

Common: *Common* landmarks that have the same name, location and image on both maps

Table B.1: *HCRC Edinburgh Map Task corpus Map difference between role — Route type 01*

Route Type 02								
Items	Map 01		Map 05		Map 09		Map 13	
Role	IG	IF	IG	IF	IG	IF	IG	IF
START	✓	✓	✓	✓	✓	✓	✓	✓
Item 1	Common	Common	Common	Common	Common	Common	Common	Common
Item 2	—	**Abs/Pre**	**Abs/Pre**	**Abs/Pre**	**2 : 1**	—	**Abs/Pre**	—
Item 3	Common	Common	Common	Common	Common	Common	Common	Common
Item 4	**2 : 1**	—	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**
Item 5	Common	Common	Common	Common	Common	Common	Common	Common
Item 6	—	**Contrast**	**2 : 1**	—	**Contrast**	—	—	**Abs/Pre**
Item 7	Common	Common	Common	Common	**Name**	**Name**	Common	Common
Item 8	**Abs/Pre**	—	—	**Abs/Pre**	—	**Abs/Pre**	**Abs/Pre**	—
Item 9	**Contrast**	**Contrast**	**Name**	**Name**	**Contrast**	**Contrast**	**Contrast**	**Contrast**
Item 10	—	**Abs/Pre**	**Abs/Pre**	—	—	**Abs/Pre**	**2 : 1**	—
Item 11	Common	Common	**Contrast**	**Contrast**	Common	Common	**Contrast**	**Contrast**
Item 12	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	—	—	**Abs/Pre**
Item 13	—	—	—	**Abs/Pre**	—	—	—	—
Item 14	Common	Common	Common	Common	Common	Common	Common	Common
FINISH	✓	—	✓	—	✓	—	✓	—
Item Other	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**
Abs/Pre	2	2	3	4	1	3	2	3
Name	0	0	1	1	1	1	0	0
2 : 1	2	1	2	1	2	1	2	1
Contrast	1	2	1	1	2	1	2	2
Locations	13	13	14	14	13	13	13	13
Landmarks	10	10	11	11	10	10	10	10
Same	7	7	6	6	6	6	7	7
Different	3	3	5	5	4	4	3	3
Empty	3	3	3	3	3	3	3	3

Abs/Pre: *Absent/Present* landmarks were found on one map but not the other

Name: *Name Change* landmarks were identical in form and location but had different labels on the two maps

2 : 1: *2 : 1* landmarks appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant irrelevant one

Contrast: *Contrast Feature* a pair of landmarks with similar names which might elicit contrastive pronunciations (Green bay and Crane Bay)

Common: *Common* landmarks that have the same name, location and image on both maps

Table B.2: *HCRC Edinburgh Map Task corpus Map difference between role — Route type 02*

Route Type 03								
Items	Map 02		Map 06		Map 10		Map 14	
Role	IG	IF	IG	IF	IG	IF	IG	IF
START	✓	✓	✓	✓	✓	✓	✓	✓
Item 1	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**
Item 2	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	–
Item 3	Common	Common	**Name**	**Name**	Common	Common	Common	Common
Item 4	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**
Item 5	Common	Common	Common	Common	Common	Common	Common	Common
Item 6	**2 : 1**	—	**2 : 1**	—	**2 : 1**	—	**2 : 1**	—
Item 7	**Name**	**Name**	Common	Common	Common	Common	**Contrast**	**Contrast**
Item 8	—	**Abs/Pre**	—	**Abs/Pre**	**Contrast**	—	—	**Abs/Pre**
Item 9	Common	Common	Common	Common	Common	Common	Common	Common
Item 10	—	**Contrast**	—	**Abs/Pre**	—	**Abs/Pre**	**Abs/Pre**	—
Item 11	Common	Common	Common	Common	Common	Common	Common	Common
Item 12	**Abs/Pre**	—	**Abs/Pre**	—	—	**Abs/Pre**	—	**Abs/Pre**
Item 13	Common	Common	Common	Common	Common	Common	Common	Common
FINISH	✓	—	✓	—	✓	—	✓	—
Item Other	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**
Item Other	Common	Common	Common	Common	Common	Common	—	—
Item Other	Common	Common	Common	Common	—	—	—	—
Abs/Pre	2	2	2	3	1	3	2	3
Name	1	1	1	1	0	0	0	0
2 : 1	2	1	2	1	2	1	2	1
Contrast	1	2	1	1	2	1	2	2
Locations	13	13	13	13	13	13	13	13
Landmarks	10	10	10	10	10	10	10	10
Same	6	6	6	6	7	7	7	7
Different	4	4	4	4	3	3	3	3
Empty	3	3	3	3	3	3	3	3

Abs/Pre: *Absent/Present* landmarks were found on one map but not the other

Name: *Name Change* landmarks were identical in form and location but had different labels on the two maps

2 : 1: *2 : 1* landmarks appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant irrelevant one

Contrast: *Contrast Feature* a pair of landmarks with similar names which might elicit contrastive pronunciations (Green bay and Crane Bay)

Common: *Common* landmarks that have the same name, location and image on both maps

Table B.3: *HCRC Edinburgh Map Task corpus Map difference between role — Route type 03*

Route Type 04								
Items	Map 03		Map 07		Map 11		Map 15	
Role	IG	IF	IG	IF	IG	IF	IG	IF
START	✓	✓	✓	✓	✓	✓	✓	✓
Item 1	**Name**	**Name**	Common	Common	Common	Common	Common	Common
Item 2	**Abs/Pre**	—	—	**Abs/Pre**	—	**Abs/Pre**	**Abs/Pre**	—
Item 3	Common	Common	Common	Common	Common	Common	Common	Common
Item 4	—	**Contrast**	**Abs/Pre**	—	**Contrast**	—	**Contrast**	**Contrast**
Item 5	—	—	—	—	—	—	—	**Abs/Pre**
Item 6	Common	Common	Common	Common	Common	Common	**Name**	**Name**
Item 7	—	**Abs/Pre**	—	**Abs/Pre**	**Abs/Pre**	—	**Name**	**Name**
Item 8	Common	Common	Common	Common	Common	Common	**2 : 1**	—
Item 9	**Abs/Pre**	—	**2 : 1**	—	—	**Abs/Pre**	Common	Common
Item 10	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	**Contrast**	—	**Abs/Pre**
Item 11	**2 : 1**	—	**Abs/Pre**	—	**2 : 1**	—	Common	Common
Item 12	Common	Common	Common	Common	Common	Common	**Abs/Pre**	—
Item 13	—	—	—	—	—	—	**Contrast**	**Contrast**
Item 14	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**	—	**Abs/Pre**
Item 15	Common	Common	Common	Common	Common	Common	Common	Common
FINISH	✓	—	✓	—	✓	—	✓	—
Item Other	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**	**2 : 1**
Item Other	—	—	Common	Common	Common	Common	Common	Common
Item Other	—	—	—	—	—	—	Common	Common
Item Other	—	—	—	—	—	—	Common	—
Abs/Pre	2	2	2	3	1	3	2	3
Name	1	1	0	0	0	0	2	2
2 : 1	2	1	2	1	2	1	2	1
Contrast	1	2	1	1	2	1	2	2
Locations	13	13	13	13	13	13	15	15
Landmarks	10	10	10	10	10	10	12	12
Same	6	6	7	7	7	7	7	7
Different	4	4	3	3	3	3	5	5
Empty	3	3	3	3	3	3	3	3

Abs/Pre: *Absent/Present* landmarks were found on one map but not the other

Name: *Name Change* landmarks were identical in form and location but had different labels on the two maps

2 : 1: *2 : 1* landmarks appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant irrelevant one

Contrast: *Contrast Feature* a pair of landmarks with similar names which might elicit contrastive pronunciations (Green bay and Crane Bay)

Common: *Common* landmarks that have the same name, location and image on both maps

Table B.4: *HCRC Edinburgh Map Task corpus Map difference between role — Route type 04*

B.3 Speech Rate of 16 HCRC Edinburgh Map Task corpus Dialogues

B.3.1 HCRC Edinburgh Map Task corpus Dialogues After 20 Window Normalization

The following tables contain the resulting 20 sample window values of the HCRC Edinburgh Map Task corpus dialogues that were analysed in this thesis.

Table B.5: 20 Window Values of Groupings within the HCRC Edinburgh Map Task corpus.

Table B.6: 20 Window Values from four dialogues of that used map 01.

Table B.7: 20 Window Values from four dialogues of that used map 01.

Table B.8: 20 Window Values from four dialogues of that used map 07.

Table B.9: 20 Window Values from four dialogues of that used map 07.

Grouping	Role		Gender		Setting		Map	
Specific	IG	IF	♀	♂	w/ EC	w/o EC	Map 01	Map 07
Dialog. Acts	2,203	1,587	2,006	1,784	1,787	2,003	1,543	2,247
Windows	20	20	20	20	20	20	20	20
w.01 (%)	−7.538	−0.170	−7.612	−0.096	−4.756	−2.952	−12.899	5.191
w.02 (%)	−0.226	9.328	4.982	4.119	1.396	7.706	−5.404	14.506
w.03 (%)	0.141	7.688	−0.769	8.597	7.398	0.431	−0.212	8.041
w.04 (%)	6.935	10.531	4.755	12.711	9.919	7.547	10.326	7.140
w.05 (%)	3.486	7.276	7.308	3.454	1.672	9.090	4.457	6.305
w.06 (%)	−2.428	6.009	0.131	3.450	−3.686	7.266	−2.961	6.541
w.07 (%)	1.624	10.889	8.312	4.201	3.580	8.933	2.430	10.084
w.08 (%)	6.396	16.192	5.261	17.327	11.559	11.029	15.550	7.037
w.09 (%)	5.907	5.905	4.184	7.629	8.337	3.475	5.201	6.611
w.10 (%)	3.077	5.059	6.431	1.705	1.976	6.160	−0.457	8.593
w.11 (%)	11.308	12.088	17.167	6.229	6.715	16.680	10.568	12.827
w.12 (%)	−4.678	13.407	13.665	−4.936	0.875	7.854	−1.685	10.415
w.13 (%)	−0.828	18.200	4.953	12.418	9.468	7.903	8.110	9.262
w.14 (%)	−6.883	14.888	5.645	2.360	1.353	6.652	0.051	7.954
w.15 (%)	−2.807	14.806	9.925	2.074	−1.553	13.552	4.212	7.787
w.16 (%)	2.740	14.749	10.014	7.475	6.443	11.046	9.134	8.355
w.17 (%)	−7.792	8.123	2.476	−2.146	3.882	−3.551	−3.476	3.807
w.18 (%)	−0.682	−1.407	10.480	−12.569	−11.843	9.754	−5.781	3.693
w.19 (%)	1.119	12.747	5.575	8.291	6.718	7.149	7.621	6.246
w.20 (%)	8.075	20.544	20.885	7.735	7.540	21.079	9.845	18.775
Mdn	4.33	15.59	9.01	10.07	8.19	10.38	7.36	11.01
M	0.85	10.34	6.69	4.50	3.35	7.84	2.73	8.46
SD	23.2	27.1	19.5	30.6	29.7	20.6	30.0	20.0

Table B.5: *20 Window Values of Groupings within the HCRC Edinburgh Map Task corpus*

Map *No.*	Map 01							
Eye Contact	w/ EC				w/o EC			
HCRC ID	q3ec4		q3ec6		q3nc4		q3nc6	
Dialog. *No.*	D.01		D.02		D.03		D.04	
Familiarity	Yes		No		Yes		No	
Dev. Score	154		89		45		99	
Role	IG	IF	IG	IF	IG	IF	IG	IF
Gender	♂	♂	♂	♂	♂	♀	♂	♀
Dialog. Acts	88	66	44	36	96	69	55	46
Duplication ×	5	10	5	5	5	20	4	10
Sub-Total	440	660	220	180	480	1,380	220	460
Groupings ÷	22	33	11	9	24	69	11	23
Windows	20	20	20	20	20	20	20	20
w.01 (%)	−10.378	5.965	−6.936	−114.690	−14.765	6.176	40.801	34.115
w.02 (%)	5.269	−17.327	−33.978	14.114	4.285	15.939	19.219	23.608
w.03 (%)	−64.209	35.702	−8.416	13.876	33.824	11.730	21.556	−6.860
w.04 (%)	4.708	16.541	13.266	−9.221	33.860	31.904	21.830	31.496
w.05 (%)	−25.849	32.513	−15.404	−37.827	−8.133	36.755	50.262	18.584
w.06 (%)	−40.428	29.107	−64.771	−46.950	27.329	10.582	30.676	−5.726
w.07 (%)	25.869	18.698	−60.152	−10.450	−22.399	13.883	24.627	−0.282
w.08 (%)	5.908	29.220	6.341	12.731	40.952	10.851	29.982	27.194
w.09 (%)	−3.331	11.282	−9.582	−10.590	16.551	11.527	5.398	−1.677
w.10 (%)	17.705	35.605	−10.924	−95.949	11.829	−1.137	15.458	17.644
w.11 (%)	13.627	41.956	−7.413	−55.640	7.854	14.487	11.317	34.553
w.12 (%)	6.896	27.263	−57.153	0.623	−36.511	32.386	−21.022	21.270
w.13 (%)	−10.199	33.469	−35.477	33.758	19.359	28.311	32.532	1.483
w.14 (%)	2.655	41.369	−65.944	0.455	−7.942	23.799	14.144	27.642
w.15 (%)	−11.654	31.449	−115.492	−18.720	19.869	22.788	33.484	51.142
w.16 (%)	42.183	48.912	−55.283	23.383	14.411	16.892	−28.360	37.297
w.17 (%)	15.037	30.901	−30.864	15.790	−11.832	10.408	−49.298	9.160
w.18 (%)	23.958	39.053	−38.554	−232.469	−0.587	15.799	22.051	42.537
w.19 (%)	8.895	42.136	−24.773	−50.324	6.432	25.481	29.022	23.412
w.20 (%)	−5.029	49.593	1.186	−66.127	12.031	27.897	24.688	40.809
Mdn	4.99	31.98	−27.82	−10.52	9.84	15.87	21.94	23.51
M	0.08	29.17	−31.02	−31.71	7.32	18.32	16.42	21.37
SD	23.8	15.9	31.5	62.0	20.0	9.9	24.0	17.1

Table B.6: *20 Window Values of Dialogues 01 to 04 of the HCRC Edinburgh Map Task corpus*

Map *No.*	Map 01							
Eye Contact	w/ EC				w/o EC			
HCRC ID	q7ec4		q7ec6		q7nc4		q7nc6	
Dialog. *No.*	D.05		D.06		D.07		D.08	
Familiarity	No		Yes		No		Yes	
Dev. Score	57		7		35		26	
Role	IG	IF	IG	IF	IG	IF	IG	IF
Gender	♂	♂	♂	♂	♀	♀	♀	♀
Dialog. Acts	104	69	312	235	133	86	63	41
Duplication ×	5	20	5	4	20	10	20	20
Sub-Total	520	1,380	1,560	940	2,660	860	1,260	820
Groupings ÷	26	69	78	47	133	43	63	41
Windows	20	20	20	20	20	20	20	20
w.01 (%)	−9.090	−4.345	2.393	29.954	−44.030	−2.805	−47.685	−71.061
w.02 (%)	−34.823	−57.014	7.024	20.056	−8.727	17.265	−35.202	−26.177
w.03 (%)	−34.729	8.311	1.525	25.021	2.720	−12.103	−19.897	−11.450
w.04 (%)	−32.075	1.614	6.557	17.693	0.987	21.055	6.482	−1.474
w.05 (%)	16.851	−17.206	1.143	−1.456	−9.499	3.809	3.931	22.835
w.06 (%)	−27.817	−7.148	4.575	29.508	−10.754	1.784	20.314	2.346
w.07 (%)	−13.322	−20.678	4.928	28.083	4.843	19.390	22.947	22.947
w.08 (%)	2.022	14.997	22.569	30.889	−20.921	11.784	6.025	18.260
w.09 (%)	19.762	−7.170	−10.292	27.807	4.348	−4.424	30.112	3.498
w.10 (%)	−43.695	3.588	−8.255	7.623	−10.271	−7.562	26.954	34.081
w.11 (%)	8.262	−38.327	2.533	34.082	17.747	37.842	35.247	10.964
w.12 (%)	−91.004	18.688	7.092	15.907	9.244	24.091	25.937	−10.673
w.13 (%)	−3.662	20.511	−7.456	10.247	−12.146	11.825	−0.704	7.904
w.14 (%)	−22.931	−22.931	−1.177	16.074	−46.110	10.778	0.897	15.495
w.15 (%)	−24.113	6.487	8.117	13.571	5.994	4.611	20.787	19.072
w.16 (%)	−20.419	19.435	9.607	−45.239	9.160	25.177	27.455	21.533
w.17 (%)	−35.414	22.128	0.785	9.103	−13.740	2.600	−13.347	−17.038
w.18 (%)	−33.975	6.043	−7.606	10.333	−12.714	24.759	30.945	17.925
w.19 (%)	−16.324	9.094	−6.143	28.277	−2.352	14.367	−2.388	37.121
w.20 (%)	−37.447	28.175	−4.122	22.742	−7.851	26.831	20.246	23.891
Mdn	−23.52	4.85	1.96	18.87	−8.29	11.80	13.36	9.43
M	−21.70	−0.06	1.69	16.51	−7.20	11.55	7.95	5.00
SD	24.7	21.2	7.8	17.4	16.2	13.2	22.9	24.4

Table B.7: *20 Window Values of Dialogues 05 to 08 of the HCRC Edinburgh Map Task corpus*

Map *No.*	Map 07							
Eye Contact	w/ EC				w/o EC			
HCRC ID	q4ec3		q4ec5		q4nc3		q4nc5	
Dialogue *No.*	D.09		D.10		D.11		D.12	
Familiarity	Yes		No		Yes		No	
Dev. Score	11		26		75		21	
Role	IG	IF	IG	IF	IG	IF	IG	IF
Gender	♀	♀	♀	♂	♀	♀	♀	♂
Dialogue Acts	140	89	101	82	295	253	251	176
Duplication ×	1	20	20	10	4	20	20	5
Sub-Total	140	1,780	2,020	820	1,180	5,060	5,020	880
Groupings ÷	7	89	101	41	59	253	251	44
Windows	20	20	20	20	20	20	20	20
w.01 (%)	−23.411	−1.848	1.755	22.589	2.441	23.585	0.359	25.021
w.02 (%)	−4.331	9.033	−13.279	29.381	−8.760	19.229	9.986	11.498
w.03 (%)	20.367	21.347	−4.787	38.133	8.117	24.762	0.196	0.741
w.04 (%)	12.448	25.299	−0.219	12.915	−14.607	26.643	−12.645	26.509
w.05 (%)	7.919	17.926	10.099	7.762	−23.969	16.907	−11.952	22.318
w.06 (%)	−1.461	17.744	−42.237	15.069	−13.071	−0.668	−1.161	30.211
w.07 (%)	−25.721	14.243	16.084	−17.303	−5.262	22.548	−9.667	12.312
w.08 (%)	5.912	−4.467	1.044	1.860	−3.967	27.398	−19.314	23.848
w.09 (%)	18.905	16.133	14.255	4.013	−6.085	26.713	−6.427	16.009
w.10 (%)	−7.209	14.633	15.284	30.024	−0.264	16.410	4.991	−8.914
w.11 (%)	26.912	6.902	3.487	9.570	4.317	29.146	−18.531	10.350
w.12 (%)	11.969	28.450	13.047	16.987	3.440	25.596	21.292	18.026
w.13 (%)	20.674	-1.614	5.213	18.685	−13.517	29.126	−14.199	19.583
w.14 (%)	6.262	16.981	−3.082	14.889	8.993	15.434	3.387	30.062
w.15 (%)	11.962	−10.020	4.406	28.961	−12.427	23.908	10.413	11.025
w.16 (%)	10.354	−15.540	12.740	−5.482	−25.803	16.887	−14.002	38.013
w.17 (%)	11.361	15.161	5.248	−18.789	0.224	26.143	4.564	−3.534
w.18 (%)	14.372	15.427	13.167	−56.957	0.108	36.506	−29.675	9.886
w.19 (%)	25.309	−3.456	15.215	−5.682	−1.438	35.687	−-9.992	26.787
w.20 (%)	12.225	25.905	21.600	8.450	8.238	27.434	16.775	26.121
Mdn	11.66	14.90	5.23	11.24	−2.70	25.18	−3.79	18.80
M	7.74	10.41	4.45	7.75	−4.67	23.47	−3.78	17.29
SD	14.2	12.5	13.9	21.5	10.1	8.1	13.1	12.0

Table B.8: *20 Window Values of Dialogues 09 to 12 of the HCRC Edinburgh Map Task corpus*

Map *No.*	Map 07							
Eye Contact	w/ EC				w/o EC			
HCRC ID	q8ec3		q8ec5		q8nc3		q8nc5	
Dialog. *No.*	D.13		D.14		D.15		D.16	
Familiarity	No		Yes		No		Yes	
Dev. Score	11		25		54		40	
Role	IG	IF	IG	IF	IG	IF	IG	IF
Gender	♂	♂	♂	♂	♀	♀	♀	♀
Dialog. Acts	112	81	141	87	81	42	187	129
Duplication ×	5	20	20	20	20	10	20	20
Sub-Total	560	1,620	2,820	1,740	1,620	420	3,740	2,580
Groupings ÷	28	81	141	87	81	21	187	129
Windows	20	20	20	20	20	20	20	20
w.01 (%)	−11.936	2.520	11.502	29.853	−4.538	29.878	−7.094	−17.628
w.02 (%)	15.734	43.008	22.635	16.826	52.141	25.630	−0.814	4.176
w.03 (%)	5.892	32.773	31.076	−3.519	4.249	1.937	4.767	−57.397
w.04 (%)	27.262	33.718	28.238	−0.042	2.523	−59.670	12.342	−6.478
w.05 (%)	22.359	11.234	29.877	−33.186	8.985	6.736	−0.852	8.715
w.06 (%)	15.677	35.542	28.894	−4.279	31.661	−6.147	3.722	−4.836
w.07 (%)	15.921	42.992	33.454	4.636	2.854	31.147	10.986	12.112
w.08 (%)	−21.945	36.024	30.167	11.666	2.355	23.687	15.204	−16.876
w.09 (%)	26.491	34.663	30.775	−29.728	−41.729	18.644	5.366	−22.221
w.10 (%)	13.159	33.522	30.432	−3.929	−16.144	15.679	10.186	−10.373
w.11 (%)	6.339	32.129	33.253	−10.229	20.071	13.337	15.899	22.285
w.12 (%)	5.502	42.776	34.591	−67.632	−25.749	30.502	17.582	−9.745
w.13 (%)	0.621	47.093	20.147	-0.522	−24.173	21.514	−9.732	9.825
w.14 (%)	2.046	24.644	13.088	−15.290	−33.689	22.882	19.274	1.378
w.15 (%)	−15.722	27.790	35.661	2.472	−20.796	38.139	4.605	−15.782
w.16 (%)	20.088	45.685	14.603	−1.943	40.342	25.300	−13.241	−14.326
w.17 (%)	29.494	37.295	−22.095	−23.036	8.250	18.707	−23.049	−5.039
w.18 (%)	11.095	21.703	19.140	5.788	−1.923	33.259	−20.710	−12.099
w.19 (%)	17.759	35.602	42.391	−10.488	−83.061	30.339	19.358	−34.396
w.20 (%)	21.580	46.774	−10.150	5.291	35.314	23.641	19.921	11.279
Mdn	14.42	35.10	29.39	−2.73	2.44	23.26	7.55	-8.11
M	10.37	33.37	22.88	−6.37	−2.15	17.26	5.16	−7.87
SD	14.3	11.5	15.7	20.7	31.5	21.1	12.9	18.0

Table B.9: *20 Window Values of Dialogues 13 to 16 of the HCRC Edinburgh Map Task corpus*

B.3.2 Line Charts of HCRC Edinburgh Map Task corpus Dialogues' Speech Rate

The following pages contain the representation[1] of the dialogues in the 20 sample window framework.

Figure B.2: Subject speech rate means by role.

Figure B.3: Subject speech rate means by gender, eye contact setting, map, and familiarity.

Figure B.4: Subject speech rate means of roles divided by gender and eye contact setting.

Figure B.5: Subject speech rate means of roles divided by map and familiarity.

Figure B.6: Subject speech rate means of gender divided by role and eye contact setting.

Figure B.7: Subject speech rate means of gender divided by map used and familiarity.

Figure B.8: Subject speech rate means of eye contact divided by role and gender.

Figure B.9: Subject speech rate means of eye contact divided by map used and familiarity.

Figure B.10: Subject speech rate means of map divided by role and gender.

Figure B.11: Subject speech rate means of map divided by eye contact and familiarity.

Figure B.12: Subject speech rate means of familiarity divided by role and gender.

Figure B.13: Subject speech rate means of familiarity divided by eye contact and map used.

Figure B.14: Individual dialogue figures — Dialogues that used map 01.

Figure B.15: Individual dialogue figures — Subjects that used map 07.

Figure B.16: Individual dialogue figures — Subjects that used map 01.

Figure B.17: Individual dialogue figures — Subjects that used map 07.

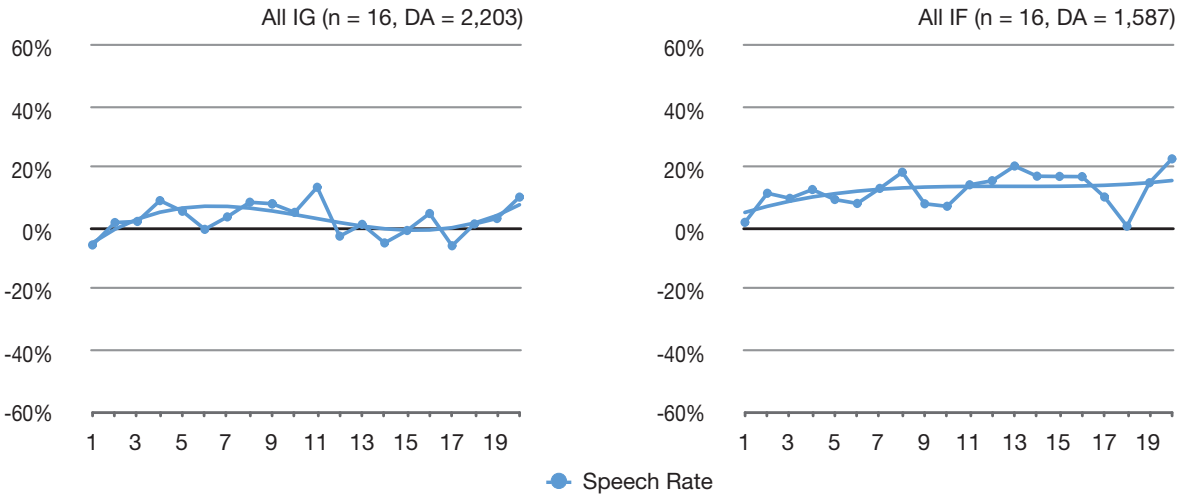


Figure B.2: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Grouping 01*

[1]For the individual dialogues of Figure B.14, to Figure B.17, the black text denotes male subjects and red text denotes female speakers.

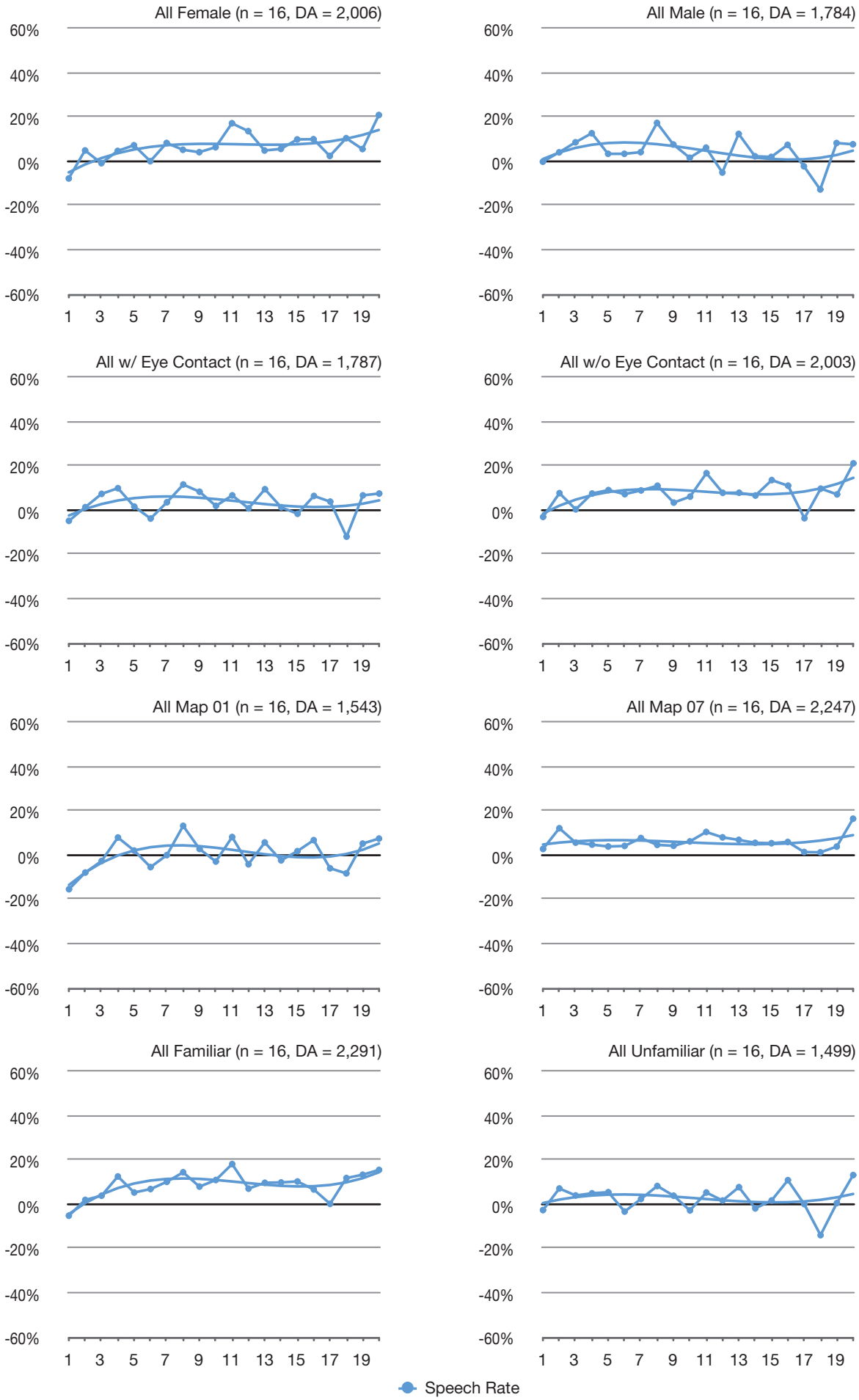


Figure B.3: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Grouping 02*

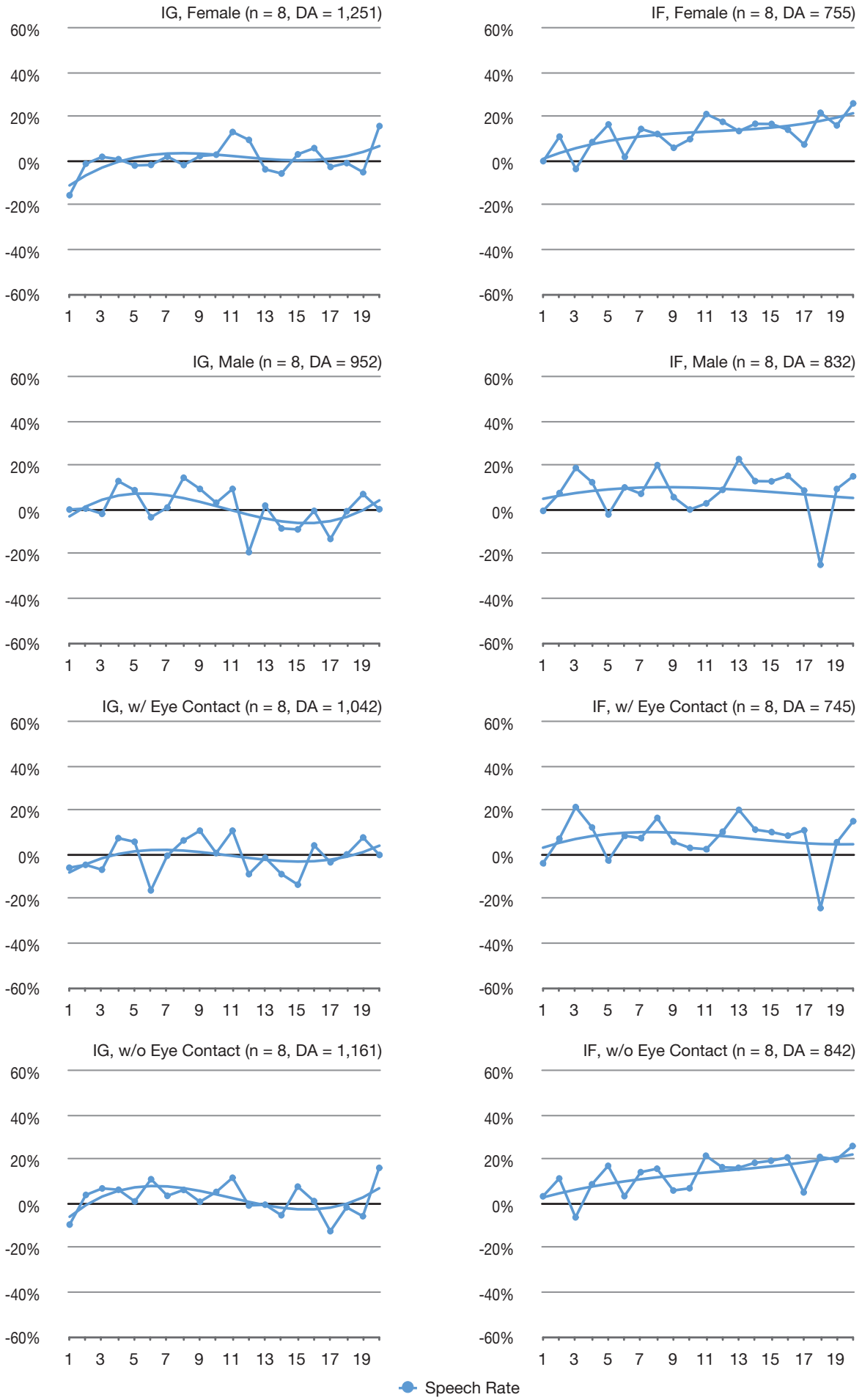


Figure B.4: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Role 01*

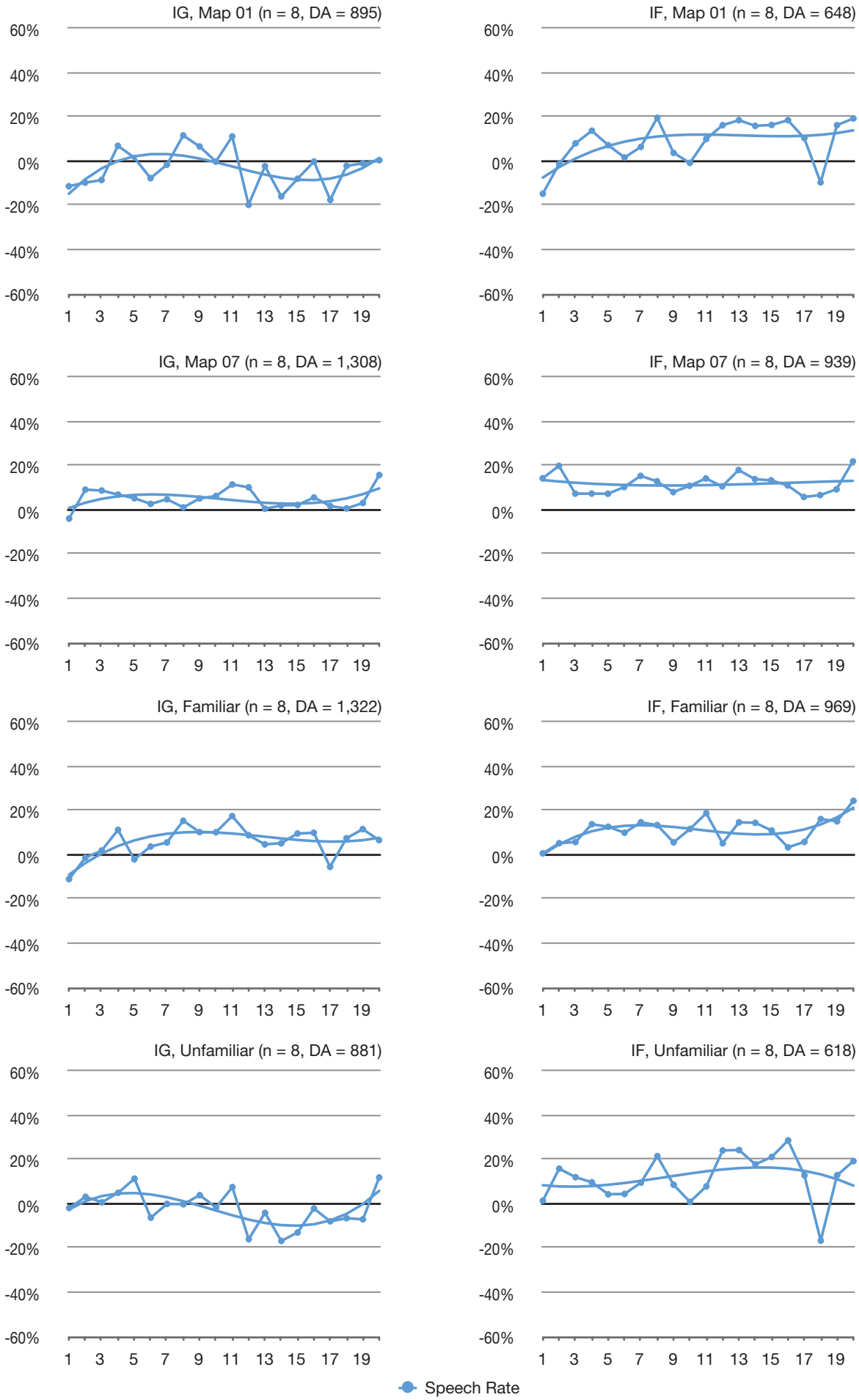


Figure B.5: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Role 02*

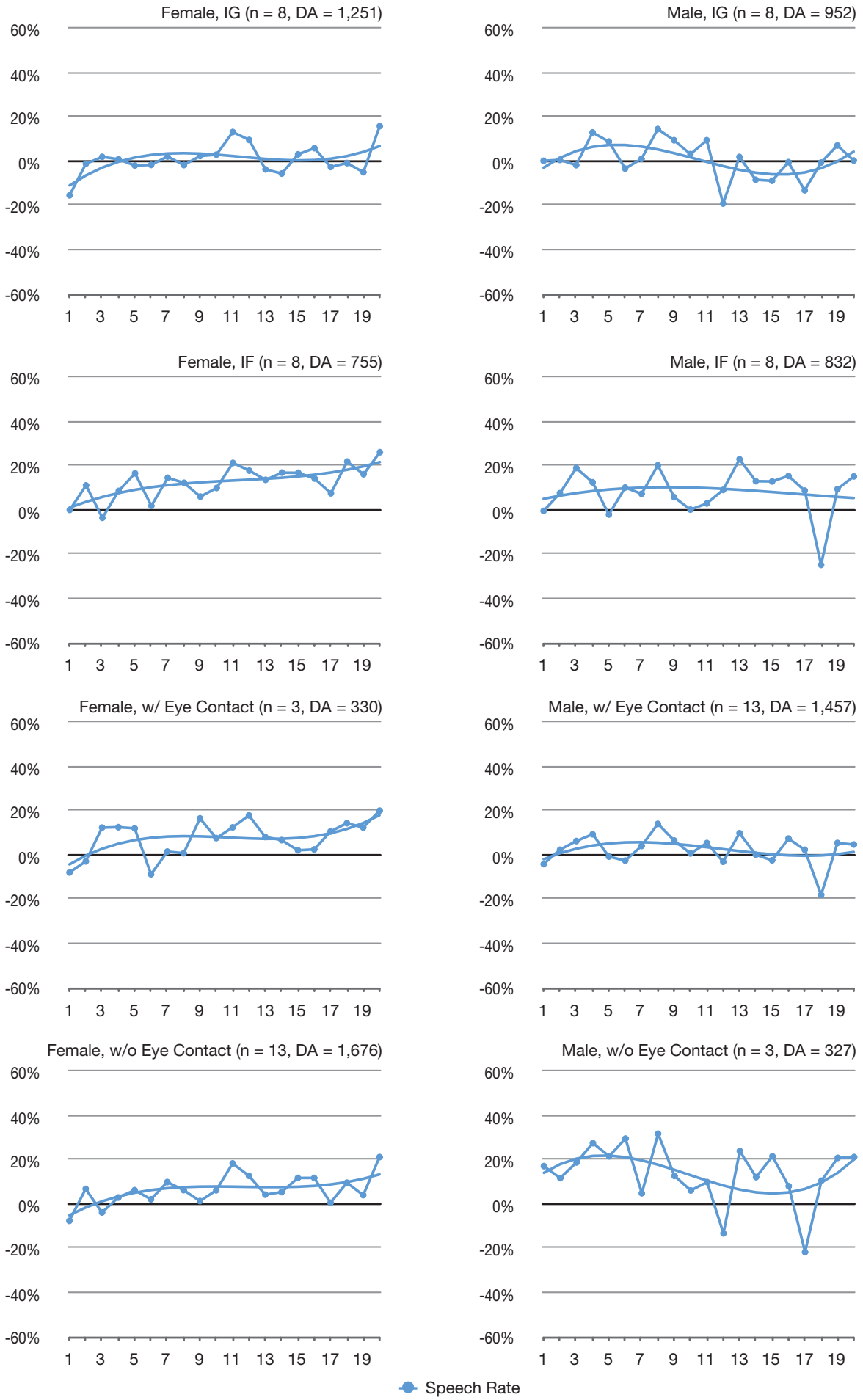


Figure B.6: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Gender 01*

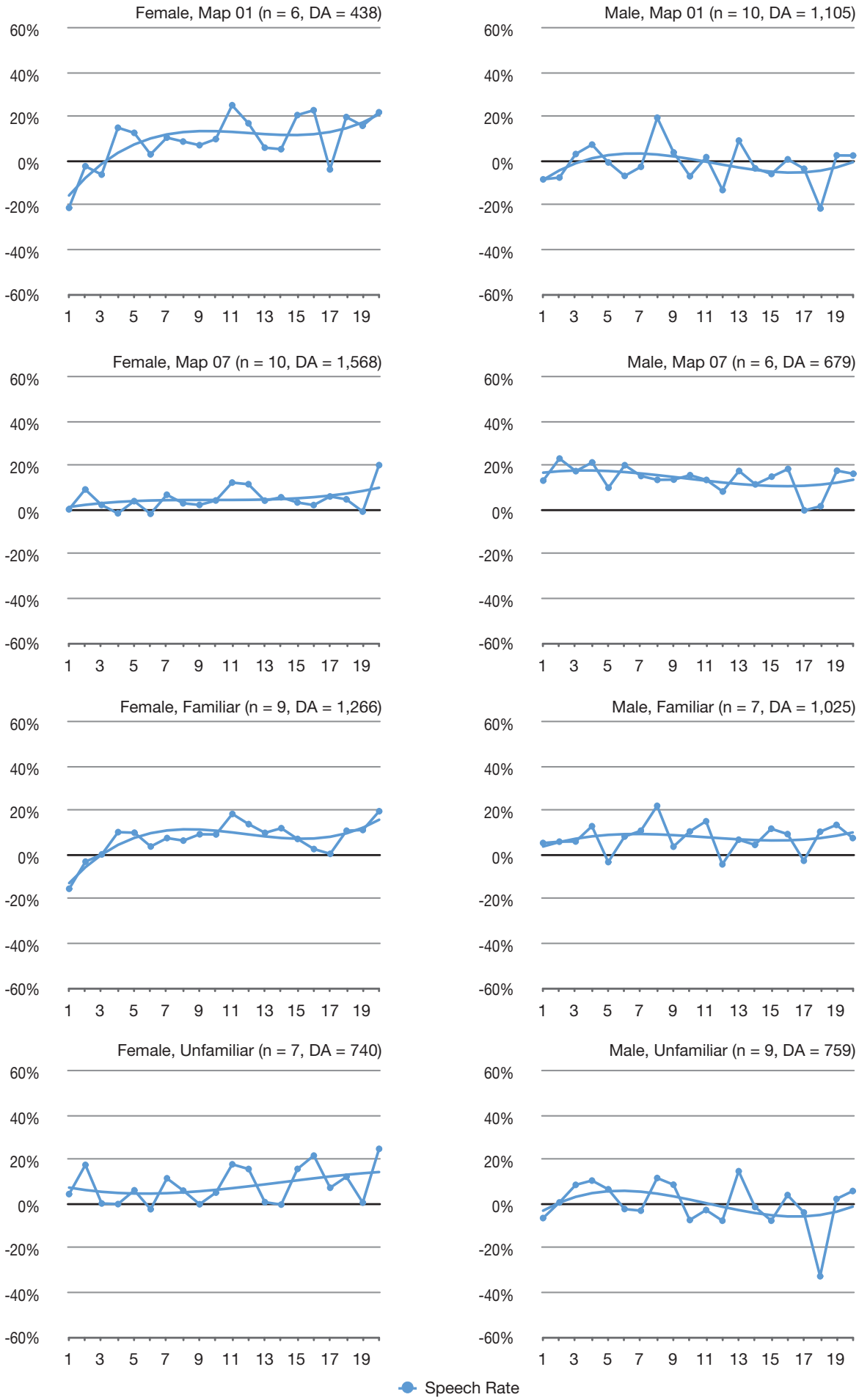


Figure B.7: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Gender 02*

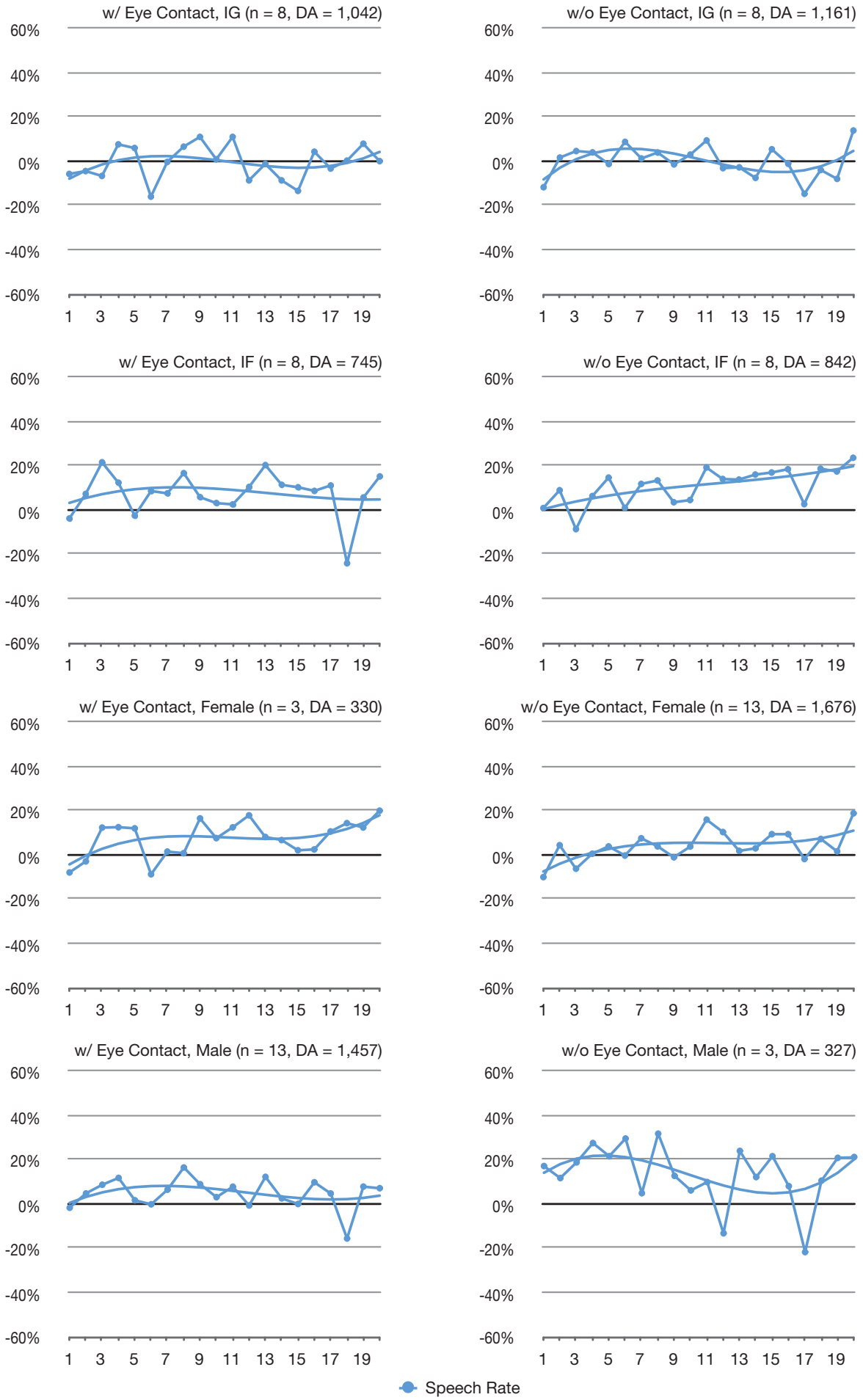


Figure B.8: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Eye Contact 01*

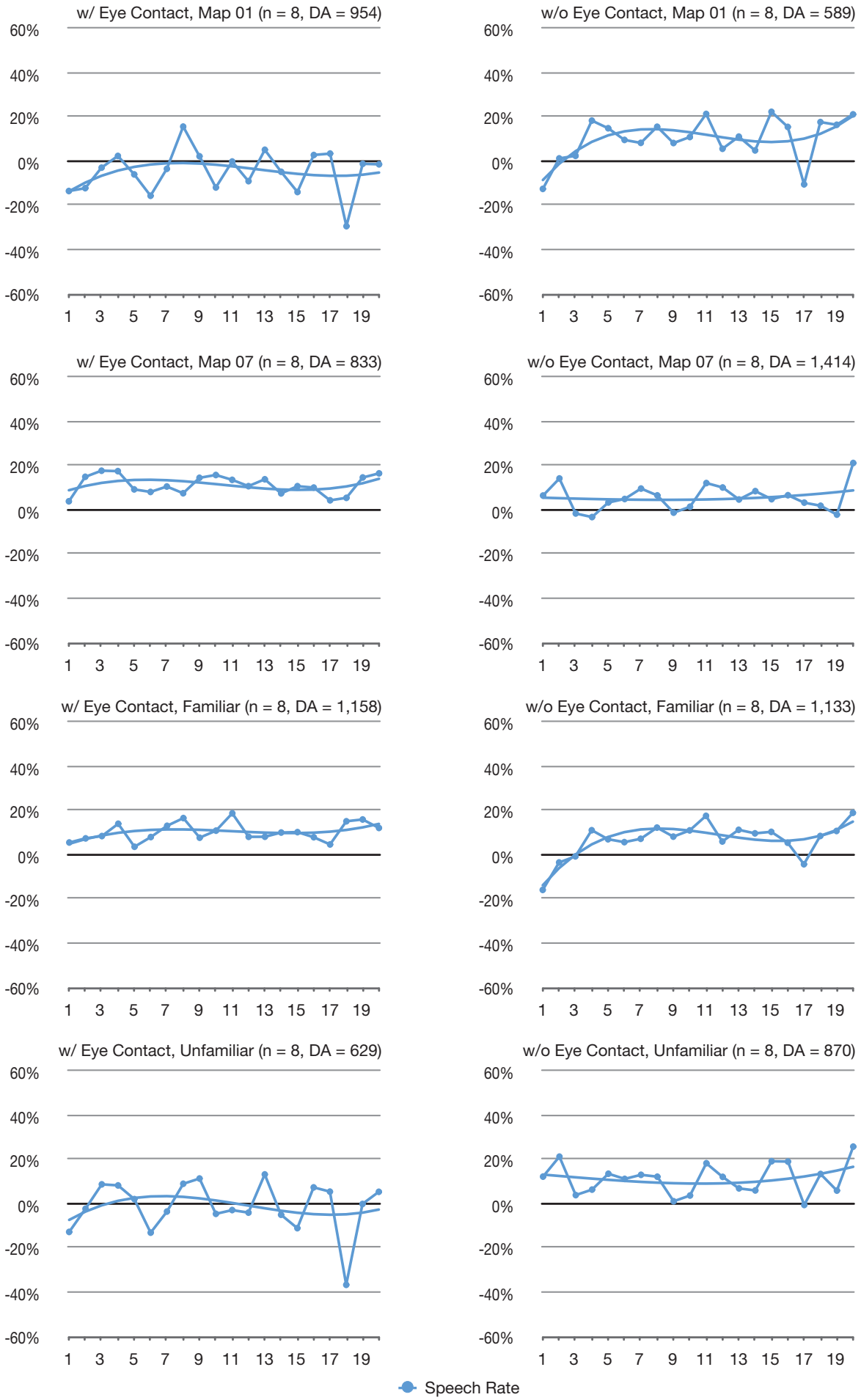


Figure B.9: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Eye Contact 02*

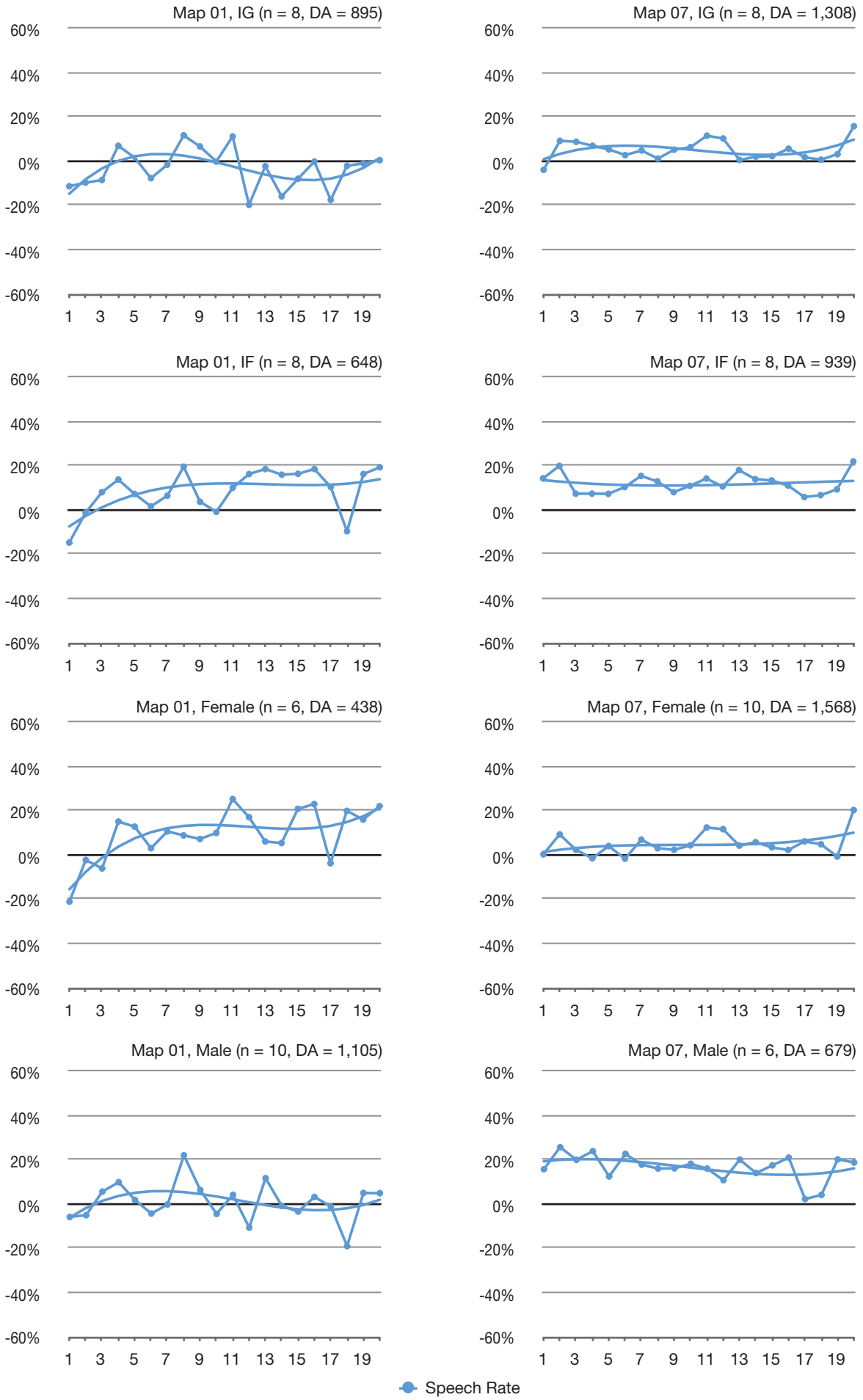


Figure B.10: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Map 01*

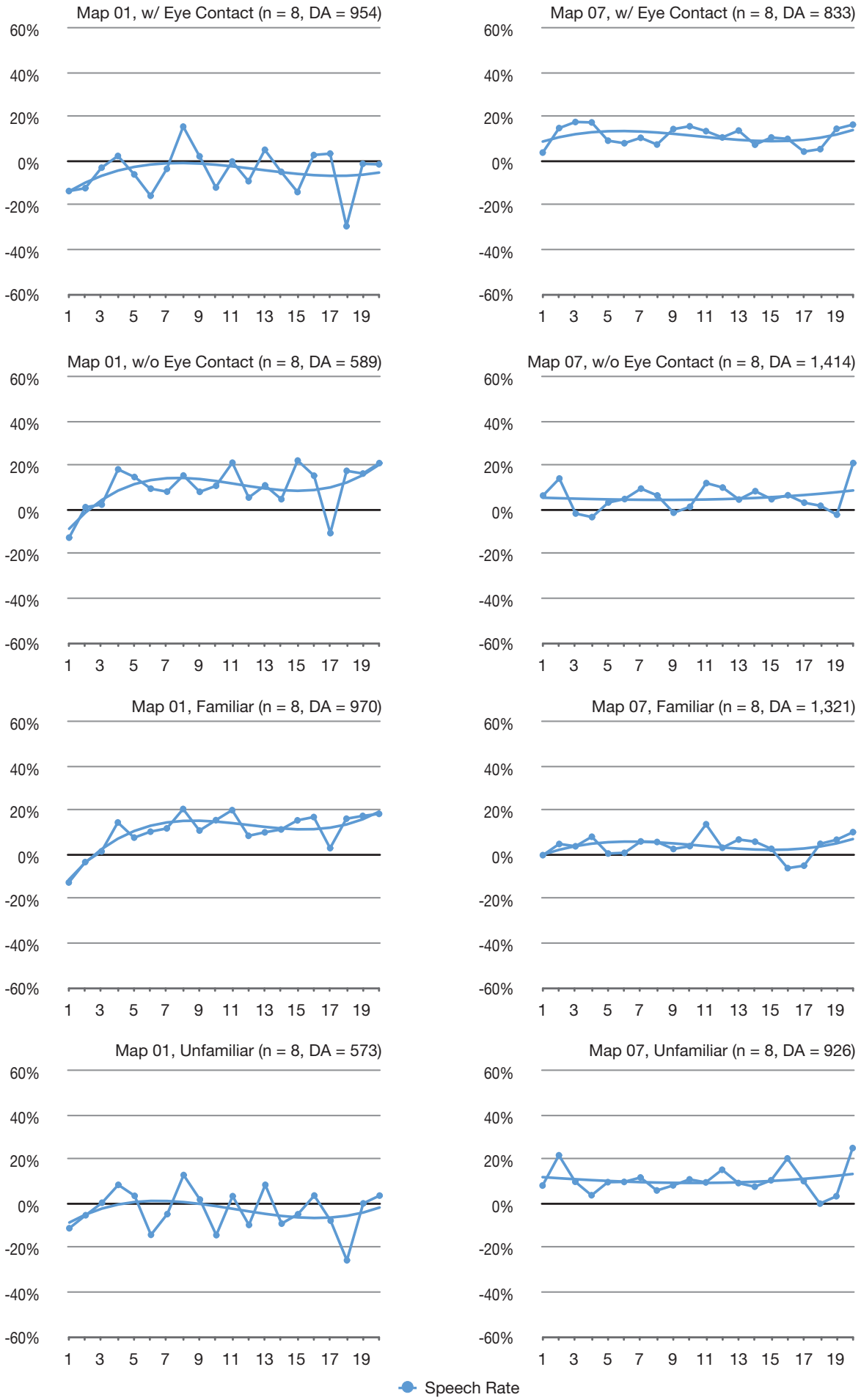


Figure B.11: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Map 02*

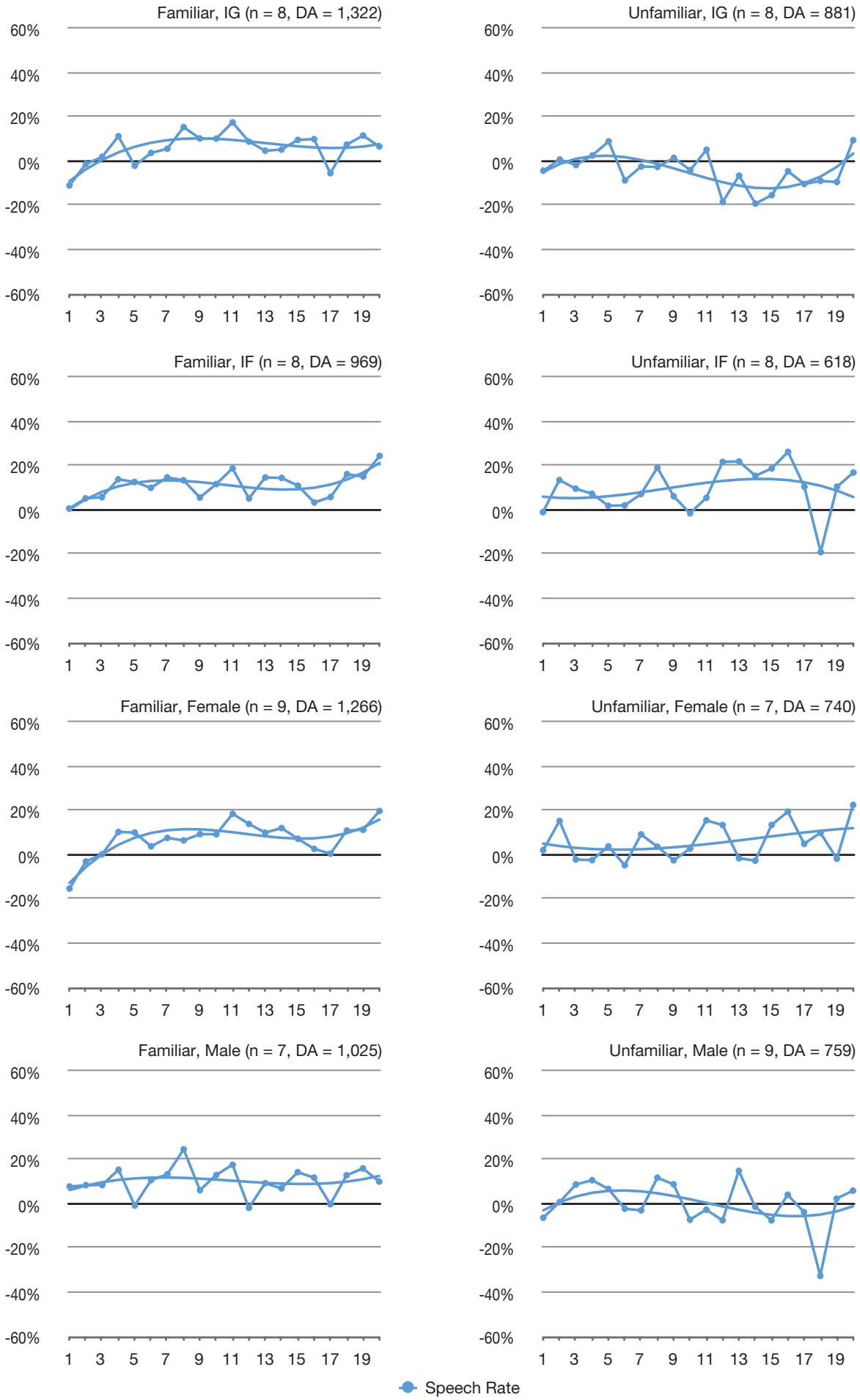


Figure B.12: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Familiarity 01*

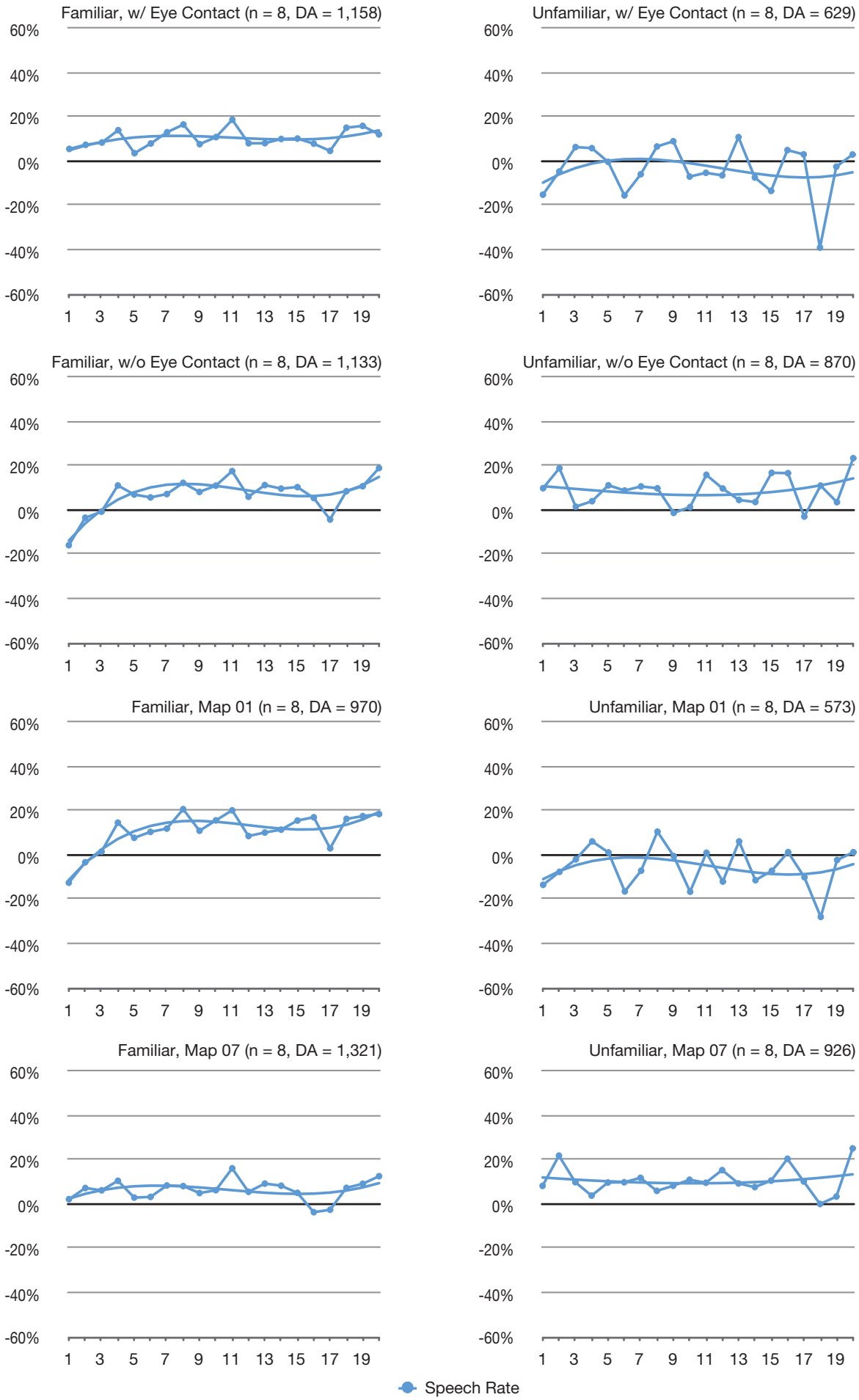


Figure B.13: *Mean Speech Rate of HCRC Edinburgh Map Task corpus — Dialogues by Familiarity 02*

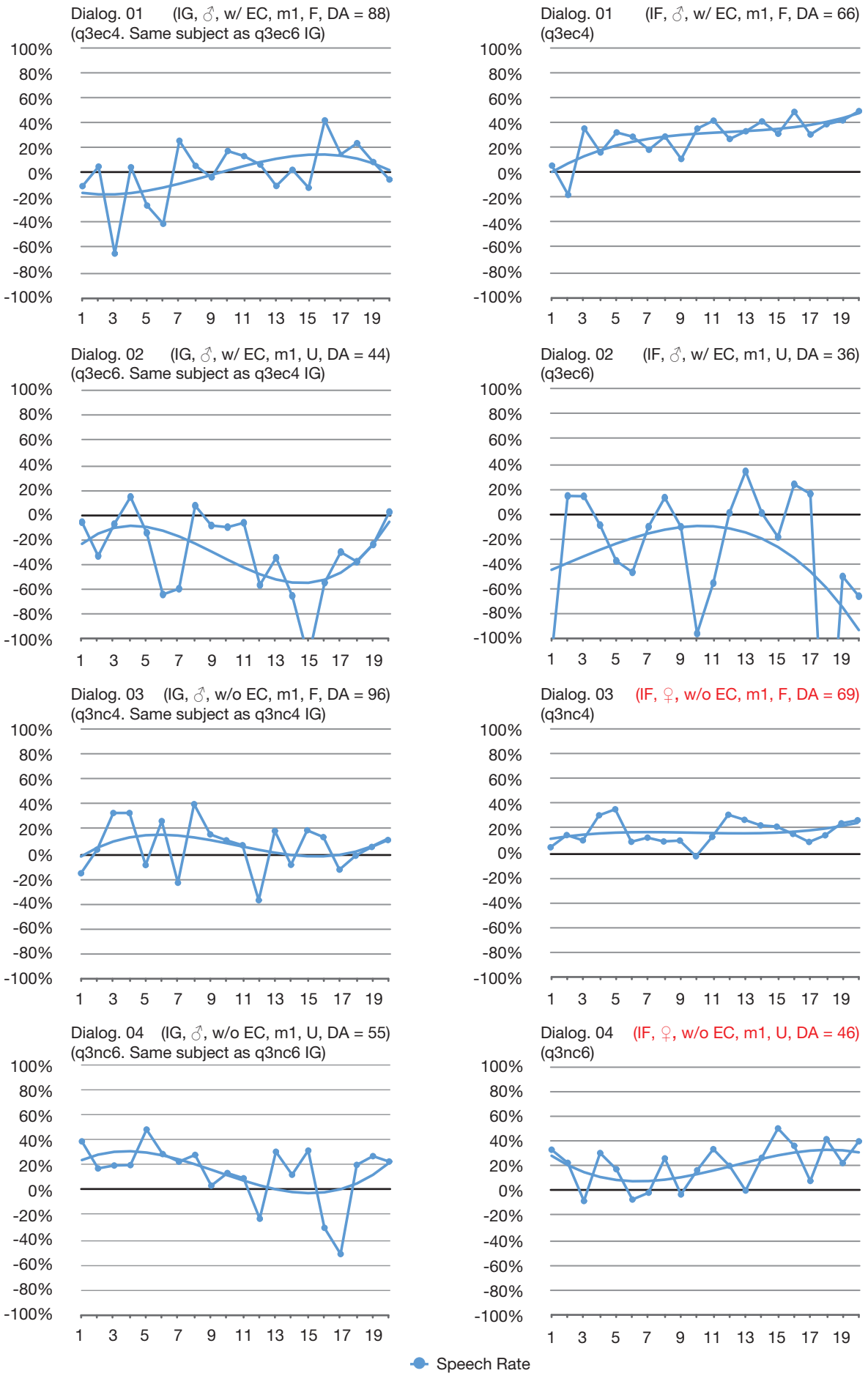


Figure B.14: *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 1 to 4*

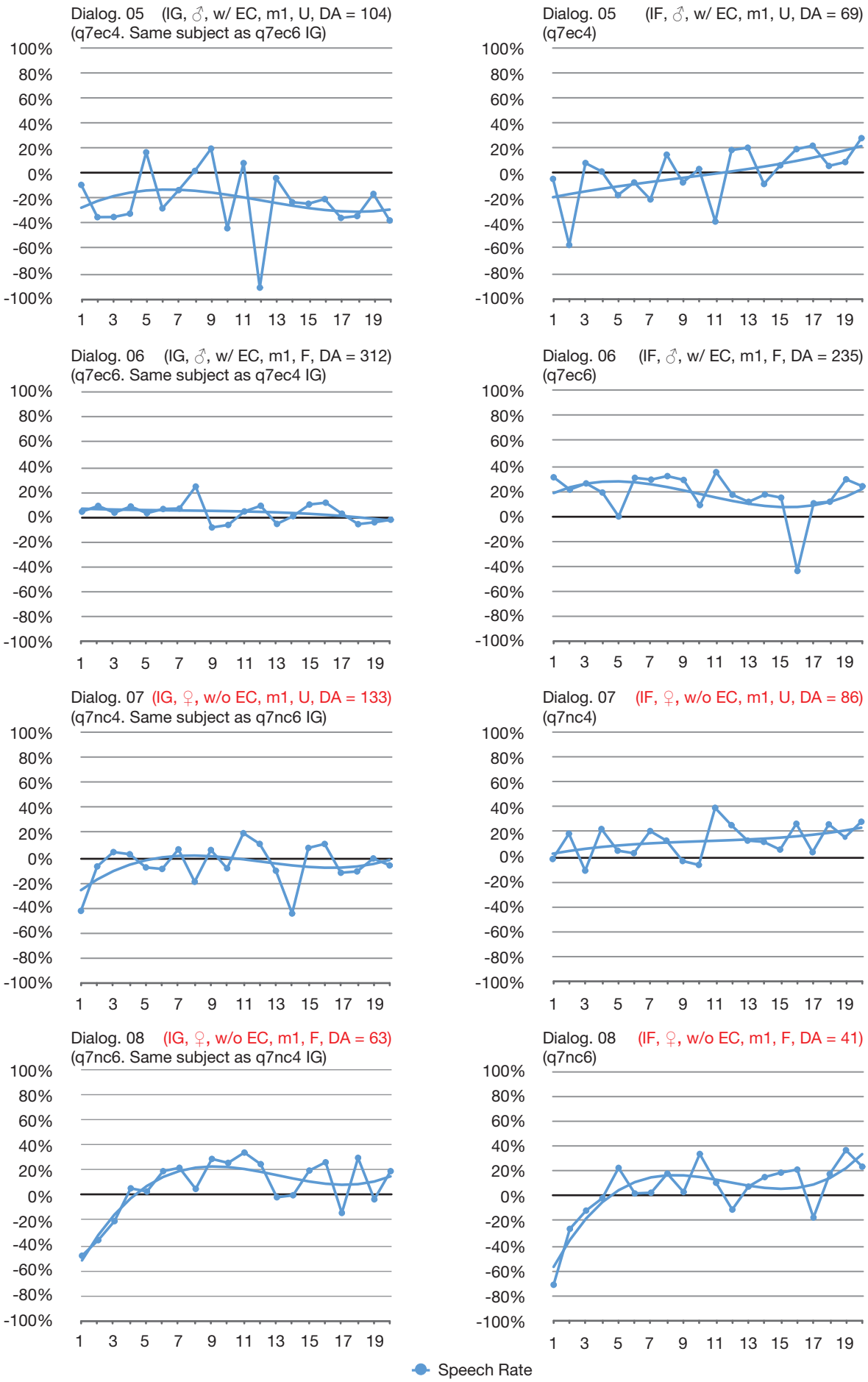


Figure B.15: *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 5 to 8*

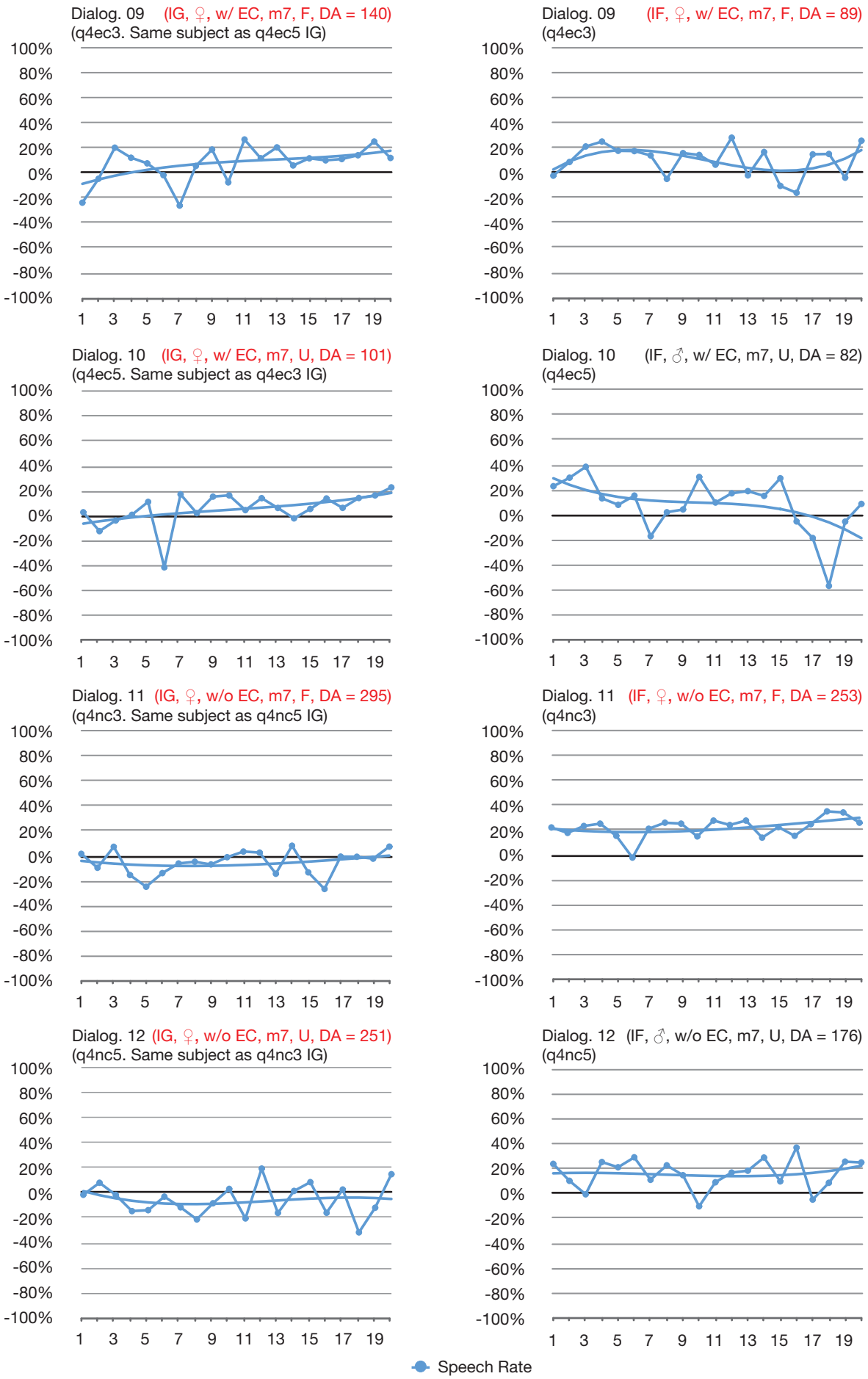


Figure B.16: *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 9 to 12*

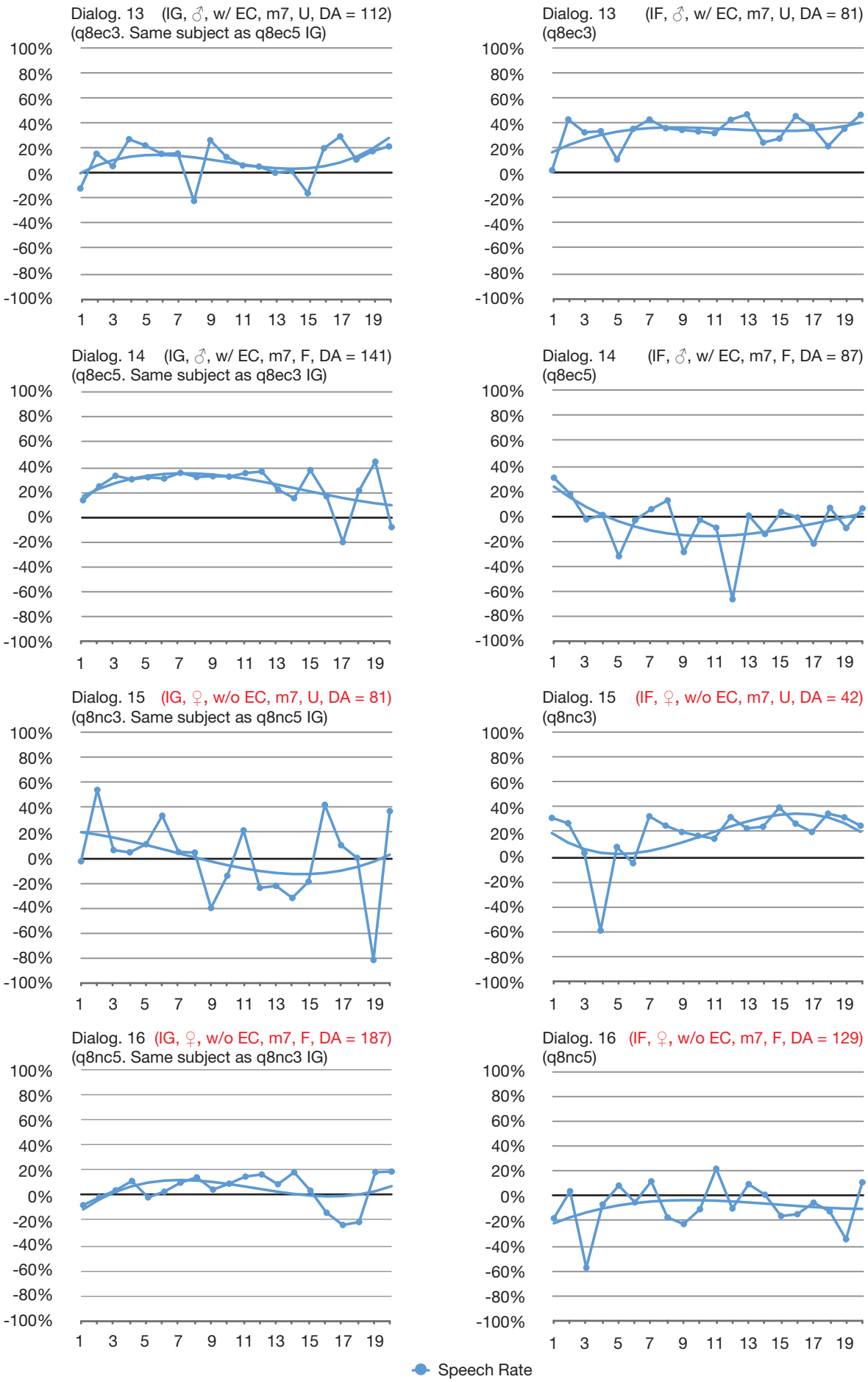


Figure B.17: *Plots of subject's Speech Rate of HCRC Edinburgh Map Task corpus — D. 13 to 16*

Appendix C

Dialogue Act Summary

Table C.1: The values of the summary and standard deviation of Dialogue Acts Word Count by Role within the HCRC Edinburgh Map Task corpus — The "Count" value is a count of dialogue acts in the dialogue and the "Sum" value is the total number of words of the dialogue act that were uttered within the dialogue, with "C-Ratio" indicating the percentage of the specific dialogue act uttered by the subject within the whole dialogue of the two subjects Information Giver and Information Follower and "S-Ratio" is the percentage of the specific dialogue act word count uttered by the subject within the whole dialogue.

Table C.2: The values of the summary and standard deviation of Dialogue Acts Word Count by Role within the ILMT- s2s corpus — The "Count" value is a count of dialogue acts in the dialogue and the "Sum" value is the total number of words of the dialogue act that were uttered within the dialogue, with "C-Ratio" indicating the percentage of the specific dialogue act uttered by the subject within the whole dialogue of the two subjects Information Giver and Information Follower and "S-Ratio" is the percentage of the specific dialogue act word count uttered by the subject within the whole dialogue.

Table C.3: Dialogue act count per corpora separated by subject role — Similar to Table C.1 but the percentage is calculated from within each role (Information Giver or Information Follower) and not the whole dialogue.

Table C.4: Dialogue act word count per corpora separated by subject role — Similar to Table C.2 but the percentage is calculated from within each role (Information Giver or Information Follower) and not the whole dialogue.

Figure C.1: Graphical comparison of the values in Table C.1 — The bar chart on the top is a graphical representation of the "C-Ratio", the bar chart in the middle is a representation of the "S-Ratio", and the boxplot a the bottom is a representation of the summary of the word count of the dialogue acts.

Figure C.2: Graphical comparison of the values in Table C.2 — The bar chart on the top is a graphical representation of the "C-Ratio", the bar chart in the middle is a representation of the "S-Ratio", and the boxplot a the bottom is a representation of the summary of the word count of the dialogue acts.

Figure C.3: Graphical comparison of the HCRC Edinburgh Map Task corpus percentage values in Table C.3 and Table C.4 — The order has been modified to show the dialogue acts with the highest proportion at the top and also connected with lines to indicate the change of order between roles of the subject: Grey highlighted section indicate 90% coverage.

Figure C.4: Graphical comparison of the ILMT-s2s corpus percentage values in Table C.3 and Table C.4 — The order has been modified to show the dialogue acts with the highest proportion at the top and also connected with lines to indicate the change of order between roles of the subject: Grey highlighted section indicate 90% coverage.

Figure C.5: Graphical comparison of the Information Giver percentage values in Table C.3 — The bar chart differs from Figure C.1 and Figure C.2 by comparing the subjects with the role of Information Giver in the two corpora as compared to the separate roles of the same corpus that are compared in Figure C.1 and Figure C.2.

Figure C.6: Graphical comparison of the Information Giver summary values in Table C.1 and Table C.2 — The boxplot differs from Figure C.1 and Figure C.2 by comparing the subjects with the role of Information Giver in the two corpora as compared to the separate roles of the same corpus that are compared in Figure C.1 and Figure C.2.

Figure C.7: Graphical comparison of the Information Follower percentage values in Table C.3 — The bar chart differs from Figure C.1 and Figure C.2 by comparing the subjects with the role of Information Follower in the two corpora as compared to the separate roles of the same corpus that are compared in Figure C.1 and Figure C.2.

Figure C.8: Graphical comparison of the Information Follower summary values in Table C.1 and Table C.2 — The boxplot differs from Figure C.1 and Figure C.2 by comparing the subjects with the role of Information Follower in the two corpora as compared to the separate roles of the same corpus that are compared in Figure C.1 and Figure C.2.

	Min.	1st Q.	*Mdn*	*M*	3rd Q.	Max.	*SD*	Count	C-Ratio	Sum	S-Ratio
HCRC IG											
Acknowledge	1	1	1	1.75	2	9	1.18	222	6.40%	388	1.88%
Align	1	1	2	4.62	7	24	5.00	301	8.68%	1,390	6.73%
Check	2	3	5	5.03	6	11	2.44	30	0.87%	151	0.73%
Clarify	1	4	8	8.98	12	32	6.01	182	5.25%	1,635	7.92%
Explain	2	5	7	9.21	12	38	6.58	124	3.58%	1,142	5.53%
Instruct	1	6	11	13.26	17	56	9.89	536	15.46%	7,108	34.43%
Query W	1	2	5.5	6.54	9.25	20	4.82	48	1.38%	314	1.52%
Query Yes/No	1	6	6	7.75	8	26	3.99	166	4.79%	1,287	6.23%
Reply No	1	1	1	2.38	3.5	11	2.13	55	1.59%	131	0.63%
Reply W	1	4	6	8.02	10	24	5.51	45	1.30%	361	1.75%
Reply Yes	1	1	1	1.87	2	13	1.78	223	6.43%	418	2.02%
IG-Total	1	1.75	5	7.42	10	56	7.73	1,932	55.71%	14,325	69.39%
HCRC IF											
Acknowledge	1	1	1	1.72	2	13	1.63	534	15.40%	918	4.45%
Align	1	4	7	8.17	12	22	5.19	30	0.87%	245	1.19%
Check	1	4	6	7.27	10	31	4.81	279	8.04%	2,029	9.83%
Clarify	2	4.75	8.5	10.20	13.25	39	8.45	20	0.58%	204	0.99%
Explain	1	5	7	8.08	9	31	4.78	128	3.69%	1,034	5.01%
Instruct	4	4	4	4	4	4	NA	1	0.03%	4	0.02%
Query W	1	2	5	5.80	8	20	4.38	56	1.61%	325	1.57%
Query Yes/No	1	4	5	6.60	8	21	3.98	63	1.82%	416	2.02%
Reply No	1	1	1	2.00	1	13	2.66	52	1.50%	104	0.50%
Reply W	1	4	6	7.66	11	22	5.03	67	1.93%	513	2.48%
Reply Yes	1	1	1	1.72	1	12	1.74	306	8.82%	527	2.55%
IF-Total	1	1	2	4.11	6	39	4.45	1,536	44.29%	6,319	30.61%
HCRC Total	1	1	4	5.95	8	56	6.69	3,468	100%	20,644	100%

Table C.1: *Summary and standard deviation of Dialogue Acts Word Count by Role within the HCRC Edinburgh Map Task corpus.*

	Min.	1st Q.	*Mdn*	*M*	3rd Q.	Max.	*SD*	Count	C-Ratio	Sum	S-Ratio
ILMT-s2s IG											
Acknowledge	1	1	1	1.54	1	9	1.45	43	3.13%	66	0.89%
Align	1	1	1	3.47	4	14	4.05	17	1.24%	59	0.80%
Check	2	2	2	2.00	2	2	0	3	0.22%	6	0.08%
Clarify	1	1	1	2.25	3	13	2.63	40	2.91%	90	1.21%
Explain	1	3	4	4.82	6	27	3.38	129	9.38%	622	8.39%
Instruct	1	4	6	8.23	10	178	13.88	304	22.09%	2,503	33.78%
Query W	1	3	5	5.13	6.5	13	3.08	39	2.83%	200	2.70%
Query Yes/No	1	4	5	5.84	7	13	2.30	130	9.45%	759	10.24%
Reply No	1	1	3	3.33	5	7	2.55	9	0.65%	30	0.40%
Reply W	1	2	4.5	5.13	6.25	22	2.55	48	3.49%	246	3.32%
Reply Yes	1	1	1	1.62	1	7	1.43	45	3.27%	73	0.99%
IG-Total	1	2	5	5.77	7	178	9.10	807	58.65%	4,654	62.81%
ILMT-s2s IF											
Acknowledge	1	1	1	1.93	2.75	6	1.43	54	3.92%	104	1.40%
Align	1	5	6	6.81	9	21	4.02	52	3.78%	354	4.78%
Check	3	6	7	8.88	10	24	5.49	17	1.24%	151	2.04%
Clarify	1	2	4	4.37	5.5	17	3.34	35	2.54%	153	2.06%
Explain	1	3	5	5.53	7	18	3.15	158	11.48%	874	11.79%
Instruct	1	2.5	4	4.67	6	9	2.61	15	1.09%	70	0.94%
Query W	1	2	4	4.93	7	21	4.01	94	6.83%	463	6.25%
Query Yes/No	2	4	5	5.52	6.75	12	2.31	50	3.63%	276	3.72%
Reply No	1	1	3	3.35	5	8	2.3	26	1.89%	87	1.17%
Reply W	1	4	5	5.69	7	15	3.33	13	0.94%	74	1.00%
Reply Yes	1	1	2	2.73	4	9	2.07	55	4.00%	150	2.02%
IF-Total	1	2	4	4.84	6	24	3.52	569	41.35%	2,756	37.19%
ILMT-s2s Total	1	2	4	5.39	7	178	7.34	1,376	100%	7,410	100%

Table C.2: *Summary and standard deviation of Dialogue Acts Word Count by Role within the ILMT-s2s corpus.*

Dialogue Act type	HCRC Edinburgh Map Task corpus				ILMT-s2s corpus (Eng)			
	IG Count	IG %	IF Count	IF %	IG Count	IG %	IF Count	IF %
Acknowledge	222	11.5%	534	34.8%	43	5.3%	54	9.5%
Align	301	15.6%	30	2.0%	17	2.1%	52	9.1%
Check	30	1.6%	279	18.2%	3	0.4%	17	3.0%
Clarify	182	9.4%	20	1.3%	40	5.0%	35	6.2%
Explain	124	6.4%	128	8.3%	129	16.0%	158	27.8%
Instruct	536	27.7%	1	0.1%	304	37.7%	15	2.6%
Query W	48	2.5%	56	3.6%	39	4.8%	94	16.5%
Query Yes/No	166	8.6%	63	4.1%	130	16.1%	50	8.8%
Reply No	55	2.8%	52	3.4%	9	1.1%	26	4.6%
Reply W	45	2.3%	67	4.4%	48	5.8%	13	2.3%
Reply Yes	223	11.5%	306	19.9%	45	5.6%	55	9.7%
Total	1,932	100%	1,536	100%	807	100%	569	100%

Table C.3: *Dialogue act occurrence count per corpora separated by subject role*

Dialogue Act type	HCRC Edinburgh Map Task corpus				ILMT-s2s corpus (Eng)			
	IG Count	IG %	IF Count	IF %	IG Count	IG %	IF Count	IF %
Acknowledge	388	2.7%	918	14.5%	66	1.4%	104	3.8%
Align	1,390	9.7%	245	3.9%	59	1.3%	354	12.8%
Check	151	1.1%	2,029	32.1%	6	0.1%	151	5.5%
Clarify	1,635	11.4%	204	3.2%	90	1.9%	153	5.6%
Explain	1,142	8.0%	1,034	16.4%	622	13.4%	874	31.7%
Instruct	7,108	49.6%	4	0.1%	2,503	53.8%	70	2.5%
Query W	314	2.2%	325	5.1%	200	4.3%	463	16.8%
Query Yes/No	1,287	9.0%	416	6.6%	759	16.3%	276	10.0%
Reply No	131	0.9%	104	1.6%	30	0.6%	87	3.2%
Reply W	361	2.5%	513	8.1%	246	5.3%	74	2.7%
Reply Yes	418	2.9%	527	8.3%	73	1.6%	150	5.4%
Total	14,325	100%	6,319	100%	4,654	100%	2,756	100%

Table C.4: *Dialogue act word count per corpora separated by subject role*

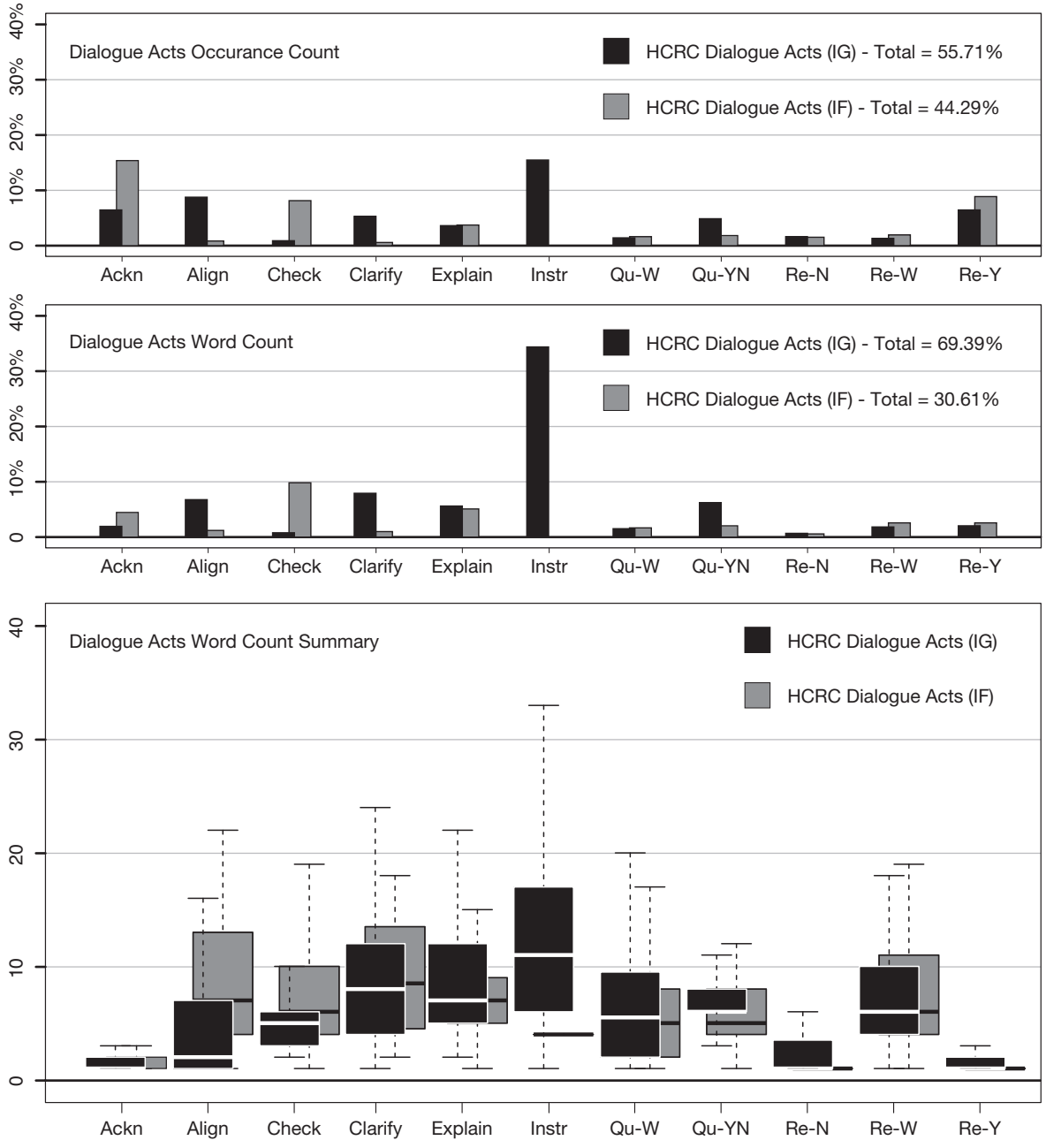


Figure C.1: *Bar Chart Role Ratio Comparison of Dialogue Acts within HCRC Edinburgh Map Task Corpus (Top and Middle) and Boxplot Comparison of said corpus Dialogue Act Word Count (Bottom).*

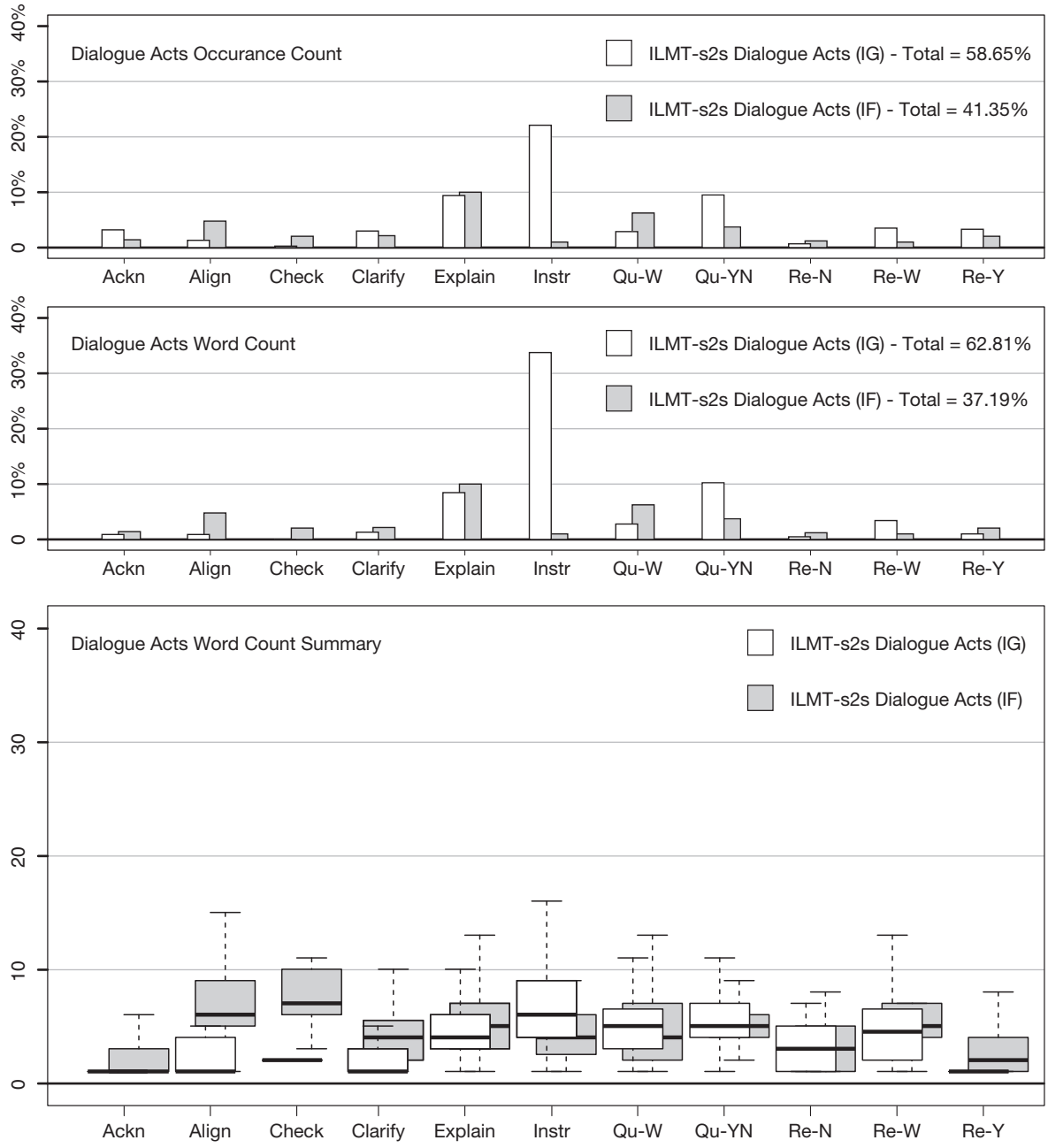


Figure C.2: *Bar Chart Role Ratio Comparison of Dialogue Acts within ILMT-s2s Corpus (Top and Middle) and Boxplot Comparison of said corpus Dialogue Act Word Count (Bottom).*

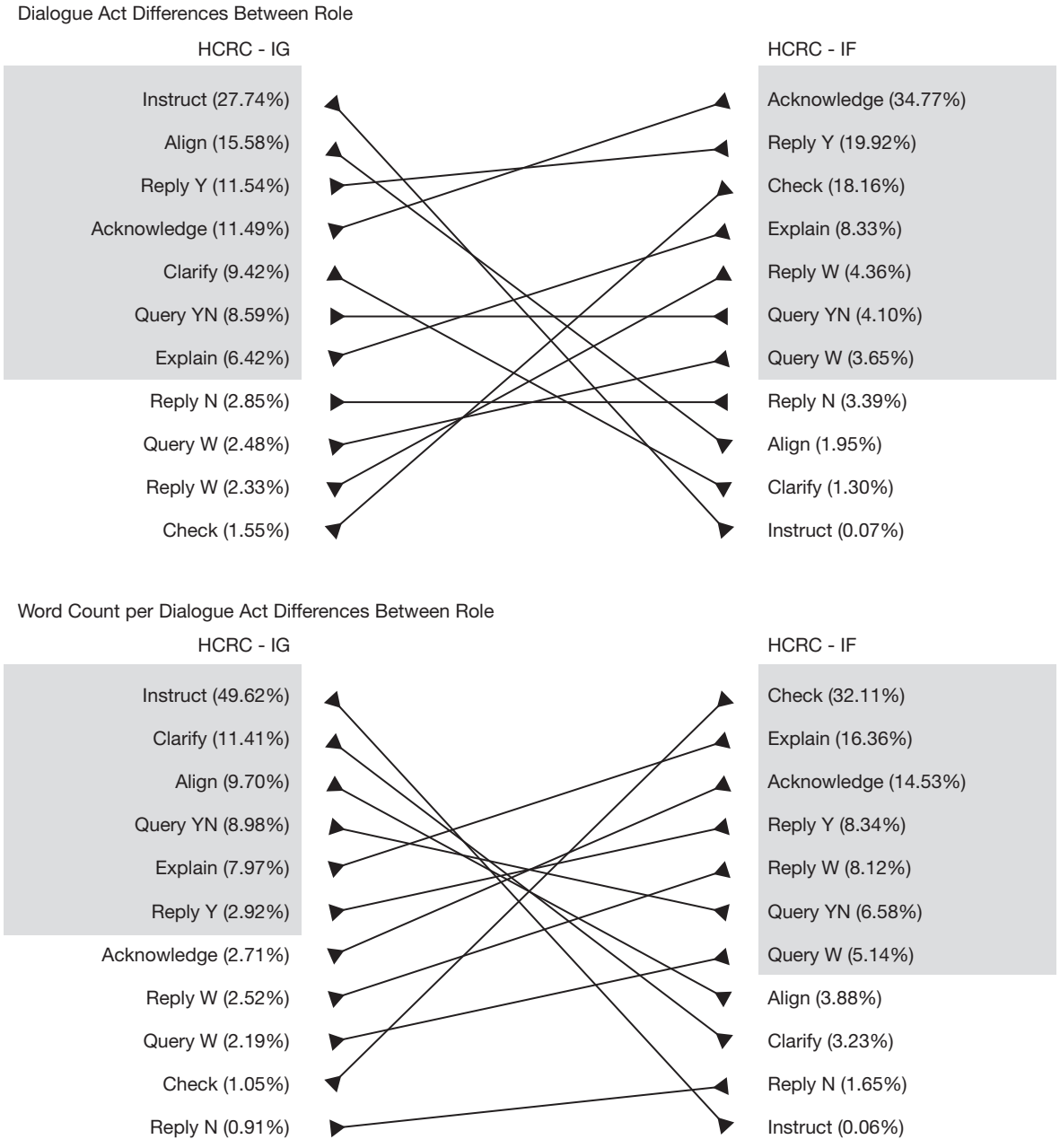


Figure C.3: *Frequency of Occurrence Differences Between the Role within the HCRC Map Task corpus: Grey highlighted sections indicate 90% coverage.*

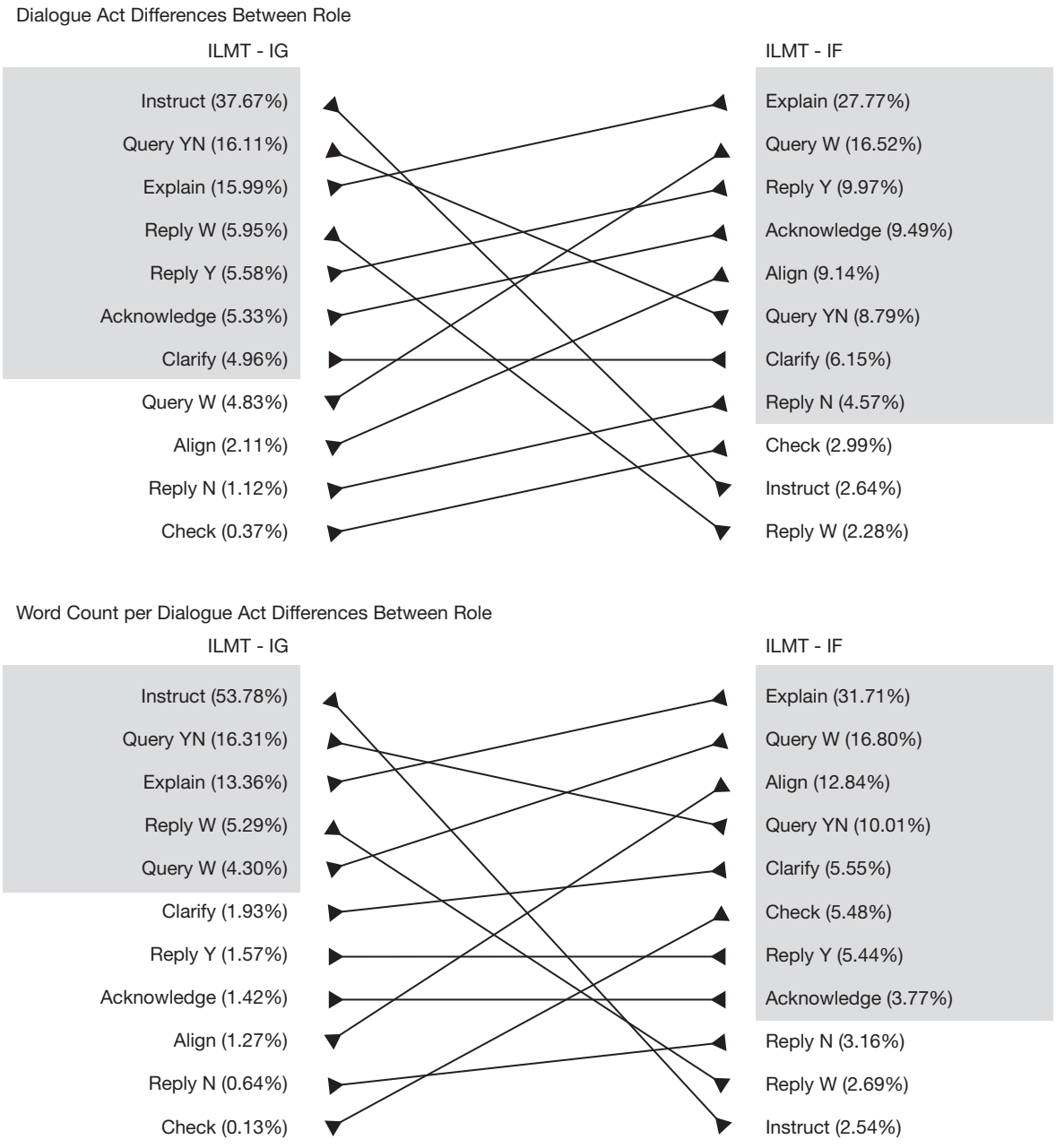


Figure C.4: *Frequency of Occurrence Differences Between the Role within the ILMT-s2s corpus: Grey highlighted sections indicate 90% coverage.*

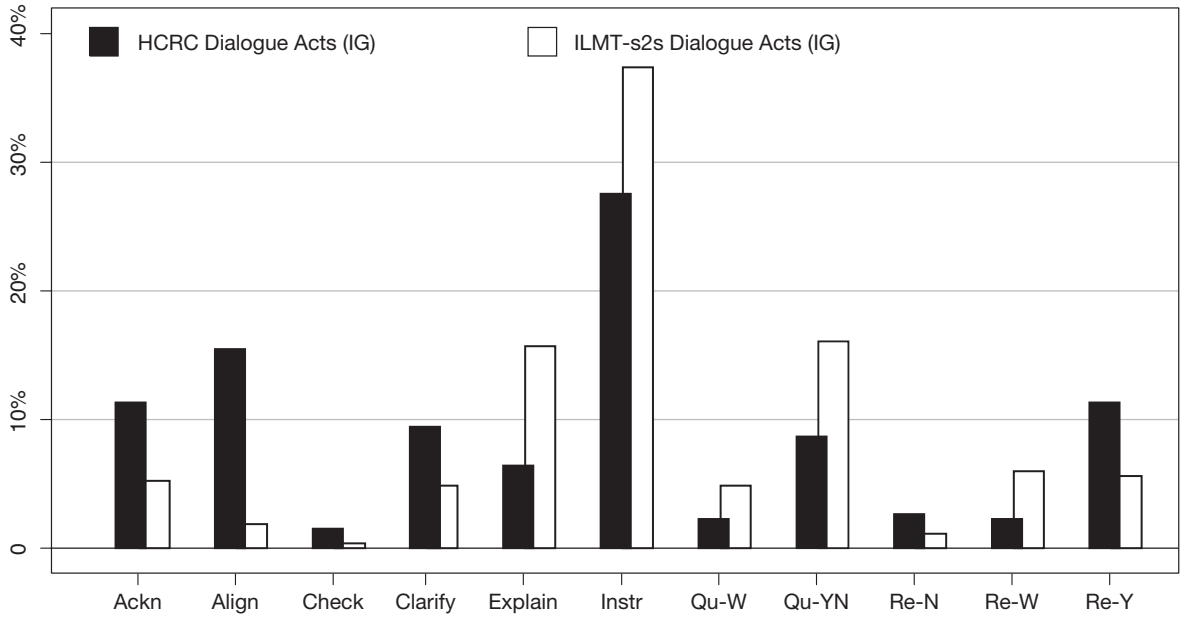


Figure C.5: *Bar Chart Comparison of IG Dialogue Act Percentage of Each Corpus*

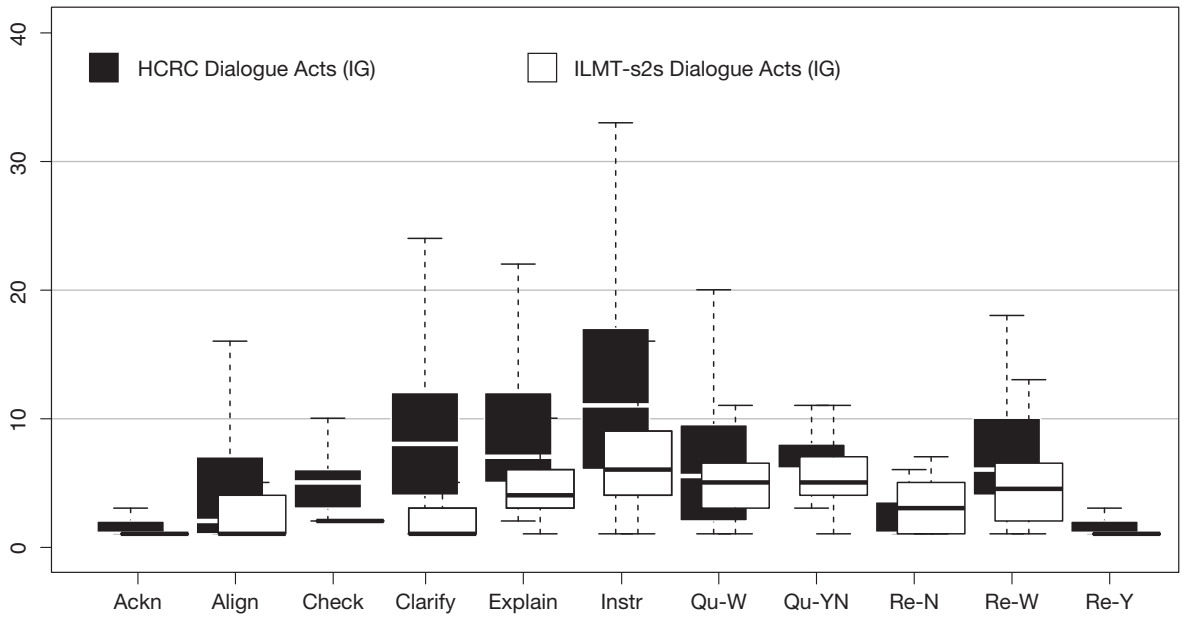


Figure C.6: *Word Count Boxplot of IG Dialogue Acts of Each Corpus*

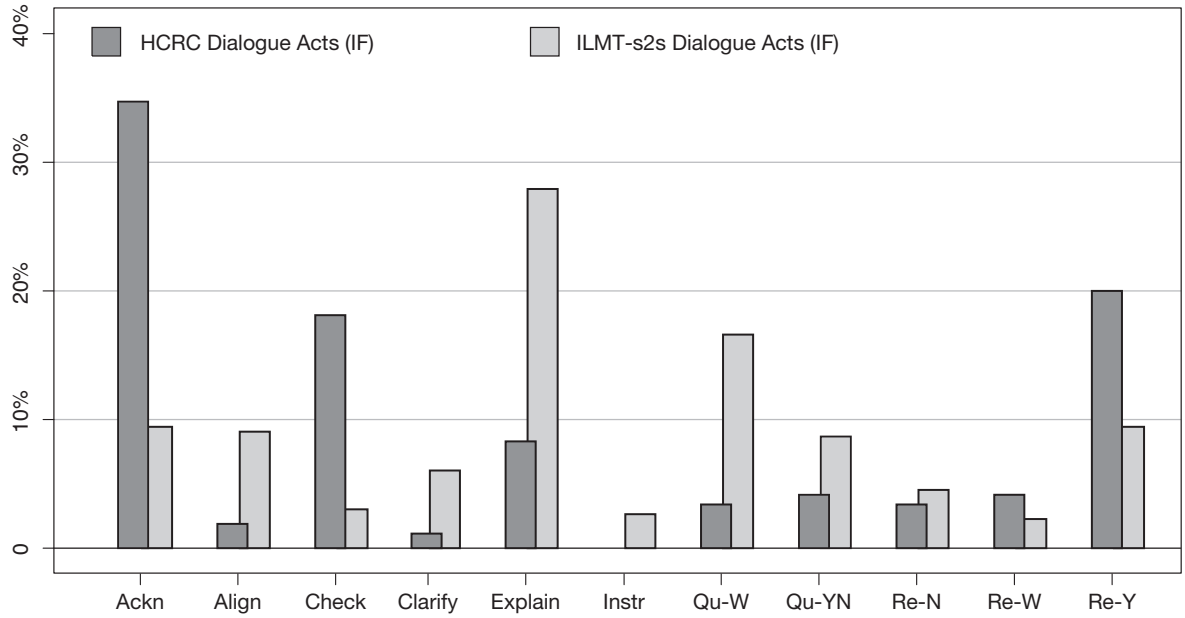


Figure C.7: *Bar Chart Comparison of IF Dialogue Act Percentage of Each Corpus*

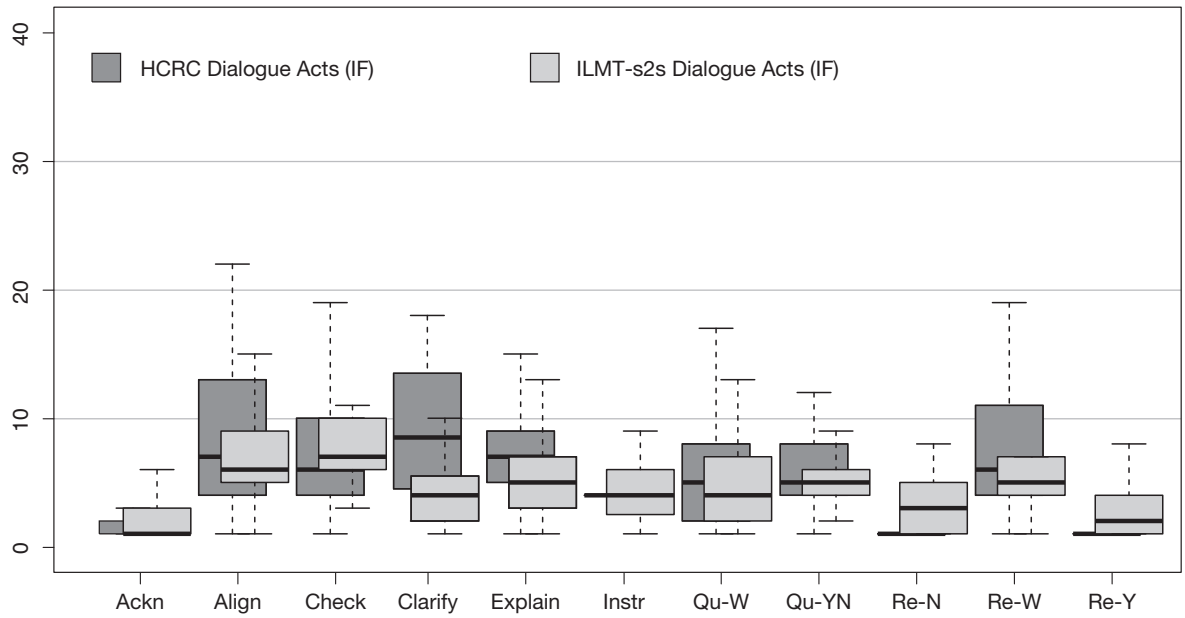


Figure C.8: *Word Count Boxplot of IF Dialogue Acts of Each Corpus*

Glossary

Biosignal Type

biosignal Refers to either EEG, BVP, HR or SC signals recorded using the Mind Media Nexus-4 during the ILMT-s2s corpus collection .... ii, 8, 9, 44, 45, 60–62, 75–79, 102, 119, 219, 220, 222, 227

BVP Refers to Blood Volume Pulse measurements 44, 45, 61, 62, 75, 102

EEG Refers to Electroencephelography measurements 44, 45, 61, 62, 75, 102, 267

HR Refers to Heart Rate measurements calculated from the BVP 44, 45, 61, 62, 75

SC Refers to Skin Conductance measurements 44, 45, 61, 62, 75, 102

Dialogue Details

female Refers to female subjects of the ILMT-s2s corpus collection and/or data and results from female subjects 5, 6, 15, 29, 30, 70, 119, 133, 135, 136, 139–142, 144–147, 163, 166, 167, 170, 171, 173, 176, 194, 199, 201, 213, 214, 217, 220, 221, 229, 280, 288, 289, 291, 309

♀ Abbreviation of "female" that is used in figures and tables of this thesis .... 28, 29, 34, 59, 70, 75, 121, 122, 125–129, 166, 167, 170–172, 246, 247, 250–252, 261–276, 278, 279, 304–308, *see also* female

male Refers to male subjects of the ILMT-s2s corpus collection and/or data and results from male subjects 5, 6, 15, 29, 30, 70, 119, 133, 135, 136, 139–142, 144–147, 163, 166, 167, 170, 171, 173, 176, 194, 199, 201, 214, 217, 220, 221, 229, 280, 288, 289, 291, 309

♂ Abbreviation of "male" that is used in figures and tables of this thesis .... 28, 29, 34, 59, 70, 75, 121, 122, 125–129, 166, 167, 170–172, 246, 247, 250–252, 261–276, 278, 279, 304–308, *see also* male

Information Follower

Refers to subjects with the role of following instructions in the ILMT-s2s corpus

collection and/or data and results from Information Follower subjects — also found to be referred to as Instruction Follower in literature. 5, 6, 10, 17, 28, 30, 31, 60, 62, 65, 66, 68, 81, 115, 119, 132, 133, 135, 136, 139, 140, 142, 144, 146, 147, 149, 151–155, 158–160, 162, 163, 165, 168, 170, 171, 173, 176, 178, 191, 193, 194, 199–201, 203, 204, 206, 208–211, 214, 215, 220, 221, 280, 288, 290, 293–295, 325, 326

IF Abbreviation of "Information Follower" that is used in figures and tables of this thesis .... 28, 29, 31, 33, 35, 65, 66, 70, 75, 79, 92, 121–123, 125–129, 132, 134, 141, 150, 152, 165–167, 170–172, 188, 189, 202, 204, 208–210, 224, 225, 246–248, 250–253, 260–280, 292, 299–302, 304–308, 327–330, *see also* Information Follower

Information Giver

Refers to subjects with the role of giving instructions in the ILMT-s2s corpus collection and/or data and results from Information Giver subjects — also found to be referred to as Instruction Giver in literature. 5, 6, 17, 28, 30, 31, 49, 60, 62, 65, 67, 68, 81, 115, 119, 132, 133, 135, 136, 139, 140, 142, 144, 146, 147, 149, 151–153, 155, 158–160, 162, 163, 165, 168, 170, 171, 173, 176, 178, 191, 193, 194, 198–201, 203, 204, 206, 208–211, 214, 215, 220, 221, 280, 288, 290, 293–295, 325, 326

IG Abbreviation of "Information Giver" that is used in figures and tables of this thesis .... 28, 29, 31, 33, 35, 65, 66, 70, 75, 79, 92, 121–123, 125–129, 132, 134, 141, 150, 152, 165–167, 170–172, 188, 189, 202, 204, 208–210, 224, 225, 246–248, 250–253, 260–280, 292, 299–302, 304–308, 327–330, *see also* Information Giver

Eng Refers to subjects that spoke English for the ILMT-s2s corpus collection and/or data and results from English speaking subjects. Three letter abbreviation — ISO 639-3 .... 75, 128, 129, 207, 210, 278, 279, 329, 330, 338, *see also* En

En Two letter abbreviation of Eng — ISO 639-1 .... 38, 81, 96, 100, 104, 105, 109, 112, 113, 121, 122, 125–127, 162–164, 166, 167, 170–172, 186, 188, 189, 200, 204, 246, 247, 250–252, 261–276, 288, 289, 291, *see also* Eng

Por Refers to subjects that spoke Portuguese for the ILMT-s2s corpus collection and/or data and results from Portuguese speaking subjects. Three letter abbreviation — ISO 639-3 .... 75, 128, 129, 207, 210, 278, 279, 338

Pt Two letter abbreviation of Por — ISO 639-1 .... 38, 81, 96, 104, 105, 109, 112, 113, 121, 122, 125–127, 163, 164, 166, 167, 170–172, 186, 188, 189, 200, 204, 246, 247, 250–252, 261–276, 288, 289, 291

Map 01 Refers to subjects that used Map 01 in the ILMT-s2s corpus collection and/or

data and results from subjects that used that map 29–31, 35, 70, 72, 79, 194, 199, 201, 224, 277–280, 288, 289, 292, 304

Map 07 Refers to subjects that used Map 07 in the ILMT-s2s corpus collection and/or data and results from subjects that used that map 29–31, 35, 70, 72, 79, 194, 199, 201, 224, 277–280, 288, 289, 292, 304

w/ Bio Refers to subjects that used the biosignal recording devices during the ILMT-s2s corpus data collection and/or data and results from subjects that collected biosignals 61, 62, 77, 119, 128, 129, 226, 261–276, 278, 279

w/o Bio Refers to subjects that **did not** use the biosignal recording devices during the ILMT-s2s corpus data collection and/or data and results from subjects that **did not** collected biosignals 61, 62, 77, 119, 128, 129, 226, 261–276, 278, 279

w/ Crash Refers to subjects that did experience an ILMT-s2s System crash during the ILMT-s2s corpus data collection and/or data and results from subjects that did experience a system crash 75, 170–172, 261–268, 270–275, 278, 279

w/o Crash Refers to subjects that **did not** experience an ILMT-s2s System crash during the ILMT-s2s corpus data collection and/or data and results from subjects that **did not** experience a system crash 170, 171, 261–276, 278, 279

w/ Video Refers to subjects that could see their interlocutor during the ILMT-s2s corpus data collection and/or data and results from subjects that could see their interlocutor .... xxii, 6, 59, 78, 81, 87–90, 119, 128, 129, 136, 139, 140, 163, 167, 168, 170, 171, 178, 180, 181, 186, 188, 189, 192–194, 199, 201–205, 208–210, 212, 215, 221, 261–275, 280, 288, 290, 292–295

w/ Vid Abbreviation of "w/ Video" that is used in figures and tables of this thesis .... 35, 75, 79, 166, 167, 171, 172, 178, 205, 208–211, 278, 279, *see also* w/ Video

w/o Video Refers to subjects that **could not** see their interlocutor during the ILMT-s2s corpus data collection and/or data and results from subjects that **could not** see their interlocutor 6, 59, 78, 81, 87, 88, 90, 119, 128, 129, 136, 139, 140, 163, 167, 168, 170, 178, 180, 181, 186, 188–190, 192–194, 198, 199, 201–205, 208–212, 215, 221, 261–276, 280, 288, 290, 292–295

w/o Vid Abbreviation of "w/o Video" that is used in figures and tables of this thesis .... 35, 75, 79, 166, 167, 171, 172, 178, 205, 208, 209, 211, 278, 279, *see also* w/o Video

ILMT-s2s System Related

ASR Abbreviation of Automatic Speech Recognition. For the ILMT-s2s System the Google Speech API was used. 4–6, 8, 16, 20, 24, 39, 42–47, 49–53, 55,

57, 82, 88, 91–94, 97–103, 130–133, 161–164, 166–169, 172, 173, 176, 178–184, 186, 189, 193, 207, 214, 215, 217–223, 229, 261, 263–267, 269, 271–275

ASR Performance measures

***F*-Score** Refers to a measurement of ASR performance. It differs from WER by having a range between 0–1 with 0 indicating that all words are wrong and 1 indicating that all words have been correctly recognised 39, 97–99, 162, 163

Reverse *F*-Score

Refers to the *F*-Score but with the measurements reversed so that 0 indicates all words have been recognised correctly and 1 indicates that all words have been incorrectly recognised by the ASR component of the ILMT-s2s System .... 97, 98, 101, 164, 166, 167, 171–173, 189, 229, *see also* *F*-Score

WER Abbreviation of Word Error Rate. In this thesis it is referring to the WER of the ASR component in the ILMT-s2s System 16, 39, 97–99, 103, 130, 131, 161–163, 178, 180, 183, 189, 190, 207, 217

MT Abbreviation of Machine Translation. For the ILMT-s2s System the Microsoft Bing translation system was used. 2, 18, 19, 24, 43, 44, 46, 47, 50–53, 55, 58, 82, 88, 91–94, 102, 207, 219, 222, 223

TTS Abbreviation of Text-to-Speech. For the ILMT-s2s System the macOS TTS system voice was used. 21, 33, 38, 40, 43, 44, 46, 47, 51, 53–55, 58, 72, 82, 91–94, 96, 98, 101, 102, 104, 108, 110–115, 130, 144, 145, 147, 164, 179, 181, 182, 186, 196, 219, 222, 229, 340

speech rate In this thesis, it refers to the percentage (%) difference of the subject's utterance when compared with the TTS output of the ILMT-s2s System.

wpm Abbreviation of Word Per Minute 33, 38, 96, 103–111, 113–115, 130, 144–147, 229

Perceived Emotions

Amused Refers to the perceived emotions where the subject is seen to be in a state of amusement 16, 24, 25, 37, 82–84, 182–184, 186, 187, 189, 190, 215, 217, 223

Frustrated Refers to the perceived emotions where the subject is seen to be in a state of frustration 16, 24, 25, 37, 82–84, 182, 183, 186, 187, 223

Neutral Refers to a state where the subject is **not** seen to be in a state of *Amused* / *Frustrated* / *Surprised* 180, 183, 186–189

Surprised Refers to the perceived emotions where the subject is seen to be in a state of surprise 16, 24, 25, 37, 82–84, 88, 182, 183, 187, 188, 190, 222, 223

Recording Equipment

Audio-Technica HYP-190H

Refers to Microphone used during the ILMT-s2s corpus data collection — used by subject without biosignal recording devices 61, 62

Mind Media Nexus-4

Refers to Device used to record biosignal during the ILMT-s2s corpus data collection from Mind Media B.V..... 61, 62, 75, 226, 227

Panasonic HX-A100

Refers to Wearable camcorder used during the ILMT-s2s corpus data collection — used by subject without biosignal recording devices 61, 62, 226

Sennheiser MKE104

Refers to Microphone used during the ILMT-s2s corpus data collection — used by subject with biosignal recording devices 61, 62

SMI Eye Tracking Glasses

Refers to Eye Tracking Glasses used during the ILMT-s2s corpus data collection — used by subject with biosignal recording devices 61, 62, 75, 79, 226

SONY HDR-CX370

Refers to the make and model name of the camcorder used to record the back view of one of the subjects during the recording of the ILMT-s2s corpus data 226

SONY HDR-XR500

Refers to the make and model name of the camcorder used to record the subjects face and upper torso during the recording of the ILMT-s2s corpus data 61, 62, 78, 226, 228

Video Recording Details

1080i Refers to interlace video recorded at a resolution of 1080 horizontal lines 61, 62

1080p Refers to progressive video recorded at a resolution of 1080 horizontal lines 61, 62

29.97 fps Refers to video frame recording speed of 29.97 frames per second 61, 62, 76, 226

30 fps Refers to video frame recording speed of 30 frames per second 61, 62, 76, 226

960p Refers to progressive video recorded at a resolution of 960 horizontal lines 61, 62, 76

HD Video Refers to High-definition Video (HD Video) is video of a resolution of 1080 or 720 horizontal lines 62

Talk Type

On-Talk Refers to utterances made by the subject towards their interlocutor but mediated by the S2S-MT system 9, 14, 16, 20, 35, 37–39, 46, 82, 86–88, 90, 91, 93, 94, 98, 100, 117–128, 130–136, 139, 140, 142, 144, 150, 153, 154, 158–161, 163–166, 173, 176, 179, 180, 182, 183, 215, 222, 224, 225, 227, 245–247, 249, 252

Off-Talk Refers to utterances made by the subject towards their interlocutor but ***not*** mediated by the S2S-MT system 9, 20, 22, 25, 26, 35, 38, 82, 85–88, 90, 91, 93, 94, 117–122, 124–126, 130, 158–160, 179, 182, 217, 222, 225, 245, 246, 249

Off-Talk Other

Refers to utterances made by the subject towards their interlocutor (direct Human-to-Human communication) 14, 15, 37, 38, 82, 90, 91, 93, 94, 117–119, 122–124, 126–128, 130, 131, 133, 134, 138, 139, 142, 144, 161, 173, 176, 245–247, 249, 252

Off-Talk Self

Refers to utterances made by the subject towards themselves (talking to themselves) 14, 15, 37, 38, 82, 90, 91, 93, 94, 117–119, 122–124, 126–128, 130, 131, 139, 144, 161, 176, 245–247, 249, 252

OOT Abbreviation of "Other forms of *Off-Talk*". A term used by Oppermann et al. and Batliner et al. 26, 91, 117

ROT Abbreviation of "Read *Off-Talk*". A term used by Oppermann et al. and Batliner et al. 26, 91, 117

Other

API Abbreviation for Application Programming Interface 46, 47, 50, 93, 102

H2C Abbreviation for Human-to-Computer 2, 3, 45

H2H Abbreviation for Human-to-Human 1–4, 12, 15, 19, 24, 44, 45, 133, 142, 159, 219, 220, 223, 264

HCRC Edinburgh Map Task corpus

Refers to the direct Human-to-Human task oriented corpus collected at Edinburgh University 4, 12, 14, 20, 24, 27, 31, 33, 36, 42, 44, 59, 63–68, 70,

72, 86, 90, 102, 104, 115, 132, 134, 136, 137, 139, 141, 148, 150, 151, 153–155, 158, 176, 214, 219, 224, 225, 303, 325, 326, 329, 330

HCRC Abbreviation of "HCRC Edinburgh Map Task corpus" that is used in figures and tables of this thesis .... 33, 104, 105, 109, 112–114, 139, 141, 150, 305–308, 327, *see also* HCRC Edinburgh Map Task corpus

ILMT-s2s corpus

Refers to the interlingual machine translation mediated Human-to-Human task oriented corpus collected using the ILMT-s2s System that was collected at Trinity College Dublin, The University of Dublin (refer to § 4 for more details) .... ii, 4, 12, 14, 27, 34, 36–38, 42–45, 49, 59, 61, 63, 68, 76–78, 84, 86, 90, 91, 94, 100, 102, 104, 115, 134, 135, 137–141, 146, 148, 150, 151, 153–155, 158, 159, 192, 214, 219, 220, 222, 224, 225, 227, 326, 329, 330

ILMT-s2s Abbreviation of "ILMT-s2s corpus" that is used in figures and tables of this thesis .... 38, 96, 99, 104, 105, 109, 112–114, 139, 150, 163, 164, 219, 328, *see also* ILMT-s2s corpus

ILMT-s2s System

Refers to Inter Lingual Map Task Speech-to-Speech System or Inter Lingual Machine Translation Speech-to-Speech System 2, 4–6, 8, 12, 14, 15, 20, 22, 24, 26, 39, 43, 46–53, 55–57, 60–62, 70, 72, 73, 78, 79, 82, 84, 86, 90–94, 97, 98, 102–104, 108, 117, 118, 130–133, 142, 158–163, 176–178, 193, 196, 198, 204–207, 209, 211, 212, 214, 215, 219–221, 224–228, 340

L2 Refers to a foreign language (second, third, fourth or fifth foreign language learnt) 18

M Abbreviation of "Mean" (average) that is used in figures and tables of this thesis 28, 29, 33, 38, 39, 70, 96, 99, 101, 104–106, 109, 111–114, 120, 124, 132, 138–141, 146, 150, 155, 163–167, 171, 172, 181, 186, 188, 189, 195, 196, 199–201, 203–208, 210, 220, 249, 279, 290–295, 304–308, 327, 328

Mdn Abbreviation for "Median" that is used in figures and tables of this thesis 28, 29, 33, 38–40, 96, 99, 101, 104–106, 108, 109, 111–114, 119, 120, 124, 138–141, 146, 150, 155, 163–167, 172, 181, 186, 188, 189, 195, 196, 199–201, 203–208, 210, 220, 249, 278, 290–295, 304–308, 327, 328

S2S Abbreviation for Speech-to-Speech that refers to computer interaction that takes voice as the input and also the output 48

S2S-MT Abbreviation for Speech-to-Speech Machine Translation that refers to machine translation that takes voice as the input and also the output .... iii, 1–5, 18, 19, 24–26, 43, 44, 46, 49, 51, 52, 59, 70, 73, 102, 177, 192, 213–215, 217, 219,

220

S2T Abbreviation for Speech-to-Text that refers to computer interaction that takes voice as the input and text as the output 46, 50, 51, 161

S2T-MT Abbreviation for Speech-to-Text Machine Translation that refers to machine translation that takes voice as the input and text as the output 2, 46

SD Abbreviation for "Standard Deviation" that is used in figures and tables of this thesis 28, 29, 33, 38, 39, 96, 99, 101, 104–106, 109–114, 120, 124, 132, 139–141, 146, 150, 163–167, 171, 172, 181, 186, 188, 189, 195, 196, 199–201, 203–208, 220, 249, 290–295, 304–308, 327, 328

T2T-MT Abbreviation for Text-to-Text Machine Translation that refers to machine translation that takes text as the input and also the output 18, 19

UI Abbreviation for User Interface 46, 48, 54–56, 159, 177

Bibliography

Abe, M., Shikano, K., & Kuwabara, H. (1990, Apr). Cross-language voice conversion. In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-90)* (Vol. 1, pp. 345–348). doi: 10.1109/ICASSP.1990.115676

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3), 273–287. doi: 10.1007/s10579-007-9061-5

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*(4), 351-366. doi: 10.1177/002383099103400404

Apple. (2013). Mac OS X Human Interface Guidelines. *Apple Inc, Cupertino*. Retrieved from https://web.archive.org/web/20140603021344/https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/AppleHIGuidelines/OSXHIGuidelines.pdf

Arai, T., & Greenberg, S. (1997). The Temporal Properties of Spoken Japanese are Similar to those of English. In *EUROSPEECH '97 – 5th European Conference on Speech Communication and Technology, September 22–25, Rhodes, Greece, Proceedings* (pp. 1011–1014). Rhodes, Greece: ISCA.

Baranyi, P., & Csapó, Á. (2012). Definition and Synergies of Cognitive Infocommunications. *Acta Polytechnica Hungarica*, *9*(1), 67–83.

Baranyi, P., Csapó, Á., & Várlaki, P. (2014, Jul). An overview of research trends in CogInfoCom. *IEEE 18th International Conference on Intelligent Engineering Systems INES 2014*. doi: 10.1109/ines.2014.6909365

Batliner, A., Fischer, K., Huber, R., Spiker, J., & Nöth, E. (2000). Desperately Seeking Emotions or: Actors, Wizards, and Human Beings. In *ISCA Tutorial and Research Workshop (ITRW) on on Speech and Emotion, September 5–7, Newcastle, Northern Ireland, UK, Proceedings* (pp. 195–200).

Batliner, A., Hacker, C., & Nöth, E. (2006). To talk or not to talk with a computer: On-Talk vs. Off-Talk. In *Workshop on 'How People Talk to Computers, Robots, and Other Artificial Communication*

Partners', April 21—23, Hansewissenschaftskolleg, Delmenhorst, Germany, Proceedings (pp. 79–100). SFB/TR 8 Spatial Cognition Report.

Batliner, A., Hacker, C., & Nöth, E. (2009). To talk or not to talk with a computer. *Journal on Multimodal User Interfaces*, *2*(3), 171–186. doi: 10.1007/s12193-009-0016-6

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., & Niemann, H. (2001). Whence and whither prosody in automatic speech understanding: a case study. In *Proceedings of the Workshop on Prosody and Speech Recognition* (pp. 3–12). ISCA.

Beatty, J. (1982, Mar). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. doi: 10.1037/0033-2909.91.2.276

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. doi: 10.1016/j.jml.2008.06.003

Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A., & Girand, C. (2002). Which Predictability Measures Affect Content Word Durations? In *Proceedings of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology* (pp. 1–5). ISCA.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human-computer interaction. In *Proceedings of ICPhS XV (15^{th} International Congress of Phonetic Sciences)* (pp. 2453–2456). Barcelona, Spain.

Black, A. W., Brown, R., Frederking, R., Singh, R., Moody, J., & Steinbrecher, E. (2002). Tongues: Rapid development of a speech-to-speech translation system. In *Proceedings of the second international conference on human language technology research* (pp. 183–186). San Francisco, California, USA: Morgan Kaufmann Publishers Inc.

Blanchon, H., & Boitet, C. (2000). Speech Translation for French within the C-STAR II Consortium and Future Perspectives. In *Proceedings of ICSLP 2000: the Sixth International Conference on Spoken Language Processing* (pp. 412–417). Beijing, China: ISCA.

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9-10), 341–347.

Borsel, J. V., & Maesschalck, D. D. (2008). Speech rate in males, females, and male-to-female transsexuals. *Clinical Linguistics & Phonetics*, *22*(9), 679–685. (PMID: 18608249) doi: 10.1080/02699200801976695

Botherel, V., & Maffiolo, V. (2006). Regulation of emotional attitudes for a better interaction: Field study in call centres. In *HFT 2006: 20th International Symposium on Human Factors in Telecommunication.* Sophia Antipolis, France.

Branigan, H., Lickley, R., & McKelvie, D. (1999). Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech. In *Proceedings of ICPhS XIV (14th International Congress of Phonetic Sciences)* (pp. 387–390). San Francisco, California, USA.

Branigan, H., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. doi: 10.1177/135910457000100301

Brown, G., Anderson, A., Shillcock, R., & Yule, G. (1985). *Teaching talk: Strategies for production and assessment.* Cambridge University Press.

Bub, T., Wahlster, W., & Waibel, A. (1997, Apr). Verbmobil: the combination of deep and shallow processing for spontaneous speech translation. In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)* (Vol. 1, pp. 71–74). doi: 10.1109/ICASSP.1997.607199

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational linguistics*, *23*(1), 13–31.

Cerrato, L., Hayakawa, A., Campbell, N., & Luz, S. (2016). Future and Emergent Trends in Language Technology: First International Workshop, FETLT 2015, Seville, Spain, November 19-20, 2015, Revised Selected Papers. In F. J. Quesada, F.-J. Martín Mateos, & T. Lopez-Soto (Eds.), (pp. 53–64). Cham: Springer International Publishing. doi: 10.1007/978-3-319-33500-1_5

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC, US: American Psychological Association. doi: 10.1037/10096-006

Clark, H. H., & Marshall, C. K. (1981). Definite knowledge and mutual knowledge. In A. Joshi, B. H. Weber, & I. A. Sag (Eds.), *Elements of discourse understanding.* Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. doi: 10.1016/0010-0277(86)90010-7

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. doi: 10.1037/0033-2909.114.3.494

Davies, B. (2000). Grice's cooperative principle: Getting the meaning across. *Leeds Working Papers in Linguistics and Phonetics*, *8*, 1–26.

Davies, B. (2006). Testing dialogue principles in task-oriented dialogues: An exploration of co-operation, collaboration, effort and risk. *Leeds Working Papers in Linguistics and Phonetics*, *11*, 30–64.

de Jong, N. H., & Wempe, T. (2009, May). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. doi: 10.3758/BRM.41.2.385

Diderichsen, P. (2007, Jan). Givenness revisitedindefinite one-anaphora in unscripted Danish dialogue. *Acta Linguistica Hafniensia*, *39*(1), 179—208. doi: 10.1080/03740463.2007.10414605

Esposito, A., Esposito, A. M., & Vogel, C. (2015). Needs and challenges in human computer interaction for processing social emotional information. *Pattern Recognition Letters*, *66*, 41–51. doi: 10.1016/j.patrec.2015.02.013

Feely, A. J., & Harzing, A. (2003). Language management in multinational companies. *Cross Cultural Management: An International Journal*, *10*(2), 37–52. doi: 10.1108/13527600310797586

Finn, K. E., Sellen, A. J., & Wilbur, S. B. (Eds.). (1997). *Video-Mediated Communication.* Lawrence Erlbaum Associates, Inc.

Flournoy, R. S., & Callison-Burch, C. (2001). Secondary Benefits of Feedback and User Interaction in Machine Translation Tools. In *In Proceedings of MT Summit VIII: the eighth Machine Translation Summit.* Santiago de Compostela, Spain: AAMT.

Frederking, R., Black, A., Brown, R., Moody, J., & Steinbrecher, E. (2002). Field Testing the Tongues Speech-to-Speech Machine Translation System. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)* (pp. 160–164). Paris, France: European Language Resources Association (ELRA).

Frederking, R., Rudnicky, A., Hogan, C., & Lenzo, K. (2000, Jun 01). Interactive Speech Translation in the Diplomat Project. *Machine Translation*, *15*(1), 27–42. doi: 10.1023/A:1011172330853

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971. doi: 10.1145/32206.32212

Giles, H., Mulac, A., Bradac, J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. In M. L.McLaughlin (Ed.), *The first decade and beyond, in communication yearbook 10* (pp. 13–48). Newbury Park: Sage.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 517–520. doi: 10.1109/icassp.1992.225858

Greenberg, S. (1997). *The Switchboard Transcription Project* (Tech. Rep.). Baltimore, Maryland, USA: Johns Hopkins University.

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*(2–4), 159–176. doi: 10.1016/S0167-6393(99)00050-3

Grice, H. P. (1975). The logic of conversation. *Syntax and Semantics*, *3*, 41–58.

Grice, M., & Savino, M. (2003). Map Tasks in Italian: Asking Questions about Given, Accessible and New Information. *Catalan Journal of Linguistics*, *2*(0), 153–180.

Haider, F. (2018). *Improving social intelligence of machines in the context of public speaking situations* (Unpublished doctoral dissertation). Trinity College Dublin.

Haider, F., Hayakawa, A., Luz, S., Vogel, C., & Campbell, N. (2018). On-Talk and Off-Talk Detection: A Discrete Wavelet Transform Analysis of Electroencephalogram. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)* (pp. 960–964). Calgary, Alberta, Canada: IEEE.

Haider, F., Hayakawa, A., Vogel, C., Campbell, N., & Luz, S. (2019). Analysing patterns of right brain-hemisphere activity prior to speech articulation for identification of system-directed speech. *Speech Communication*, *107*, 18–25.

Hara, K., & Iqbal, S. T. (2015). Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)* (pp. 3473–3482). New York, NY, USA: ACM. doi: 10.1145/2702123.2702407

Hayakawa, A., Campbell, N., & Luz, S. (2014). Interlingual Map Task Corpus Collection. In *Proceedings of INTERSPEECH'14: the 15th Annual Conference of the International Speech Communication Association* (pp. 189–191). Singapore: ISCA.

Hayakawa, A., Cerrato, L., Campbell, N., & Luz, S. (2015). A Study of Prosodic Alignment in Interlingual Map-Task Dialogues. In The Scottish Consortium for ICPhS (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, UK: The University of Glasgow. (Paper number 0760.1-9)

Hayakawa, A., Haider, F., Cerrato, L., Campbell, N., & Luz, S. (2015). Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems. In *Proceedings of INTERSPEECH'15: the 16th Annual Conference of the International Speech Communication Association* (pp. 2539–2543). Dresden, Germany: ISCA.

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., & Campbell, N. (2016). Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of Speech Prosody 2016 (SP8)* (pp. 776–780). Boston, Massachusetts, USA: ISCA. doi: 10.21437/SpeechProsody.2016-159

Hayakawa, A., Luz, S., & Campbell, N. (2016). Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of INTERSPEECH'16: the 17th Annual Conference of the International Speech Communication Association* (pp. 1422–1426). San Francisco, California, USA: ISCA. doi: 10.21437/Interspeech.2016-1623

Hayakawa, A., Luz, S., Cerrato, L., & Campbell, N. (2016, May). The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus. In N. C. C. Chair) et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 605–612). Paris, France: European Language Resources Association (ELRA).

Hayakawa, A., Vogel, C., Luz, S., & Campbell, N. (2017a). Perception Changes With and Without the Video Channel: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of CogInfoCom 2017: the 8th IEEE International Conference on Cognitive InfoCommunications* (pp. 401–406). Debrecen, Hungary: IEEE. doi: 10.1109/coginfocom.2017.8268279

Hayakawa, A., Vogel, C., Luz, S., & Campbell, N. (2017b). Speech Rate Comparison when Talking to a System and Talking to a Human: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of INTERSPEECH'17: the 18th Annual Conference of the International Speech Communication Association* (pp. 3286–3290). Stockholm, Sweden: ISCA. doi: 10.21437/Interspeech.2017-1584

Hayakawa, A., Vogel, C., Luz, S., & Campbell, N. (2018, May). Speech Rate Calculations with Short Utterances: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In N. C. C. chair) et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3176–3183). Paris, France: European Language Resources Association (ELRA).

Helgason, P. (2006). SMTC-A Swedish Map Task Corpus. In *Proceedings of Fonetik* (pp. 57–60).

Höge, H. (2002). Project Proposal TC-STAR Make Speech to Speech Translation Real. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)* (pp. 136–141). Paris, France: European Language Resources Association (ELRA).

Horiuchi, Y., Nakano, Y., Koiso, H., Ishizaki, M., Suzuki, H., Okada, M., ... Akira, I. (1999). The Design and Statistical Characterization of the Japanese Map Task Dialogue Corpus. *Journal of Japanese Society for Artificial Intelligence*, *14*(2), 261-272.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26–37. doi: 10.1037/0096-3445.116.1.26

Juel Henrichsen, P., & Allwood, J. (2013). Predicting the Attitude Flow in Dialogue Based on Multi-Modal Speech Cues. In J. Allwood, E. Ahlsén, P. Paggio, & K. Jokinen (Eds.), *Nealt2012: Proceedings of the 4th nordic symposium on multimodal communication* (pp. 47–53). Göteborg Universitet.

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 56–88. doi: 10.1080/14639220210159735

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ, USA: Prentice-Hall.

Kale, E., & Syed, H. R. (2010). Language barriers and the use of interpreters in the public health services. A questionnaire-based survey. *Patient Education and Counseling*, *81*(2), 187–191. doi: 10.1016/j.pec.2010.05.002

Kane, B., & Luz, S. (2006a, Dec). Multidisciplinary Medical Team Meetings: An Analysis of Collaborative Working with Special Attention to Timing and Teleconferencing. *Computer Supported Cooperative Work (CSCW)*, *15*(5–6), 501–535. doi: 10.1007/s10606-006-9035-y

Kane, B., & Luz, S. (2006b). Probing the use and value of video for multi-disciplinary medical teams in teleconference. In *Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS'06)* (pp. 518–523). Salt Lake City, Utah, USA. doi: 10.1109/CBMS.2006.131

Kaptein, M., & Robertson, J. (2012). Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1105–1114). New York, NY, USA: ACM. doi: 10.1145/2207676.2208557

Klem, G., Lüders, H., Jasper, H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*, *52*, 3–6.

Koch, A. S., Forgas, J. P., & Matovic, D. (2013). Can negative mood improve your conversation? Affective influences on conforming to Grice's communication norms. *European Journal of Social Psychology*, *43*(5), 326–334. doi: 10.1002/ejsp.1950

Korczynski, M. (2003). Communities of Coping: Collective Emotional Labour in Service Work. *Organization*, *10*(1), 55–79. doi: 10.1177/1350508403010001479

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual review of psychology*, *62*, 621–647. doi: 10.1146/annurev.psych.093008.131123

Lavie, A., Metze, F., & Pianesi, F. (2002). Enhancing the Usability and Performance of NESPOLE!: A Real-world Speech-to-speech Translation System. In *Proceedings of the Second International Conference on Human Language Technology Research* (pp. 269–274). San Francisco, California, USA: Morgan Kaufmann Publishers Inc.

Lavie, A., Pianesi, F., & Levin, L. (2006, Sept). The NESPOLE! System for multilingual speech communication over the Internet. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1664–1673. doi: 10.1109/TSA.2005.858520

Levin, L., Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T., & Woszczyna, M. (2000). Evaluation of a Practical Interlingua for Task-oriented Dialogue. In *Proceedings of the 2000 NAACL-ANLP Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP* (Vol. 2, pp. 18–23). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1117554.1117557

Levin, L., Gates, D., Lavie, A., & Waibel, A. (1998). An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. In *Proceedings of ICSLP 1998: the Fifth International Conference on Spoken Language Processing.* Sydney, Australia: ISCA.

Levin, L., Gates, D., Wallace, D., Peterson, K., Lavie, A., Pianesi, F., ... Mana, N. (2002). Balancing Expressiveness and Simplicity in an Interlingua for Task Based Dialogue. In *Proceedings of the ACL-02 Workshop on Speech-to-speech Translation: Algorithms and Systems* (Vol. 7, pp. 53–60). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1118656.1118664

Levin, L., Lavie, A., Woszczyna, M., Gates, D., Gavaldá, M., Koll, D., & Waibel, A. (2000). The Janus-III Translation System: Speech-to-Speech Translation in Multiple Domains. *Machine Translation*, *15*(1), 3–25. doi: 10.1023/A:1011186420821

Lewis, J. (1993). *IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use* (Technical Report No. 54.786). Boca Raton, FL, USA: IBM Corporation.

Lewis, W. D. (2015). Skype translator: Breaking down language and hearing barriers. *Translating and the Computer (TC37).*

Lickley, R. J., Schepman, A., & Ladd, D. R. (2005). Alignment of "Phrase Accent" Lows in Dutch Falling Rising Questions: Theoretical and Methodological Implications. *Language and Speech*, *48*(2), 157-183. (PMID: 16411503) doi: 10.1177/00238309050480020201

Lipartito, K. (2003). Picturephone and the Information Age: The Social Meaning of Failure. *Technology and Culture*, *44*(1), 50–81.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Matsuda, S., Hu, X., Shiga, Y., Kashioka, H., Hori, C., Yasuda, K., ... Nakamura, S. (2013, June). Multilingual Speech-to-Speech Translation System: VoiceTra. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management* (Vol. 2, pp. 229–233). Milan, Italy: IEEE. doi: 10.1109/MDM.2013.99

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., ... others (2005). The AMI meeting corpus. In *Measuring Behavior 2005 — the 5th International Conference on Methods and Techniques in Behavioral Research, August 30 – September 2, Wageningen, The Netherlands, Proceedings* (pp. 137–140).

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., & Bourlard, H. (2005). *On the Use of Information Retrieval Measures for Speech Recognition Evaluation* (Tech. Rep.). Martigny, Switzerland: IDIAP Research Institute.

Morikawa, O., & Maesako, T. (1998). HyperMirror: Toward Pleasant-to-use Video Mediated Communication System. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (pp. 149–158). New York, NY, USA: ACM. doi: 10.1145/289444.289489

Morita, N. (2004). Negotiating Participation and Identity in Second Language Academic Communities. *TESOL Quarterly*, *38*(4), 573–603. doi: 10.2307/3588281

Müller, M., Fünfer, S., Stüker, S., & Waibel, A. (2016, May). Evaluation of the KIT Lecture Translation System. In N. C. C. Chair) et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1856–1861). Paris, France: European Language Resources Association (ELRA).

Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., ... Yamamoto, S. (2006, March). The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 365–376. doi: 10.1109/TSA.2005.860774

Newlands, A., Anderson, A. H., & Mullin, J. (2003). Adapting communicative strategies to computer-mediated communication: an analysis of task performance and dialogue structure. *Applied Cognitive Psychology*, *17*(3), 325–348. doi: 10.1002/acp.868

Niehues, J., Nguyen, T. S., Cho, E., Ha, T.-L., Kilgour, K., Müller, M., ... Waibel, A. (2016, Sep). Dynamic Transcription for Low-Latency Speech Translation. In *Proceedings of INTERSPEECH'16: the 17th Annual Conference of the International Speech Communication Association* (pp. 2513–2517). San Francisco, California, USA: ISCA. doi: 10.21437/interspeech.2016-154

Norman, D. A. (2002). *The design of everyday things.* Basic books.

Nöth, E., Batliner, A., Kießling, A., Kompe, R., & Niemann, H. (2000, Sep). Verbmobil: the use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio Processing*, *8*(5), 519–532. doi: 10.1109/89.861370

O'Conaill, B., Whittaker, S., & Wilbur, S. (1993, Dec). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human–Computer Interaction*, *8*(4), 389—428. doi: 10.1207/s15327051hci0804_4

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., & Bruce, V. (1996a). Comparison of face-to-face and video-mediated interaction. *Interacting with computers*, *8*(2), 177–192.

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., & Bruce, V. (1996b). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, *8*(2), 177 – 192. doi: 10.1016/0953-5438(96)01027-2

Oppermann, D., Schiel, F., Steininger, S., & Beringer, N. (2001). Off-Talk — a Problem for Human-Machine-Interaction? In *Proceedings of EUROSPEECH 2001 Scandinavia: the 7th European Conference on Speech Communication and Technologyand the 2nd INTERSPEECH Event* (pp. 2197–2200). Aalborg, Denmark: ISCA.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., ... Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. doi: 10.1093/cercor/bht355

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech & Language*, *9*(1), 19–35. doi: 10.1006/csla.1995.0002

Oviatt, S. (1997, March). Multimodal Interactive Maps: Designing for Human Performance. *Human-Computer Interaction*, *12*(1), 93–129. doi: 10.1207/s15327051hci1201\&2_4

Oviatt, S. (1999). Mutual Disambiguation of Recognition Errors in a Multimodel Architecture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 576–583). New York, NY, USA: ACM. doi: 10.1145/302979.303163

Oviatt, S., Bernard, J., & Levow, G.-A. (1998). Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech*, *41*(3-4), 419–442. doi: 10.1177/002383099804100409

Oviatt, S., Coulston, R., & Lunsford, R. (2004). When Do We Interact Multimodally?: Cognitive Load and Multimodal Communication Patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 129–136). New York, NY, USA: ACM. doi: 10.1145/1027933.1027957

Oviatt, S., Darves, C., & Coulston, R. (2004). Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *11*(3), 300–328.

Oviatt, S., DeAngeli, A., & Kuhn, K. (1997). Integration and Synchronization of Input Modes During Multimodal Human-Computer Interaction. In *Referring Phenomena in a Multimedia Context and*

Their Computational Treatment (pp. 1–13). Stroudsburg, PA, USA: Association for Computational Linguistics.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003, Mar). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71. doi: 10.1207/s15326985ep3801_8

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 141–144). New York, NY, USA: ACM. doi: 10.1145/1743666.1743701

Pearson, J., Hu, J., Branigan, H., Pickering, M. J., & Nass, C. I. (2006). Adaptive language behavior in hci: How expectations and beliefs about a system affect users' word choice. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1177–1180). New York, NY, USA: ACM. doi: 10.1145/1124772.1124948

Picard, R. W. (1995, November). *Affective computing* (Technical Report No. 321). Cambridge, Massachusetts, USA: M.I.T Media Laboratory Perceptual Computing Section.

Picard, R. W., Vyzas, E., & Healey, J. (2001, Oct). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. doi: 10.1109/34.954607

Pickering, M. J., & Garrod, S. (2004, 4). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190. doi: 10.1017/S0140525X04000056

Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions. *ISSCO WorkingPaper 62*.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, *61*(1), 5–18. (Psychophiosology and Cognitive Neuroscience) doi: http://dx.doi.org/10.1016/j.ijpsycho.2005.10.024

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Stanford, California, USA: CSLI Publications/Cambridge University Press.

Reitter, D., & Moore, J. D. (2007). Predicting Success in Dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 808–815). Prague, Czech Republic: Association for Computational Linguistics.

Schmidt, T., Hedeland, H., Lehmberg, T., & Wörner, K. (2010). Hamtac – the hamburg maptask corpus. *Retrieved from www. exmaralda. org/files/HAMATAC.pdf (last accessed May 2020)*.

Schneider, A., & Luz, S. (2011). Speaker Alignment in Synthesised, Machine Translated Communication. In *Proceedings of the 2011 International Workshop on Spoken Language Translation, IWSLT 2011* (pp. 254–260). San Francisco, California, USA.

Shigenobu, T. (2007). Evaluation and Usability of Back Translation for Intercultural Communication. In N. Aykin (Ed.), *Proceedings of UI-HCII 2007 - Second International Conference on Usability and Internationalization: Held as Part of HCI International 2007, the 12th International Conference on Human-Computer Interaction, Beijing, China, July 22-27* (pp. 259–265). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-73289-1_31

Sjölander, K., & Beskow, J. (2000). Wavesurfer — an open source speech tool. In *Proceedings of ICSLP 2000: the Sixth International Conference on Spoken Language Processing* (Vol. 4, pp. 464–467). Beijing, China: ISCA.

Stallard, D., Choi, F., Krstovski, K., Natarajan, P., Prasad, R., Saleem, S., & Suleiman, R. (2006, Dec). Design and evaluation of the 2006 BBN English/Iraqi two-way speech translation system. In *2006 IEEE Spoken Language Technology Workshop* (pp. 254–257). doi: 10.1109/SLT.2006.326803

Stent, A. J., Huffman, M. K., & Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication*, *50*(3), 163–178. doi: 10.1016/j.specom.2007.07.005

Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science*, *19*(2), 131-141. doi: 10.1080/09540090701369125

Sztahó, D., Kiss, G., & Vicsi, K. (2015). Estimating the Severity of Parkinson's Disease from Speech Using Linear Regression and Database Partitioning. In *Proceedings of INTERSPEECH'15: the 16th Annual Conference of the International Speech Communication Association* (pp. 498–502). Dresden, Germany: ISCA. Retrieved from http://www.isca-speech.org/archive/interspeech_2015/i15_0498.html

Takahashi, K. (2004, Dec). Remarks on emotion recognition from multi-modal bio-potential signals. In *IEEE International Conference on Industrial Technology (IEEE ICIT '04)* (Vol. 3, pp. 1138–1143). doi: 10.1109/ICIT.2004.1490720

Trancoso, I., Viana, M. C., Duarte, I., & Matos, G. (1998). Corpus de diálogo CORAL. In *Proceedings of PROPOR'98.*

van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999, Nov). Semantic Integration in Sentences and Discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671. doi: 10.1162/089892999563724

Veinott, E. S., Olson, J., Olson, G. M., & Fu, X. (1999). Video Helps Remote Work: Speakers Who Need to Negotiate Commono Ground Benefit from Seeing Each Other. *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. doi: 10.1145/302979.303067

Walton, D., & Krabbe, E. C. (1995). Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. *State University of New York Press, Albany*.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 1556–1559). Genoa, Italy: European Language Resources Association (ELRA).

Yamashita, N., & Ishida, T. (2006a). Automatic Prediction of Misconceptions in Multilingual Computer-mediated Communication. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (pp. 62–69). New York, NY, USA: ACM. doi: 10.1145/1111449.1111469

Yamashita, N., & Ishida, T. (2006b). Effects of Machine Translation on Collaborative Work. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (pp. 515–523). New York, NY, USA: ACM. doi: 10.1145/1180875.1180955

Yuan, J., & Jurafsky, D. (2005, Nov). Detection of questions in Chinese conversational speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* (pp. 47–52). doi: 10.1109/ASRU.2005.1566536

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an Integrated Understanding of Speaking Rate in Conversation. In *Proceedings of INTERSPEECH'06: the 7th Annual Conference of the International Speech Communication Association* (pp. 2197–2200). Pittsburgh, PA, USA: ISCA.

Zipf, G. K. (1945). The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, *33*(2), 251–256. doi: 10.1080/00221309.1945.10544509