Exploring the Effect of Cultural Background and Interaction Language on Preferences for Repair Strategies in Spoken Dialogue Systems

Essam Saleh Alghamdi

Computer and Information Sciences
University of Strathclyde

A thesis submitted for the degree of
Doctor of Philosophy

Glasgow 2025

Declaration of Authenticity and Author's Rights

Signed: Essam

Date: 10/02/2026

Acknowledgment

I would like to begin by thanking Allah (the Most Gracious and the Most Merciful) for granting me strength, patience, and guidance throughout my life and my studies.

I extend my sincere thanks and appreciation to my supervisors, Professor Martin Halvey and Dr Emma Nicol, for their continuous guidance, encouragement, and support throughout my PhD journey. Their academic insight and thoughtful feedback were instrumental in shaping my research, refining its direction, and guiding me through the publication process. I am particularly grateful for their understanding, flexibility, and support over the years, as well as for their availability and encouragement through regular discussions and meetings.

I would also like to thank my viva examiners, Professor Benjamin Cowan and Dr Mark Dunlop, for their valuable feedback and constructive comments, which helped improve the quality and clarity of my thesis. My thanks also go to the CIS Department and the IB Group for providing a supportive and stimulating research environment. In particular, I would like to thank Professor Ian Ruthven. I am grateful for the regular group meetings and opportunities to exchange ideas and present my work.

I would like to thank my mother (Safiah) for her patience, belief in me, and constant emotional support throughout my educational journey. Her encouragement and love have been a source of strength. I also dedicate this work to the memory of my father (Saleh), who encouraged me to value education and learning from an early age; I believe he would have been proud of this achievement. I would also like to thank my brothers and sisters for their continued support, and in particular my youngest sister (Amal) whose intelligence and enthusiasm for technology have always inspired me.

I would like to express my gratitude to my wife, Dr Nouf Alatifi, for her unwavering support, patience, and partnership throughout this long journey. She has been by my side since the beginning of my pursuit of postgraduate education, from my master's studies to the completion of this doctorate, and her encouragement has been invaluable, especially during the final year. I thank my lovely four-year-old daughter (Deema) whose patience, presence, and joyful smile have been a constant source of motivation and strength. Her smile, often resting on my desk, gave me the strength to take one more step and keep moving forward, even during difficult moments of this journey.

I would also like to thank all my friends for their encouragement and support throughout this journey. In particular, I would like to thank my friend (Sami Althmali) for his encouragement during vulnerable moments, which helped me continue my education. I am also grateful to all participants who took part in the studies presented in this thesis for their time and valuable contributions. My thanks also extend to my former teachers whose early encouragement shaped my academic path, particularly my undergraduate C++ instructor (Tayyab) and my English teachers.

I would also like to acknowledge the financial and institutional support that made this PhD possible. I am grateful to the Ministry of Higher Education in KSA and to Prince Sattam bin Abdulaziz University for sponsoring my studies and providing the opportunity to pursue this research. I would also like to thank my daughter's nursery centre in the UK for providing a safe and supportive environment during my studies.

Contents

I Introduction, Background, and Systematic Scoping Review **1**

1 Introduction **2**

1.1 Motivation . . . 2
1.2 Context . . . 3
1.3 Research Questions . . . 3
1.4 Contributions . . . 4
1.5 Thesis Outline . . . 5

2 Relevant Background **10**

2.1 Spoken Dialogue Systems (SDSs) . . . 10
2.1.1 Components of Spoken Dialogue Systems . . . 11
2.2 Conversational Breakdowns . . . 13
2.3 Conversational Repair Strategies . . . 16
2.3.1 System Repair Strategies . . . 17
2.3.2 User Repair Strategies . . . 19
2.4 Communication Styles in HHI and HCI . . . 20
2.4.1 Communication in Human-Human Interaction (HHI) . . . 20
2.4.1.1 Hall's theory of low-context and high-context cultures . . . 21
2.4.1.2 Hofstede's cultural dimensions . . . 22
2.4.2 Communication in Human-Computer Interaction (HCI) . . . 25
2.4.2.1 Computers as Social Actors (CASA) . . . 25
2.4.2.2 Cultural Background . . . 26
2.5 Chapter Summary . . . 27

3 System and User Repair Strategies: A Systematic Scoping Review **28**

3.1 Importance . . . 28
3.2 Method . . . 30

3.2.1 Identifying relevant studies . . . 31
3.2.1.1 Databases . . . 31
3.2.1.2 Search Query . . . 31
3.2.2 Study selection . . . 33
3.2.2.1 Inclusion criteria . . . 33
3.2.2.2 Exclusion criteria . . . 33
3.3 Results . . . 35
3.3.1 The characteristics of included studies and SDSs . . . 35
3.3.1.1 Publication Overview . . . 35
3.3.1.2 Methodological Overview . . . 37
3.3.1.3 SDS Characteristics Overview . . . 38
3.3.2 System-Repair Strategies . . . 42
3.3.2.1 Confirmation . . . 42
3.3.2.2 Information . . . 43
3.3.2.3 Social . . . 43
3.3.2.4 Solve . . . 43
3.3.2.5 Ask . . . 44
3.3.2.6 Disclosure . . . 44
3.3.3 User-Repair Strategies . . . 44
3.3.3.1 Clarification . . . 46
3.3.3.2 Information Adjustment . . . 46
3.3.3.3 Modulation . . . 47
3.3.3.4 System-Centric . . . 47
3.3.3.5 Contextual and Affective Responses . . . 47
3.4 Discussion and Future Directions . . . 48
3.4.1 System and User Repair Strategies in SDSs . . . 48
3.4.1.1 Frameworks of repair strategies . . . 48
3.4.1.2 Interaction of Repair Strategies . . . 50
3.4.2 Research Methods and Defining Characteristics of SDSs . . . 51
3.4.3 Cultural Background and Interaction Language . . . 52
3.5 Conclusion . . . 53
3.6 Chapter Summary . . . 53

II Two Empirical Studies on User-Centred Evaluation of Repair Strategies in Spoken Dialogue Systems 55

4 Methodology for User-Centred Empirical Studies 56

4.1 Methodological Rationale 56
4.2 Repair Strategy Framework 58
4.3 Research Design Overview 60
4.3.1 Goal-Oriented Task 61
4.3.2 Simulated Work Task Scenario 62
4.3.3 Type of Communication Breakdown 63
4.3.4 Participant Overview 64
4.4 Consistent Experimental Design Across Studies (RQ2 & RQ3a) 66
4.5 Measurement of Individual-Level Factors (RQ3b) 70
4.5.1 Computer Self-efficacy 70
4.5.2 Prior Experience with Technology and SDSs 71
4.5.3 Language Proficiency 72
4.6 Qualitative Interview Design (RQ4) 73
4.7 Data Processing and Analysis 75
4.7.1 Preference Score Calculation 75
4.8 Preparatory Procedures 76
4.8.1 Pilot Study 76
4.8.1.1 Aim 76
4.8.1.2 Method 77
4.8.1.3 Results 79
4.8.2 Translation Process 79
4.9 Chapter Summary 80

5 Study 1: Effect of Cultural Background on Repair Strategy Preferences 81
5.1 Introduction 81
5.2 Method 82
5.2.1 Repair Strategies Framework 83
5.2.2 Participants 83
5.2.3 Simulated Task Scenario 84
5.2.4 Paired Comparison Experiment 85
5.2.5 Study Design 86
5.3 Results 87
5.3.1 Overall Descriptive Statistics 87
5.3.2 Within-Group Comparisons 88
5.3.2.1 Overall Within-Group Differences: Friedman Test 88
5.3.2.2 Pairwise Preference Comparisons: Wilcoxon Signed-Ranks Test . . . 89
5.3.3 Between-Group Comparisons: Mann–Whitney Tests 90

5.4 Discussion . . . 91
5.4.1 Overview of Key Findings . . . 92
5.4.2 Cultural Background and Preference for Information Repair Strategies (RQ2) 92
5.4.3 Understanding the Absence of Cultural Differences in Other Repair Strategies 93
5.5 Conclusion . . . 95
5.6 Chapter Summary . . . 95

6 Study 2: Effect of Interaction Language on Repair Strategy Preferences 96
6.1 Introduction . . . 96
6.2 Method . . . 98
6.2.1 Repair Strategies Framework . . . 98
6.2.2 Participants . . . 99
6.2.3 Simulated Task Scenario . . . 100
6.2.4 Paired Comparison Experiment . . . 101
6.2.5 Individual-level Factors . . . 102
6.2.6 Qualitative Interview Design . . . 103
6.3 Results . . . 104
6.3.1 Effect of Interaction Language (RQ3a) . . . 105
6.3.1.1 Overall Descriptive Statistics . . . 105
6.3.1.2 Overall Within-Group Differences: Friedman Test . . . 105
6.3.1.3 Within-Groups Pairwise Comparisons: Wilcoxon Signed-Ranks Test 107
6.3.1.4 Between-Group Comparisons: Mann–Whitney Tests . . . 107
6.3.2 Effect of Individual Factors (RQ3b) . . . 109
6.3.2.1 Experience with Technology and DVAs . . . 110
6.3.2.2 Computer Self-Efficacy . . . 110
6.3.2.3 Language Proficiency . . . 110
6.3.2.4 Gender and Age . . . 110
6.3.3 Users' perspectives of Repair Strategies (RQ4) . . . 111
6.3.3.1 Theme 1: User Experiences and Adaptation with DVAs . . . 111
6.3.3.2 Theme 2: Repair Strategy Preferences and Perspectives . . . 114
6.3.3.3 Theme 3: DVA Adaptation to Language Proficiency and Speech Input 116
6.4 Discussion . . . 118
6.4.1 Repair Strategy Preferences (RQ3a & RQ4) . . . 118
6.4.2 Interaction Language Effect (RQ3a) . . . 120
6.4.3 Individual-level Factors (RQ3b) . . . 120
6.5 Conclusion . . . 122
6.6 Chapter Summary . . . 123

III Integrated Discussion and Conclusions **124**

7 Integrated Analysis and General Discussion **125**

7.1 Integrated Analysis of Three Participant Groups . . . 125

7.2 Theoretical Implications in Cross-Cultural and Linguistic Contexts . . . 128

7.2.1 Strategies with Significant Between-Group Differences . . . 128

7.2.1.1 Cultural Communication Style . . . 129

7.2.1.2 Cognitive Load in L2 Interaction . . . 130

7.2.2 Strategies Without Statistically Significant Between-Group Differences . . . 133

7.2.2.1 Ask . . . 133

7.2.2.2 Social . . . 134

7.2.2.3 Solve . . . 134

7.3 Practical Design Recommendations . . . 135

7.3.1 Suitability for First-Attempt Repair . . . 136

7.3.2 User Background Consideration . . . 137

7.3.3 System's Confidence Threshold Considerations . . . 137

7.4 Summary of Findings . . . 138

7.5 Chapter Summary . . . 139

8 Conclusions **141**

8.1 Conclusions . . . 141

8.2 Summary of Contributions . . . 143

8.3 Limitations of the Thesis . . . 144

8.4 Future Research Directions . . . 145

8.5 Closing Remarks . . . 147

Bibliography **148**

A Participant Information Sheets and Consent Forms for the Pilot Study (Ethical ID: 2769) **173**

A.1 Participant Information Sheet . . . 173

A.2 Questionnaire Consent Form . . . 180

A.3 Questionnaires . . . 182

A.3.1 Demographic Questionnaire . . . 182

A.3.2 Verifying the Distinctiveness of Repair Strategies . . . 184

A.3.3 Evaluating the Realism of the Task Scenario . . . 186

B Participant Information Sheets and Consent Forms for the Empirical Studies (Ethical ID: 2781) **188**
B.1 Participant Information Sheet 188
B.2 Questionnaires Consent Form 196
B.3 Interview Consent Form 199
B.4 Questionnaires and Interview 200
B.4.1 Demographic Questionnaire 200
B.4.2 Prior Experience with Technology and DVAs 202
B.4.3 Computer Self-Efficacy 205
B.4.4 Pairwise Comparison Experiment on Preferences 208
B.4.5 Self-Assessment of English Language Proficiency 209
B.4.6 Interview Questions 211

C Supplementary Statistical Analyses **214**
C.1 Group Comparability Analyses (t-tests) 214
C.2 Binary logistic regression predicting preferences for the five repair strategies 215
C.2.1 Social strategy: 215
C.2.2 Confirmation strategy: 215
C.2.3 Information strategy: 216
C.2.4 Solve strategy: 216
C.2.5 Ask strategy: 216

List of Figures

1.1 Overview of the thesis structure and research questions 6

2.1 Basic spoken dialogue systems architecture 12
2.2 Comparison of Misunderstanding and Non-understanding Scenarios in SDSs 15
2.3 Communicating scale of low- and high-context cultures 21
2.4 Comparison of Hofstede's cultural dimension scores for Saudi Arabia and the UK. . 24

3.1 PRISMA-ScR flowchart for systematic review process. 34
3.2 Distribution of relevant literature according to year 35
3.3 Heat map of methodology details for the selected studies 38
3.4 System repair strategies: flowchart of SDS breakdown and framework. 45
3.5 User repair strategies: flowchart of SDS breakdown and framework. 49

4.1 Complexity of SDS applications and flight reservation task. 62
4.2 Participant groups: BL1, SL1, and SL2 in cultural and language studies. 65
4.3 Text-based scenario of flight booking task with repair strategies. 69
4.4 Geometric representation of paired comparisons among repair strategies 70
4.5 Computer Self-Efficacy Constructs 71
4.6 Task scenario and predefined categories of the five repair strategies used in the pilot study. 78

5.1 Overview of the experimental procedure completed by participants in Study 1. . . . 87
5.2 Friedman Mean Rank preferences scores for repair strategies: BL1 vs. SL1. 89
5.3 Mann–Whitney U preference scores for repair strategies: BL1 vs. SL1. 91

6.1 Overview of the survey-based and experimental procedures completed by participants in Study 2 101
6.2 Friedman Mean Rank preferences scores for repair strategies: SL1 vs. SL2. 106
6.3 Mann–Whitney U preference scores for repair strategies: SL1 vs. SL2. 109

7.1 Overall ranking of repair strategy preferences across participants and groups. 126
7.2 Kruskal–Wallis preference scores for repair strategies across all groups: BL1, SL1, and SL2. 127

List of Tables

3.1 General Query Structure for All Searches 32
3.2 Publication Venues 36
3.3 A Literature Matrix of Repair Strategies in SDSs 41

4.1 Descriptions and Attributes of Repair Strategies 59
4.2 Adapted Items for Measuring Computer Self-Efficacy and Experiences with Technology and DVAs 72
4.3 Common European Framework of Reference for Languages 73
4.4 Example of Preference Score Calculation for a Single Participant 76
4.5 Overview of Analytical Methods and Their Purpose Across Studies 76

5.1 Mean preference scores for repair strategies (BL1 vs. SL1). 88
5.2 Adjusted significance levels from Wilcoxon signed-rank tests (BL1 vs. SL1) 90

6.1 Participant demographics and factors by interaction language. 100
6.2 Mean preference scores for repair strategies (SL1 vs. SL2) 105
6.3 Adjusted significance levels from Wilcoxon signed-rank tests (SL1 vs. SL2) 108
6.4 Summary of themes and sub-themes from thematic analysis 111

7.1 Design recommendations for the use of different repair strategies in SDSs. 136

List of Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
BL1	**B**ritish participants interacting in their native language, English (**L1**)
CAs	**C**onversational **A**gents
CASA	**C**omputers **A**s **S**ocial **A**ctors
CEFR	**C**ommon **E**uropean **F**ramework of **R**eference for Languages
DVAs	**D**igital **V**oice **A**ssistants
HCI	**H**uman–**C**omputer **I**nteraction
HHI	**H**uman–**H**uman **I**nteraction
ITC	**I**nternational **T**est **C**ommission
L1	Native (First) **L**anguage
L2	Non-native (Second) **L**anguage
LLMs	**L**arge **L**anguage **M**odels
NLP	**N**atural **L**anguage **P**rocessing
SDSs	**S**poken **D**ialogue **S**ystems
SL1	**S**audi participants interacting in their native language, Arabic (**L1**)
SL2	**S**audi participants interacting in their a non-native language, English (**L2**)
SLU	**S**poken **L**anguage **U**nderstanding
SWTS	**S**imulated **W**ork **T**ask **S**cenario

Glossary

Ask Repair Strategy	The system directly requests the user to repeat or clarify their input without providing any additional information, explanation, or options.
Collectivism	A cultural orientation emphasizing group harmony, shared responsibilities, and interdependence [115].
Communication Breakdown	A failure in understanding between a user and the system, often requiring repair strategies to restore mutual understanding.
Computers as Social Actors (CASA)	A paradigm proposing that people apply social rules, norms, and expectations to computers and digital systems, treating them as if they were human communicators [214].
Confirmation Repair Strategy	The system attempts to confirm its understanding.
Digital Voice Assistant (DVA)	A type of SDS that operates on consumer devices (e.g., Amazon Alexa, Google Assistant, Apple Siri). DVAs perform different tasks through voice commands such as booking flights, setting reminders, or providing information.
High-Context Culture	Cultures where communication relies heavily on implicit cues, shared background, and situational context [103].
Individualism	A cultural orientation in which individuals prioritize independence, self-expression, and personal goals [115].
Information Repair Strategy	The system provides an explanation of the cause of the breakdown.
Low-Context Culture	Cultures where communication is direct, explicit, and less dependent on shared context [103].
Pairwise Comparison	A research method in which participants are presented with two options at a time and asked to choose their preference. This approach allows for fine-grained comparison between alternatives.

Repair Strategy	A conversational mechanism used by the system or user to recover from communication breakdowns.
Social Repair Strategy	The system involves using human-like and socially oriented expressions, such as polite apologies, to acknowledge breakdowns.
Solve Repair Strategy	The system offers multiple options or solutions to help the user recover from the breakdown.
Spoken Dialogue System (SDS)	A computer system designed to interact with users through natural language in spoken form, capable of interpreting speech input and generating spoken responses.
Uncertainty Avoidance	The extent to which members of a culture feel uncomfortable with ambiguity or uncertainty and rely on rules, traditions, or predictability [115].

Abstract

Spoken dialogue systems (SDSs) are increasingly integrated into everyday life, appearing in forms such as embodied robots, in-car systems, and, most commonly, digital voice assistants (DVAs) like Amazon Alexa, Apple's Siri, and Google Assistant. Despite their widespread use, conversational breakdowns remain inevitable and problematic. Repair strategies, defined as system responses designed to recover from such breakdowns, play a crucial role in shaping user experience. While prior research has focused on technical solutions and the development of effective repair strategies, limited attention has been paid to how user characteristics influence preferences and perspectives on communication repair strategies in voice-based interactions.

This thesis addresses this gap by examining how user characteristics shape preferences and perspectives on repair strategies in goal-oriented tasks. Specifically, it considers (a) cultural background (high- vs. low-context cultures), (b) interaction language (native vs. non-native), and (c) individual-level factors. To achieve this, the thesis begins with a systematic scoping review, followed by two user-centered empirical studies. The overarching goal is to provide a deeper understanding of how user characteristics influence repair strategy preferences and perspectives, thereby informing the design of adaptive SDSs that deliver improved user experiences.

The scoping review synthesized the literature on repair strategies in SDSs and produced two comprehensive frameworks: one categorizing system-initiated repair strategies and another addressing user-initiated repair strategies. Building on these foundations, the first empirical study examined how cultural background influences preferences by comparing users from the United Kingdom and Saudi Arabia. The second empirical study examined the influence of interaction language (native vs. non-native) alongside individual-level factors, including prior experience and computer self-efficacy, on repair strategy preferences. Complementing these quantitative studies, a qualitative investigation using semi-structured interviews explored user perspectives on communication breakdowns and repair strategies when interacting with SDSs.

Overall, the findings reveal cultural differences, particularly in preferences for elaborative and explanation-based repair strategies. However, the effects of interaction language and individual-level factors often outweighed cultural influences. The qualitative insights provided deeper explanations of these preferences, highlighting how contextual factors and task demands shaped user perspectives on breakdowns and repair. Collectively, this thesis extends the theoretical understanding of repair strategies in SDSs and offers practical recommendations for developing adaptive, user-centered systems that are sensitive to cultural, linguistic, and individual diversity.

Part I

Introduction, Background, and Systematic Scoping Review

Chapter 1

Introduction

1.1 Motivation

Spoken Dialogue Systems (SDSs) play a critical role in enabling natural and efficient human-machine interaction through speech. These systems come in various forms, including embodied robots, in-car systems, and most commonly digital voice assistants (DVAs) such as Amazon Alexa, Apple's Siri, and Google Assistant. Due to their intuitive, hands-free nature, SDSs have become a widely adopted mode of interaction with digital devices in everyday settings. However, despite their growing popularity, SDSs often struggle to manage complex or ambiguous dialogues, frequently leading to communication breakdowns. In response to these challenges, research has increasingly focused on developing effective communication repair strategies, which are essential for improving system performance and enhancing user satisfaction and trust.

As SDSs continue to be deployed globally and support multiple languages, they serve a highly diverse user base. Despite their growing use, communication breakdowns remain common, and relatively little is known about how users from different cultural and linguistic backgrounds prefer SDSs to handle such breakdowns. This highlights a gap in our understanding of how repair strategies should be adapted to accommodate user diversity.

The aim of this thesis is to investigate how user characteristics, including (a) cultural background (high- vs. low-context cultures), (b) interaction language (native vs. non-native), and (c) individual-level factors, influence user preferences and perspectives toward communication repair strategies in SDSs. Specifically, this PhD thesis seeks to answer the overarching question: **How do user characteristics influence user preferences and perspectives on communication repair strategies in SDSs during conversational breakdowns?**

1.2 Context

In this thesis, SDSs are examined within the context of goal-oriented task scenarios, where users interact with DVAs such as Amazon Alexa or Apple's Siri to accomplish specific objectives such as booking a flight. DVAs are the most widely used and familiar form of SDSs, making them a practical and relevant focus for studying real-world interaction patterns. While SDSs can support multiple interaction modalities, this thesis focuses exclusively on voice-based interaction, with no visual feedback, to isolate the effects of spoken communication. In this context, the term user characteristics refers to a combination of sociocultural factors, including cultural background and interaction language, and individual-level factors, such as demographics, computer self-efficacy, and prior experience with technology and DVAs. The aim of this PhD thesis is to investigate how these user characteristics influence user preferences and perspectives on communication repair strategies during interactions with DVAs.

1.3 Research Questions

In order to provide the answer to the overarching question: **How do user characteristics influence user preferences and perspectives on communication repair strategies in SDSs during conversational breakdowns?**", this thesis investigates the following research sub-questions concerning user preferences and perspectives in repair strategies of SDSs.

1. **RQ1:** What are the state-of-the-art strategies of conversational breakdown when interacting with SDSs? (Systematically reviewed in Chapter 3)
   - RQ1**a**: What types of user-facing strategies are employed by SDSs to repair breakdowns?
   - RQ1**b**: What strategies or tactics do users apply to repair breakdowns in SDSs?
   - RQ1**c**: What research methods have been utilised to investigate and evaluate repair strategies in SDSs, and what are the defining characteristics of SDSs in this context?
2. **RQ2:** How does users' cultural background influence their preference for communication repair strategies in DVAs? (Quantitatively explored in Chapter 5)
   - RQ2**a**: Which repair strategies are preferred in low-context cultures when a conversational breakdown occurs with DVAs?
   - RQ2**b**: Which repair strategies are preferred in high-context cultures when a conversational breakdown occurs with DVAs?
3. **RQ3:** What is the influence of interaction language and individual-level user characteristics on user preferences for repair strategies in DVAs? (Quantitatively explored in Chapter 6)

- RQ3**a**: Which repair strategies are preferred by users when interacting with a DVA in their native versus non-native language during goal-oriented tasks?
- RQ3**b**: How do individual-level factors, such as (**a**) computer self-efficacy, (**b**) prior experience with technology and DVAs, and (**c**) language proficiency, influence user preferences for repair strategies?

4. **RQ4:** What are the perspectives of users regarding different repair strategies and breakdowns in DVAs? (Qualitatively explored in Chapter 6)

1.4 Contributions

The principal contributions of this PhD thesis are based on a systematic scoping review and two empirical studies. This research demonstrates how users' sociocultural background and individual-level factors influence their preferences and perspectives regarding communication repair strategies when interacting with SDSs. The key contributions of this thesis are summarized below.

This thesis:

1. Proposes two comprehensive frameworks for system-initiated and user-initiated repair strategies in SDSs through a systematic scoping review: an adapted system-initiated repair framework grounded in prior literature and a novel user-initiated repair framework developed by the author (addressing RQ1).

2. Investigates the effect of users' cultural background on preferences for repair strategies in DVAs by comparing evaluations from British participants (BL1, representing a low-context culture) and Saudi participants (SL1, representing a high-context culture) interacting in their native language, addressing RQ2.

3. Investigates the effect of interaction language (Arabic as a native language vs. English as a non-native language) on repair strategy preferences by comparing evaluations from Saudi participants interacting in Arabic (SL1) and the same population interacting in English (SL2), addressing RQ3a.

4. Explores the effect of individual-level factors, including computer self-efficacy and prior experience with technology and DVAs, on users' repair strategy preferences, (addressing RQ3b).

5. Provides qualitative insights into user perspectives on repair strategies, highlighting experiences, preferences, and expectations through semi-structured interviews, (addressing RQ4).

6. Extends the use of the pairwise comparison method to evaluate user preferences for repair strategies in spoken (voice-based) dialogue systems, demonstrating its effectiveness in capturing nuanced cross-cultural and linguistic preferences, (detailed in Chapter 4).

7. Proposes a set of recommendations for implementing repair strategies in goal-oriented SDS tasks, based on the integrated findings from the conducted studies, (discussed in Chapter 7).

1.5 Thesis Outline

The current thesis is structured into the following three main parts and corresponding chapters, each addressing a distinct but interconnected research aim. Figure 1.1 provides a visual summary of the thesis structure. It begins with a systematic scoping review to map existing research on system- and user-repair strategies in SDSs (RQ1). This review forms the foundation for the two empirical studies by providing a conceptual framework for analysing repair strategy types.

The second part of the thesis consists of two empirical studies designed to investigate user preferences for system repair strategies in the context of communication breakdowns. These studies examine both cross-cultural differences and within-cultural variation in user preferences. Study 1 employs a quantitative approach to explore the effect of cultural background on preferences for repair strategies, comparing participants from low-context culture (United Kingdom) and high-context culture (Saudi Arabia) in native-language interactions (RQ2).

Study 2 adopts a mixed-methods approach to examine within-group variation among participants from a high-context cultural background (Saudi Arabia), comparing their preferences and perspectives when evaluating system repair strategies presented in a native (Arabic) versus non-native (English) interaction language (RQ 3a). Study 2 also examines the role of individual user differences, such as computer self-efficacy and previous experience, on repair strategy preferences (RQ3b). Furthermore, as part of its mixed-methods design, Study 2 includes qualitative interviews to gain deeper insight into the user perspective of repair strategies (RQ4).

Across both empirical studies, participants are grouped according to their cultural background and interaction language, as follows: (1) British participants interacting in their native language, English (BL1); (2) Saudi participants interacting in their native language, Arabic (SL1); and (3) Saudi participants interacting in their non-native language, English (SL2). This structure enables a comparative and in-depth investigation of how cultural context and interaction language influence user preferences and perspectives.

Part I: Introduction, Background, and Systematic Scoping Review, this part introduces the thesis by outlining its structure, motivation, and research context. It also presents a systematic scoping review of existing repair strategies in SDSs, identifying key gaps and informing a conceptual framework for the subsequent two empirical studies.

Chapter 1 - Introduction: introduces the research topic, outlines the motivation and context of the PhD thesis, and presents the main research questions. It also defines the scope and context of the thesis, summarizes its key contributions, and provides an overview of the thesis structure.

Chapter 2 - Background: provides the theoretical and conceptual background for the thesis.

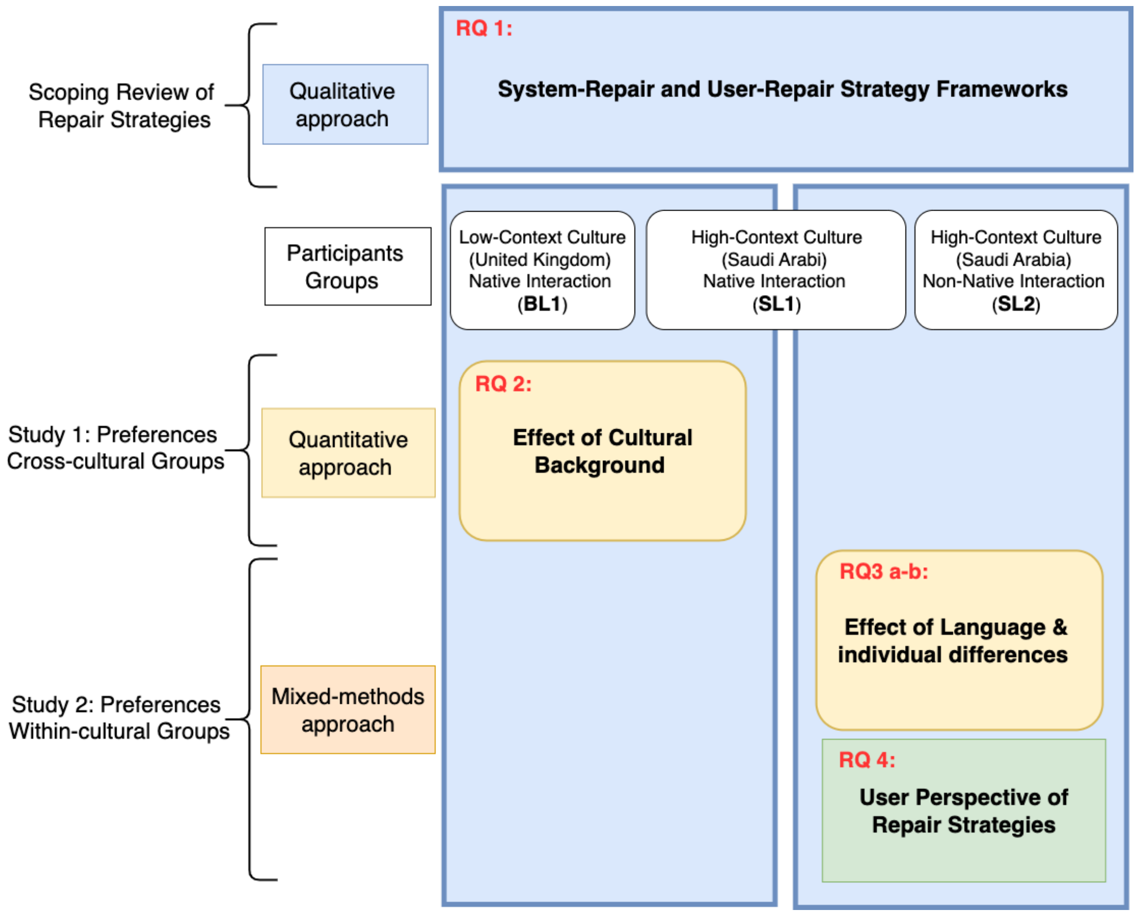


Figure 1.1: Overview of the thesis structure and research questions

It reviews the relevant literature on SDSs, communication breakdowns, and repair strategies. It also defines core concepts of communication styles and cultures in HHI and HCI, such as Hall's context theory, Hofstede's cultural dimensions, and the Computers as Social Actors (CASA) paradigm. In doing so, it provides the foundation for the scoping review and the two empirical studies presented in the following chapters. In addition, the chapter highlights the importance of developing adaptive SDSs that can respond to diverse user needs and preferences, particularly in the context of communication breakdowns.

Chapter 3 - System and User Repair Strategies: A Systematic Scoping Review: presents a systematic scoping review of the literature on communication repair strategies in SDSs. It identifies and categorizes both system-initiated and user-initiated strategies employed to repair conversational breakdowns. The review also examines the methods used to evaluate these strategies and highlights gaps in the existing literature. This chapter contributes two comprehensive frameworks: one for system-initiated and one for user-initiated repair strategies. This chapter serves as the conceptual basis for the two empirical studies.

Part II: Two Empirical Studies on User-Centred Evaluation of Repair Strategies in Spoken Dialogue Systems, this part presents two interconnected empirical studies that explore user preferences and perspectives on repair strategies in SDSs. Study 1 (Chapter 5) employs a quantitative approach to investigate the influence of cultural background on user preferences, addressing RQ 2. Study 2 (Chapter 6) adopts a mixed-methods approach to examine the influence of interaction language (RQ 3a) and individual-level user characteristics (RQ 3b), while also providing a qualitative exploration of users' perspectives on repair strategies (RQ 4). This part begins with a dedicated Methodology Chapter (Chapter 4), which outlines the overall research design, data collection procedures, and the consistent experimental framework applied across both studies

Chapter 4 - Methodology for User-Centred Empirical Studies: outlines the consistent experimental design employed across the two empirical studies, both of which use a pairwise comparison method to investigate the influence of cultural background (RQ 2) and interaction language (RQ 3a) on user preferences for repair strategies. It also describes the quantitative measures used to assess the effect of individual-level factors, such as computer self-efficacy and prior experience (RQ 3b). In addition, this chapter explains the qualitative approach used to explore user perspectives on repair strategies and communication breakdowns through semi-structured interviews (RQ 4). It also presents the preparatory procedures, including the design and implementation of a pilot study, conducted prior to the main empirical investigations.

Chapter 5 - Study 1: Effect of Cultural Background on Repair Strategy Preferences: investigates the effect of cultural background on user preferences for repair strategies in SDSs. Using a pairwise comparison method, the study quantitatively compares evaluations from participants interacting with the system in their native language. The analysis focuses on cross-cultural differences by contrasting preferences between participants from the United Kingdom (BL1, representing a low-context culture) and Saudi Arabia (SL1, representing a high-context culture) interacting in their native language. In addition to identifying these differences, the study explores how users from different cultural backgrounds align with specific repair strategies, drawing on relevant communication theories to interpret patterns of alignment and variation.

Chapter 6 - Study 2: Effect of Interaction Language on Repair Strategy Preferences: presents a mixed-methods investigation into how interaction language and individual-level user characteristics influence preferences for repair strategies in SDSs. Using a pairwise comparison method, the study quantitatively examines the effect of interaction language (RQ 3a) by comparing Saudi participants' evaluations of repair strategies presented in Arabic (SL1) versus English (SL2), while holding cultural background constant. The chapter also explores the role of individual-level factors (RQ 3b), including user demographics, self-efficacy, prior experience with digital technologies, and prior experience with DVAs, to assess how these characteristics shape preferences for repair strategies. In addition to the quantitative components, Study 2 includes a qualitative investigation (RQ 4) based on semi-structured interviews conducted with Saudi participants. This qualitative component deeply explores users' perspectives, expectations, and experiences related to repair strategies and communication breakdowns in SDSs.

Part III: Integrated Discussion and Conclusions, This part synthesizes findings from the qualitative and quantitative studies to provide an integrated discussion of key insights. It concludes with theoretical reflections, practical design implications, thesis limitations, and directions for future research.

Chapter 7 - General Discussion: synthesizes the findings from all studies, identifying overarching patterns and insights related to user preferences and perspectives on communication repair strategies in SDSs. It discusses theoretical implications and provides design recommendations for adaptive SDSs.

Chapter 8 - Conclusions: This final chapter summarizes the key findings of the thesis and revisits the research questions in light of the evidence from the conducted studies. It emphasizes the importance of user characteristics in shaping preferences and perspectives on repair strategies in SDSs. The chapter also discusses the limitations of this PhD thesis and outlines directions for future research.

Finally, the thesis includes three appendices containing complementary materials. Appendix A presents the ethics approval, participant information sheets, consent forms, and experimental questionnaires for the pilot study, while Appendix B provides the corresponding materials for the empirical studies. Appendix C contains supplementary statistical analyses, including full logistic regression outputs and additional tests that support the quantitative results.

Publications

Two research papers from this PhD thesis have been published in peer-reviewed venues, with both receiving the Best Full Paper Award at their respective conferences:

1. Alghamdi, E., Halvey, M. and Nicol, E., 2024. System and user strategies to repair conversational breakdowns of spoken dialogue systems: a scoping review. In Proceedings of the 6th ACM Conference on Conversational User Interfaces (pp. 1-13) [8]. This paper is discussed in Chapter 3.

2. Alghamdi, E., Halvey, M. and Nicol, E., 2025. The effect of interaction language on preferences for communication repair strategies in Digital Voice Assistants (DVAs): a comparative study. In Proceedings of the Mensch und Computer 2025 (pp. 146-162) [9]. This paper is discussed in Chapter 6.

Chapter 2

Relevant Background

To understand the current work on how user characteristics influence preferences for repair strategies in SDSs, it is essential to first establish key foundational concepts. This chapter provides the necessary background to contextualize the research presented in this thesis. It begins with an overview of SDSs, describing their different forms, such as DVAs and Conversational Agents (CAs) and outlining the core components that enable spoken interaction (Section 2.1). The chapter then discusses conversational breakdowns, a critical challenge in SDSs, and introduces common types of breakdowns observed in human-computer interaction (Section 2.2). Following this, the chapter explores various conversational repair strategies used by systems and users to resolve miscommunication and restore dialogue flow (Section 2.3). Finally, since communication styles shape how breakdowns are interpreted and managed, Section 2.4 is dedicated to comparing communication in human–human interaction (HHI) and human–computer interaction (HCI).

2.1 Spoken Dialogue Systems (SDSs)

A dialogue system can be broadly defined as "a computer program that supports spoken, text-based, or multimodal conversational interactions with humans" [229, p.11]. SDSs, specifically, refer to interfaces that enable users to interact with computers or other digital devices using natural spoken language [166]. These systems represent a new generation of user interfaces designed to facilitate efficient and intuitive human-machine communication through a conversational paradigm [289]. In the literature, a variety of terms are used interchangeably to refer such systems, including Personal Digital Assistant, Virtual Personal Assistant, Conversational Agents, and Chatbots [229].

SDSs can be broadly categorized into two types: task-oriented and non-task-oriented systems [122]. Task-oriented SDSs are designed to help users accomplish specific goals or tasks, ranging from relatively simple ones like booking a hotel, to more complex activities such as planning travel or managing emergency situations. In contrast, non-task-oriented systems are developed primarily for open-ended or casual conversation without a defined task or outcome. However, the boundary between these two types is not always strict; task-oriented systems may incorporate casual conversation to enhance user engagement, while non-task-oriented systems may occasionally perform functional tasks [122].

Although dialogue systems have been in development for decades, their widespread adoption into everyday life is a relatively recent phenomenon [229]. A pivotal moment in their mainstream acceptance came in 2011 with the launch of Apple's Siri, a voice-enabled personal assistant integrated into smartphones. Since then, SDSs have proliferated across various platforms and modalities [6]. They are now embedded in messaging apps (e.g., Facebook Messenger chatbots), smartphones (e.g., Apple Siri, Google Assistant, Microsoft Cortana, Samsung Bixby), smart speakers (e.g., Amazon Echo, Google Nest), and even embodied social robots (e.g., Pepper, Furhat), marking their integration into daily human-computer interaction on a global scale.

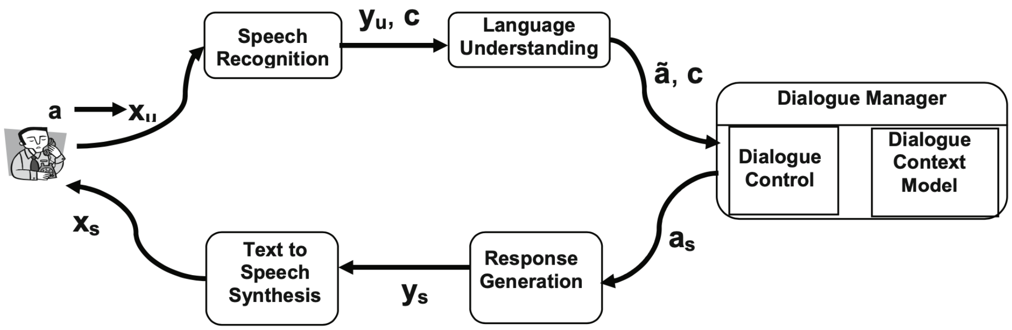
Although this PhD thesis focuses on SDSs in the voice-only domain, where interaction occurs exclusively through speech, the same scope was applied in the scoping review. In the two empirical studies, widely used DVAs, such as Apple's Siri and Amazon Alexa, were introduced as familiar, real-world examples of SDSs. These systems were selected to help participants better relate to the concept of SDSs and to ensure ease of understanding during task engagement. Their wide adoption and familiarity also support their use as representative models of modern voice-based SDSs.

2.1.1 Components of Spoken Dialogue Systems

Understanding the effectiveness of SDSs required familiarity with their core components. A SDS typically consists of five core components [122], as illustrated in Figure 2.1. These components work together to process user input and generate system responses, enabling natural spoken interaction between humans and machines.

The interaction process begins when the user produces a dialogue act (e.g., a command or question), which is expressed as an acoustic signal x_u. This signal is processed by the Automatic Speech Recognition (ASR) module, which converts the acoustic input into a hypothesized word string y_u, often accompanied by a confidence score c. The recognized string is then passed to the Spoken Language Understanding (SLU) component, which interprets the input into a dialogue act estimate $\tilde{a}$ and may extract intents and entities (e.g., Intent: BookFlight, Destination: Glasgow). Breakdowns commonly occur at these early stages of interpretation due to recognition or understanding errors [122]. For instance, ASR may not recognize user's input, or SLU may misclassify user's intent. This could lead to misunderstandings or non-understandings.

Next, the interpreted input is passed to the Dialogue Manager, the central control unit of the SDS. The Dialogue Manager consists of two key subcomponents: the Dialogue Context Model,



$\mathbf{x_u}$ – user acoustic signal
$\mathbf{y_u}$ – speech recognition hypothesis (words)
a – user dialogue act (intended)
ã – user dialogue act (interpreted)
$\mathbf{a_s}$ – system dialogue act
$\mathbf{y_s}$ – system word string
$\mathbf{x_s}$ – system acoustic signal

Figure 2.1: **Basic spoken dialogue systems architecture, based on Jokinen and McTear [122].**

which tracks dialogue history, and Dialogue Control, which decides the system's next move based on current context. This is the component responsible for detecting non-understandings and determining whether a repair strategy is needed. For example, the system may decide to clarify ambiguous input or ask the user to repeat an unclear phrase to maintain mutual understanding and complete the intended task.

Once the system dialogue act a_s is selected, the Response Generation module translates it into a word string y_s. This process involves content planning and realization, where the system selects appropriate content and structures it for output. In this thesis, this component is relevant for its potential to adapt responses based on user preferences, for example, choosing a direct clarification or a polite request depending on context. Finally, the word string is passed to the text-to-speech synthesis module, which generates the acoustic signal x_s that is spoken back to the user.

Figure 2.1 provides an overview of this architecture, showing the flow of interaction and different components where miscommunication and recovery processes may occur. Particular attention in this thesis is given to the Dialogue Manager, where decisions about initiating repair strategies are made and the Response Generation component, which may influence how such strategies are formulated and delivered based on user characteristics.

In recent years, the development of SDSs has shifted from strictly modular pipelines toward language-model-centred or hybrid designs, driven by the emergence of large language models (LLMs). Surveys of dialogue-system evolution describe this as a move from rule/statistical and "neural modular" systems toward pre-trained and then LLM-based dialogue systems, where a single generative model can cover multiple dialogue functions and generalise better via prompting and instruction-following training[258]. In practice, this often means that parts of the traditional NLU/Dialogue Manager/Response Generation stack are increasingly implemented by an instruction-tuned LLM that consumes the ASR transcript plus dialogue context [280]. This LLM-centric shift has clear implications for breakdowns and repair: LLMs make it easier to generate diverse, context-sensitive repair moves (clarification questions, explanations, apologies, or option-offering) and to adapt the formulation of repair to user characteristics; however, they also introduce new failure modes such as fluent but incorrect "hallucinated" content, overconfident confirmations, or inconsistent state tracking, which can themselves trigger miscommunication[121].

2.2 Conversational Breakdowns

Despite advances ASR and SLU components in SDSs, conversational breakdowns remain common. Miscommunication is a fundamental aspect of all communication interactions [161]. Traum and Dillenbourg [248] define miscommunication as instances of action failure (when the speaker fails to produce the intended effect) and misperception (when the hearer cannot recognize what the speaker intended to communicate) or both. While the analysis of such interactions is sometimes referred to as breakdown analysis, a breakdown represents only one extreme within a broader spectrum of miscommunication phenomena [91]. During conversation, interlocutors continually work to ensure that mutual understanding is maintained, a concept referred to as common ground [50, 51].

Miscommunication has been examined across various disciplines, including sociolinguistics, ethnography, communication science, media studies, social psychology, conversation analysis, and natural language processing (NLP)[161]. In the context of HCI, and particularly within SDSs, breakdowns typically occur when the system fails to recognize or correctly interpret user input, or when it produces responses that are inappropriate or confusing [253, 143, 161].

These breakdowns can significantly impact user experience. A failure to correctly interpret user intent or to respond meaningfully can result in user frustration, confusion, or a loss of trust in the system's reliability [160, 80, 284, 154]. Even when repair mechanisms are available, frequent breakdowns can lead to interaction abandonment [145, 80]. For instance, in a task-oriented banking chatbot scenario, Li et al. [145] found that nearly half of the users who discontinued long-term engagement had experienced only a single breakdown. They further observed that users were most likely to abandon interactions after experiencing three breakdowns, suggesting a limited tolerance threshold in task-oriented settings. In commercial settings, such failures can escalate into public relations issues for companies [206], provoking user anger [105] and leading to long-lasting negative attitudes toward the system [154].

The likelihood of breakdowns in dialogue systems can be influenced by several factors, including the length of the dialogue [39], the type of task being performed [146], and the modality used [129]. For example, Li et al. [146] found that breakdowns were more frequent when users were requesting information as opposed to providing information. In comparing text-based and voice-based systems, Kim et al. [129] reported that text-based systems frequently suffered from grammatical misunderstandings, while voice-based systems were more likely to breakdown caused by ASR errors and timing issues. Similarly, Kisser et al. [131] found that dialectal variation was a major contributor to breakdowns during interactions with Alexa. Their findings also showed that children were more likely to be misunderstood than adults and that group interactions introduced further challenges. In group settings, the voice assistant struggled to distinguish between background noise, conversations, and actual requests. These issues were less prevalent in one-to-one interactions. ASR performance and breakdown rates are also significantly affected by the interaction language used with SDSs. For instance, Bohus and Rudnicky [30] found that breakdowns occurred in 20.7% of interactions with native speakers, compared to 42.3% with non-native speakers.

To better understand the nature of breakdowns, several classification schemes have been proposed in the literature, each offering different perspectives on how miscommunication occurs. A common way to classify conversational breakdowns is by distinguishing between misunderstanding and non-understanding (e.g., [111, 264, 30]). Misunderstanding occurs when the listener assigns an interpretation to the speaker's utterance that appears correct but does not match the speaker's actual intention. In contrast, non-understanding arises when the listener is unable to construct any meaningful interpretation or encounters multiple interpretations with no clear way to choose among them. One of the key differences between these two types is that non-understandings are typically recognized immediately, prompting the system (or listener) to act, whereas misunderstandings may remain undetected until later in the dialogue, if they are noticed at all [232, 69].

In the context of SDSs, non-understanding poses a direct challenge that must be addressed promptly to maintain the coherence and effectiveness of the interaction [30, 233, 161]. When a non-understanding occurs, the system must decide how to react whether by requesting clarification, offering repair options, or providing feedback. Figure 2.2 demonstrates two contrasting breakdown scenarios in SDSs. In the first, a misunderstanding occurs when the system misinterprets a key term ("noon" as "morning"), leading to an incorrect confirmation. In the second, a non-understanding is shown when the system fails to recognize the term altogether and requests clarification. This figure helps visualize the core distinction that misunderstanding involves an incorrect interpretation, whereas non-understanding results in no interpretation at all.

Several models have been proposed to classify breakdown scenarios. For instance, based on Clark's grounding model for HHI [49], Bohu and Rudnicky [30] identified sources of breakdowns at the level at which the mismatch occurs in HCI:

1. Out-of-application [conversation level]: User input is outside the system's functional scope, either due to irrelevant topics (out-of-domain e.g. the user asks the room-reservation system

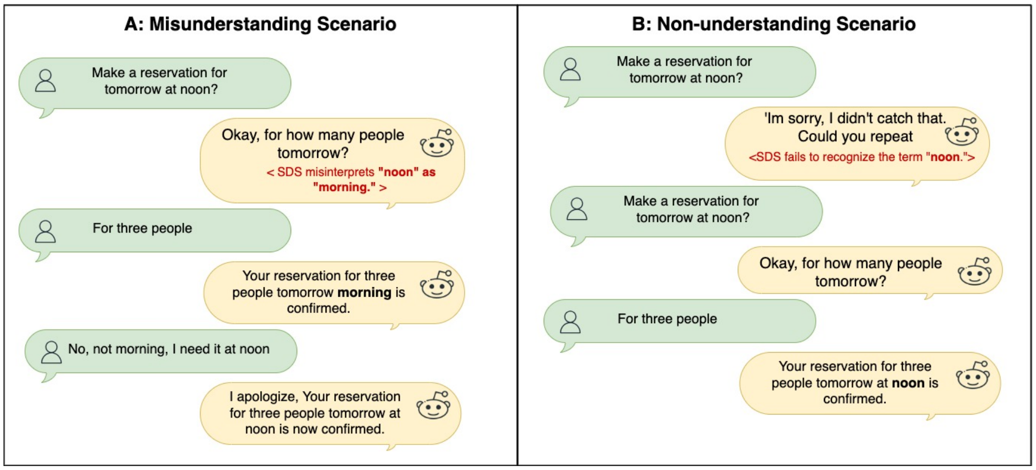


Figure 2.2: **Comparison of Misunderstanding and Non-understanding Scenarios in SDSs**

about the weather) or unsupported in-domain requests (out-of-scope e.g. the user asks if a conference room has windows).

2. Out-of-grammar [intention level]: The input is domain-relevant but does not match the system's expected phrasing or grammar (e.g., saying "erase reservation" instead of the supported "cancel reservation").

3. ASR error [signal level]: The system misrecognizes a correctly formulated utterance due to speech recognition limitations (e.g., "Thursday morning" misheard as "Friday morning").

4. End-pointer error [channel level]: The system fails to detect the correct start or end of the user's speech input (e.g., it cuts off speech or sends an empty utterance).

Similarly, in the SDS context, researchers such as Skantze [233, 232] have offered a fine-grained classification that distinguishes between:"

1. Full understanding: The full intention of the utterance was recognised.

2. Partial understanding: Only a fragment or a part of the full intention was recognised.

3. Non-understanding: No part or fragment of the intention was recognised. Single words were possibly understood

4. Misunderstanding: The listener believed in a partial or full interpretation that was not in line with the speaker's intention".

Conversational breakdowns in SDSs can arise from various sources and manifest in different forms, ranging from complete misunderstanding to total non-understanding. Among these, partial non-understanding, where only fragments of the user's intent are recognized, represents a particularly relevant and recurring challenge. While the scoping review conducted in this PhD thesis included both misunderstandings and non-understandings without focusing on a specific type, the two empirical studies presented will focus specifically on partial non-understanding. Further details regarding this focus are provided in the Methodology chapter 4. The next section explores the conversational repair strategies used to address such breakdowns.

2.3 Conversational Repair Strategies

Miscommunication can arise in any form of dialogue, including interactions with SDSs. Given the potential negative impact of breakdowns on task performance and user satisfaction, it is critical that SDSs are equipped with mechanisms for detecting and addressing such breakdowns. For example, Bohus and Rudnicky found that while low error rates had minimal impact on task success, increasing error rates significantly reduced the likelihood of task completion [30].

Despite continuous advancements in SDS development, particularly in the areas of ASR and SLU, these systems are still vulnerable to conversational breakdowns. When breakdowns occur, either the system, the user, or both may attempt to re-establish mutual understanding through what is commonly referred to as a repair dialogue or strategy [222]. Repair has been defined as "the replacement of an error or mistake by what is correct" [222, p.363]. In HCI contexts, repair is described as the effort to restore shared understanding after a misunderstanding or non-understanding has taken place. As Beneteau et al. [25, p.2] explain, "communication repair refers to the work of restoring shared understanding after conversational partners misunderstand each other," often requiring one party to rephrase or clarify their utterance [25, 172].

The overarching goal of breakdowns handling in SDSs is to ensure user satisfaction and to help guide the interaction back on track after an conversational breakdown occurs [142]. These repair mechanisms are grounded in the theory of common ground or grounding, which suggests that dialogue participants collaborate to establish shared understanding [50, 69, 249]. Within the grounding framework, repair is understood as a means of addressing detected or suspected misunderstandings while minimizing the cognitive and interactional effort required to re-establish mutual understanding.

Repair and breakdown handling remain active areas of research within the field of dialogue technology [26, 8]. McTear [161] identifies four key stages in handling dialogue breakdowns: (1) error prevention, addressed at the design stage through prompts, grammars, and dialogue flow control (e.g. [55]); (2) error detection, which involves identifying signals indicating a possible misunderstanding or non-understanding (e.g. [10]); (3) error prediction, which uses contextual cues to anticipate likely problems before they occur (e.g. [162]); and (4) error recovery, where the system takes action to repair the dialogue and restore progress (e.g. [163, 9]).

This thesis focuses specifically on error recovery, also referred to as conversational repair strategies. Although some researchers (e.g., Skantze [232]) have questioned the use of the term repair for misunderstanding and non-understanding, since it implies that the system must correct a prior misunderstanding, the term is widely used in SDS literature to describe strategies aimed at keeping the dialogue on track, regardless of whether the original breakdown is fully resolved (e.g. [64, 288, 46, 53, 163]). In linguistics, repair is defined more broadly as the set of practices used by interlocutors to address problems in speaking, hearing, or understanding during interaction [220, 221]. Importantly, this view acknowledges that repairs may be initiated even in the absence of actual breakdowns, such as when clarification is requested despite a correct statement [64].

In SDSs, repair strategies can be initiated by the system or the user when a conversational breakdown was detected. While system repairs are pre-designed into the dialogue flow, user repair strategies are often unpredictable and reflect individual traits such as prior experience and emotional and mental state [229, 163, 151]. This distinction has led to two main streams of research in HCI: one focused on system-initiated repair strategies (e.g., [286, 46]), and another on user-initiated strategies in response to system failures (e.g., [163, 53]). The following two subsections explore these streams in more detail.

2.3.1 System Repair Strategies

System repair strategies have been widely studied in the context of dialogue systems, with researchers identifying a range of approaches that systems can employ to recover from communication breakdowns. These strategies vary in form and function such as asking users to repeat an input, offering confirmation, providing alternative options, or giving explanations and apologies. Importantly, the effectiveness of a given strategy is not universal but depends on a range of contextual and moderating factors, such as the type of interaction or task [56, 19], the nature of the breakdowns [247, 170, 278], user characteristics [17, 167], and environmental conditions [95, 285, 215]. The literature therefore does not point to a single "best" repair strategy but highlights the need to consider how strategies interact with these conditions when evaluating or designing system behaviour.

One of the most common approaches is the re-prompt strategy, where the system simply repeats its previous prompt, signalling a non-understanding at a pragmatic level compared to more explicit acknowledgements such as "I didn't catch that" [192]. Closely related is the rephrasing request, in which users are asked to restate their input. These are the most frequently implemented repair strategies in practice [192, 36, 90], yet evidence shows that repeating and rephrasing are often less effective and can frustrate users. For example, repeated requests to rephrase frequently lead users to hyper-articulate, which paradoxically increases ASR errors [81]. In some cases, such as in task-oriented banking chatbots, users preferred to abandon the interaction rather than rephrase their input, perceiving that the burden of repair had been unfairly shifted onto them [145, 26].

Alternative strategies focus on assisting users directly by providing options or explanations. In text-based interactions, Ashktorab [17] found that giving users alternative choices or explanatory feedback ranked among the most preferred repair strategies. Similarly, offering suggestions has been shown to improve perceived accuracy and likeability of chatbots compared to rephrasing or politeness-based repairs [270]. Explanations can also be powerful when tailored. For example, highlighting specific misunderstood words increased satisfaction among users with positive AI attitudes but decreased satisfaction among those with negative attitudes [287]. Research on in-vehicle voice interfaces showed that providing examples of valid inputs outperformed simple repeats or requests to rephrase, as concrete examples reduced confusion and clarified system expectations [128, 213]. However, not all forms of explanation are equally effective. Raux el al. [213] show that long, generic prompts were poorly received, while concise examples were helpful . Similarly, while providing options is often perceived as quick and efficient, it can also frustrate users if the correct option is missing or the set is too limited [36]. More broadly, although suggestions and explanations reduce inappropriate system responses, they do not necessarily shorten dialogues or guarantee successful task completion [79].

Another category includes social or politeness strategies, such as apologising when errors occur. Research has found that apologies can make systems appear more usable and appealing, and reduce user frustration [284, 5, 251]. However, findings are mixed, for example, Engelhardt et al. [72] reported that apology-based strategies were rated lowest in likeability and perceived intelligence. The effect of apologies may also depend on user mood [5] or demographic factors [260]. For example, Wang et al. [260] found that when service breakdowns occurred, apologies from human staff increased satisfaction among younger travellers, while apologies from robots increased satisfaction among older travellers.

Some systems employ more anticipatory or avoidance strategies, choosing to move forward without explicitly acknowledging the breakdown. This may involve advancing to the next task step, transferring to a human agent, or remaining silent [30, 233, 36]. Such strategies can sometimes yield better outcomes, as users often reformulate or hyper-articulate their input (such as higher, louder or longer) when they face conversational breakdowns leading futhre complicatoin [124, 160, 179, 242]. Bohus and Rudnicky [30] and Hirschberg [110] show that ignoring breakdowns can increase recovery rates because users change their speech. In a voice-controlled game, Zargham et al. [286] compared traditional repeat requests with anticipatory strategies that did not flag errors, finding that the latter improved user experience and intuitiveness of control. Similarly, in customer service, Braggaar et al. [36] found that users preferred chatbots for simple, quickly resolved issues, but for complex tasks they preferred transferring to human agent as they trusted human agents more.

Finally, research suggests that relying on a single repair strategy is often insufficient (e.g.[39, 73, 26]). Instead, employing multi-strategy or adaptive repair approaches-combining, for example an apology with an explanation or promises [73], can increase resilience and sustain engagement [35, 26]. Review work confirms that combining strategies helps users persist with the system, reduces abandonment, and promotes more effective recovery from breakdowns[26].

2.3.2 User Repair Strategies

Researchers have highlighted the need to better classify user repair behaviours, as this can support the design of system responses that guide users toward inputs with a higher chance of resolution [149, 179]. User repair behaviour is more completed and unpredictable [18, 216]. When prompted to rephrase, users may simplify their input by reducing words or omitting details [87, 90, 192], or expand it by adding further information [152, 179]. Other variations include syntactic reformulation, changes in formality, or the use of synonyms and alternative terminology [160, 176, 87]. Importantly, users' behaviours are not random but shaped by how voice interactions are designed and responded [173, 39]. For example, research show that CAs' responses often signalled errors, which users then interpreted as cues for diagnosing the source of trouble [205, 204]. Similarly, when participants suspected an ASR error, they tended to repeat their utterance more slowly and concisely, whereas system boundary errors were more often met with reformulations or simplifications [128]. These findings suggest that the way CAs handle errors influences users' understanding of failures and their subsequent repair strategies [173].

Most studies of user-repair strategies in SDSs show that the most instinctive way for humans to correct mistakes is simply to repeat their utterance. Repetition and rephrasing are the most frequently documented user repair strategies for non-understandings [192]. In computer game interactions, for example, young children often relied on repetition as their initial repair strategy and rarely gave up when breakdowns occurred [46, 25]. Similarly, in collaborative dialogue systems, Weitz et al. [266, p. 7] observed that "the most successful user strategy was to rephrase their answer; the most successful system strategy was to provide the user with either new or repeated information." However, although repeating may be the most common and natural reaction when a system mishears, it is also often the least effective. Unlike non-correction dialogue turns, correction turns are poorly recognized because they are typically marked by hyper-articulation, including slower speech, exaggerated stress, and unusual prosodic features [124, 152]. Users encountering a breakdown frequently repeat their input with louder volume, slower rate, longer pauses, or higher pitch, which paradoxically leads to further recognition failures [124, 245, 110]. Myers et al. [180] also found that, while first repair attempts often involved repeating, simplifying, or expanding the input, those who relied primarily on repetition reported lower satisfaction with voice interface interactions. In a task-oriented chatbot, Dippold [70] similarly found that although rephrasing was the most frequently used strategy, it was also the least successful, whereas restarting a command was far more effective but rarely attempted by users.

In addition to repetition and rephrasing, users may also try to omit or add information to repair conversational breakdowns. For mis-recognition errors, Goldberg et al. [90] found that omission was the most common repair strategy, followed by repetition and paraphrasing. Some users also make use of system commands such as “start over” or “scratch that.” Bulyko et al. [39] reported that ASR performance significantly improved when users relied on such commands compared to repetition. However, Souvignier et al. [238] cautioned that while command-based repairs can be highly effective, they also carry the risk of altering dialogue flow in critical ways, and therefore should only be accepted when the system can process them with high confidence.

2.4 Communication Styles in HHI and HCI

2.4.1 Communication in Human-Human Interaction (HHI)

Communication styles are closely linked with cultural values, shaping how individuals express themselves, interpret meaning, and manage interaction. The way people communicate is not random but reflects shared cultural norms and expectations[88]. As Giri [88, p. 12] notes, “Culture provides its members with an implicit knowledge about how to behave in different situations and how to interpret others’ behavior in such situations”. In this sense, communication styles can be understood as both a product and a mirror of culture, serving as a framework through which individuals interpret and navigate interaction.

Norton defines communication style as “the way one verbally, non-verbally, and paraverbally interacts to signal how literal meaning should be taken, interpreted, filtered, or understood” (as cited in [88, p. 125]). Cultural knowledge and social behaviour are transmitted through language which is the primary communication tool [136]. Therefore, human beings could be partly defined by their communicative style, which reflects their cultural background and shapes how meaning is co-constructed in interaction.

Despite the absence of a single agreed definition, most scholars describe culture as a ”shared system of attitudes, beliefs, values, and behaviours” [86, p. 16]. Similarly, Soley and Pandya [235, p. 206] defines culture as “a group who share a certain system of perceptions and values”. Culture underlies all aspects of social behaviour and influences the rules and patterns of language [212] and shaping communicative preferences [193]. Hall [102] also emphasizes that culture is often subconscious, describing it as a way of life that is most visible when individuals encounter differences.

2.4.1.1 Hall's theory of low-context and high-context cultures

One of the most influential contributions to intercultural communication is Hall's theory of low-context and high-context cultures. In 1976, the terms were introduced by Hall [103] to describe how communication styles differ across cultural groups. Hall [102] defines culture as the way of life of people. Hall emphasized that "meaning and context are inextricably bound up with each other" [104, p. 36], arguing that to understand communication, one must consider not only the words themselves but also the surrounding context in which they are used. His framework offers an important lens for examining how cultural values are reflected in communication styles.

In low-context cultures, good communication is precise, simple, and explicit. Messages are expected to be clear and easily understood at face value, and repetition is often appreciated if it helps to avoid ambiguity. By contrast, high-context cultures favour communication styles that is sophisticated, layered, and detailed. Meaning is often implied rather than directly expressed, requiring listeners to "read between the lines" and rely on shared understanding, relational cues, and situational context.

Meyer [165], in her book *The Culture Map*, expands on Hall's framework by introducing the Communicating scale, which visually positions cultures along a continuum from low-context to high-context communication. As shown in Figure 2.3, countries such as the United States, Germany, and the United Kingdom are positioned toward the low-context end, while Japan, China, and Saudi Arabia appear toward the high-context end. This continuum demonstrates how cultural background shapes expectations of clarity, directness, and contextual details in communication.

Figure 2.3: **Communicating scale showing the continuum between low- and high-context cultures by Meyer [165]. The circles highlight the countries from which participants in this PhD thesis were recruited, representing their cultural backgrounds.**

Hall [103] suggested that messages themselves can be situated along a continuum ranging from low-context forms, such as direct statements and logical explanations, to high-context forms, such as greetings, gestures, or subtle cues. The extent to which a culture is characterized as high- or low-context therefore depends on the relative predominance of one type of message over the other [71]. Importantly, however, "no culture can be conceptualized as exclusively high-context or low-context", as individuals in all societies use both explicit, logical expressions and indirect, allusive forms of communication depending on the situation, regardless of the dominant cultural orientation [71, p. 86].

2.4.1.2 Hofstede's cultural dimensions

While Hall's low- and high-context theory emphasizes communication context, Hofstede's cultural dimensions complement this perspective by focusing on the underlying beliefs and values that drive communication preferences. Culture according to Hofstede is a "collective programming of the mind which distinguishes the members of one human group from another" [113, p. 302]. Amongst many models operationalize cultures, Hofstede's cultural model has been the most used and influential cultural framework in cross-cultural social scientific research [240, 164, 125].

In his research [114, 113, 115], Hofstede identified five dimensions, which have a significant impact on behaviour in across cultures:

- **Power Distance** refers to the extent to which less powerful members of a society accept and expect unequal distribution of power. In high power-distance cultures (e.g., Malaysia), individuals are less likely to challenge authority and often rely on hierarchical communication. In low power-distance cultures (e.g., Denmark), communication is more egalitarian, with subordinates feeling comfortable questioning leaders.

- **Individualism vs. Collectivism** reflects whether people define themselves primarily a independent individuals or as part of a group. In individualistic cultures (e.g., the United States, the United Kingdom), communication tends to be direct, task-focused, and assertive. In collectivist cultures (e.g., Japan, Saudi Arabia), harmony and group cohesion are prioritized, leading to more indirect communication. Relationship prevails over task and harmony should always be maintained.

- **Masculinity vs. Feminity** describes whether a culture is driven more by competition and achievement (masculinity) or by care, cooperation, and quality of life (femininity). Cultures higher on masculinity (e.g., Japan) may value assertiveness emphasis on emphasis on competition, whereas while more feminine cultures (e.g., Sweden and Norway) focus on quality of life with an importance of placed on the well-being of relationships.

- **Uncertainty Avoidance** refers how comfortable members of a culture are with ambiguity and unstructured situations. In high uncertainty-avoidance cultures (e.g., Latin and Arab countries), individuals prefer explicit instructions, detailed explanations, reduced ambiguity and clarity and structure. In contrast, low uncertainty-avoidance cultures (e.g., Denmark, Sweden and the UK) are more tolerant of ambiguity and opinions and lower stress and anxiety.

- **Long-Term vs. Short-Term Orientation** reflects whether a culture emphasizes long-term perseverance and future rewards (long-term orientation, e.g., China) or immediate results and respect for tradition (short-term orientation, e.g., the US). In communication, long-term oriented cultures may prioritize patience and saving face, while short-term oriented cultures often prefer straightforward, results-oriented exchanges.

As this thesis focuses on two distinct cultural contexts, Saudi Arabia and the United Kingdom, it is important to highlight how these countries score across Hofstede's five dimensions. Figure 2.4 presents a comparative visualization of the five Hofstede dimension scores, as provided by The Culture Factor website[1] using country comparison tool.

While Hofstede's framework includes five cultural dimensions, this thesis focuses on individualism–collectivism and uncertainty avoidance, as these dimensions are most directly associated with communication preferences and repair strategies in SDSs. Individualism–collectivism captures the balance between directness and harmony in interaction, whereas uncertainty avoidance reflects tolerance for ambiguity and the need for clarity and detail in communication. This focus is further reinforced by prior research showing that individualism and uncertainty avoidance, together with power distance, are among the cultural dimensions most consistently linked to digital engagement and technology use [4].

Individualism–collectivism is one of the most widely studied cultural dimensions in cross-cultural research. It is considered by some scholars to be more powerful in explaining cross-country differences in communication patterns than other cultural dimensions [246, 195]. Hofstede [115, 12] noted that "individualism tends to prevail in developed and Western countries, while collectivism prevails in less developed and Eastern countries." In contrast, uncertainty avoidance refers to the extent to which people "attempt to avoid experiences which they perceive as unstructured, ambiguous, or unpredictable" [113, p. 308]. Importantly, uncertainty avoidance should not be confused with risk avoidance, "it does not describe one's willingness to take or avoid risk, but rather is associated with preferences for clear rules and guidance" [114, p. 149].

The distinction between high- and low-context cultures proposed by Hall [104] aligns closely with Hofstede's dimensions. Kim [130] draws a relationship between Hofstede's individualism–collectivism and Hall's theory of high- and low-context cultures. Characterizations of high- and low-context communication systems are closely associated with the characteristics of individualism and collectivism.

[1] https://www.theculturefactor.com/country-comparison-tool?countries=

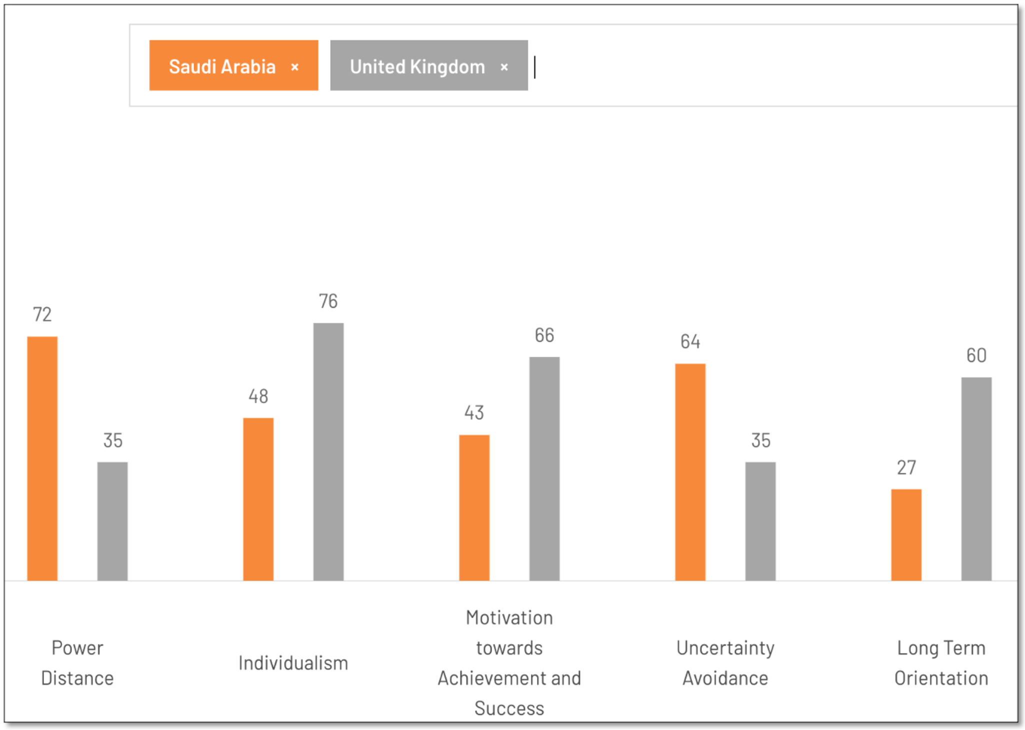


Figure 2.4: **Comparison of Hofstede's cultural dimension scores for Saudi Arabia and the United Kingdom, adapted from The Culture Factor website (Hofstede Insights). Note: the dimension labelled "Motivation towards Achievement and Success" corresponds to Hofstede's original Masculinity–Femininity dimension.**

Research also shows that individualism–collectivism shapes communication style, with individualistic cultures tending toward low-context and explicit communication, while collectivistic cultures favour high-context and indirect communication [97]. For example, Saudi Arabia, typically considered a high-context culture, also scores lower on individualism and higher on uncertainty avoidance, reflecting preferences for group harmony and for reducing ambiguity through detailed guidance or elaboration. In contrast, the United Kingdom, as a low-context culture, scores higher on individualism and lower on uncertainty avoidance, reflecting greater tolerance for ambiguity and a preference for concise and direct communication.

In the following HCI section, these differences establish a cultural foundation for exploring how users from these contexts interact with systems and perceive repair strategies in SDSs.

2.4.2 Communication in Human-Computer Interaction (HCI)

2.4.2.1 Computers as Social Actors (CASA)

The Computers as Social Actors (CASA) paradigm, derived from the media equation, proposes that people treat computers and other media much like real social partners [214]. In line with this view, users routinely import the same social and cultural norms they apply in HHI, attributing intentions, politeness, and cultural expectations to systems. Early laboratory studies showed that people display fundamentally social responses to computers politeness, reciprocity, and group bias, much as they would with other people [181, 182].

CASA explains these effects through minimal social cues: even "a very simple human feature can activate people's corresponding human–human social scripts" and lead users to ascribe human characteristics to a machine [252, p. 1231]. Consequently, people expect conversational systems to understand requests and reply appropriately, similar to expectations of a human partner [271]. When interfaces present social signals (such as voice, name, turn-taking, small talk) computers can mimic human interaction [76], and users respond accordingly [181].

Evidence from CAs reinforces this transfer of social behaviour. Liao et al. [147] document rich user practices (e.g. feedback, playful chit-chat, system inquiries, and habitual utterances) showing how everyday social behaviours migrate into human–computer exchanges. Relatedly, human-like perceptions such as anthropomorphism and personification shape how users relate to DVAs [175, 277, 20]. Recently, Schneider [224] analysed over 200,000 conversations with large language models (LLMs) and found that many users initially approach LLMs as conventional tools. After early interactions, they often shift toward more natural, polite, and context-sensitive phrasing which indicates a cognitive re-framing of the system as a conversational partner.

CASA also helps explain behaviour during conversational breakdowns. Because users typically lack insight into a system's boundaries and recognition processes, they cannot repair misunderstandings in a way that is meaningful from the voice interface's internal perspective [179]. Instead, they import HHI repair strategies(e.g. repetition, rephrasing, giving more context) when a breakdown occurs. Studies with children show the same pattern; young users draw on HHI repair strategies to fix voice-driven interface misunderstandings. Their strategies mirror those of adults, though children are often more persistent and patient [46].

While CASA paradigm demonstrates that people respond to interactive systems in socially meaningful ways, it does not suggest that users genuinely believe such systems are human. Rather, CASA research emphasises that social responses to computers are largely automatic and cue-driven which arises from minimal interactional features such as language use, voice, politeness markers, and turn-taking structures[214, 181]. Importantly, users typically remain aware of the artificial nature of the system, yet nonetheless apply social norms and conversational scripts derived from human–human interaction in a mindless manner [182].

In this thesis, CASA is therefore adopted as a framework for understanding how interface design and conversational cues prime socially grounded interactional behaviour, rather than as evidence that users anthropomorphise systems in a literal or reflective sense.

2.4.2.2 Cultural Background

Building on Hall's and Hofstede's frameworks and the CASA premise that users import human–human norms into HCI, empirical work shows that cultural background systematically shapes how people perceive, evaluate, and interact with conversational systems and social robots.

Analysing one million social media discussions about the perception of CAs, Liu et al. [153] found that Chinese users tended to adopt a more hedonic orientation and perceived voice-based/embodied CAs as warmer and more competent, whereas U.S. users emphasised functional utility and showed greater ambivalence. In both countries, perceived warmth was a key driver of positive emotion. Exploring culture-contingent preferences for communication styles, Rau et al. [212] also reported that Chinese participants favoured implicit recommendations and judged the robot as more likeable, trustworthy, and credible than German participants, who preferred explicitness and were less accepting of implicit suggestions.

Social non-verbal cues and emotion perception also interact with cultural orientation and individual factors. In service scenarios, Yu et al. [283] showed that perceptions of interpersonal warmth and satisfaction varied by culture and whether the staff was a machine vs. human. When a digital assistant displayed emotional non-verbal behaviours, Iranian participants attributed greater empathy, trustworthiness, and helpfulness than German participants [116]. The same study found that a broader set of cues (smile, head nod plus smile, head nod, sad face, head down, and dropping the arms plus sad face) signalled empathy for Iranians, whereas Germans primarily recognised head nods [116]. Similarly, compared with American listeners, German listeners gave higher overall valence ratings but judged Alexa's voice as less excited (lower arousal), indicating cross-cultural differences in mapping prosody to affect in both human and synthetic speech [85].

Finally, linguistic localisation and dialect matter for acceptance and credibility across cultures. For example, Trovato et al. [250] showed that Egyptians preferred an Arabic-speaking robot and were uncomfortable with a Japanese-speaking one, while Japanese participants showed the mirror pattern. Andrist et al. [12] also found that rhetorical linguistic features weighed more heavily in Arabic than in English when assessing a robot's credibility. Within Arabic interaction, both practical knowledge and rhetoric were most effective in the local dialect rather than Modern Standard Arabic.

Taken together, these findings are consistent with CASA where cultural background shaped expectations about communication styles, non-verbal signalling, language, and persuasion are carried into HCI. This literature points to cultural background as a potentially important factor shaping user expectations in HCI, motivating further investigation into whether dialogue system design and repair strategies may benefit from culturally adaptive approaches.

2.5 Chapter Summary

This chapter provided the necessary background to understand the key concepts of SDSs, cultural theories, and communication frameworks relevant to this PhD thesis. It introduced the field of SDSs and discussed foundational communication theories from both human-human and human-computer interaction perspectives. These concepts form the theoretical basis for the subsequent chapters. The next chapter presents a systematic scoping review of repair strategies in SDSs, focusing on how these strategies have been studied within HCI research.

Chapter 3

System and User Repair Strategies: A Systematic Scoping Review

3.1 Importance

SDSs are a type of interface designed to enable humans to interact with machines naturally through spoken language [289]. In SDSs, voice is the preferred mode of communication with machines because it offers intuitive and hands-free interactions [40, 43, 149]. SDSs are now a fundamental part of our interactions with technology, profoundly impacting domains ranging from customer service to personal assistance. The emergence of SDSs in the form of digital home devices, such as Amazon Echo, and personal assistants, such as Siri, has markedly elevated expectations regarding voice-based interfaces in everyday technology use. According to a 2021 report by Statista, the number of DVAs is projected to exceed the global population by 2024, reaching 8.4 billion units [255].

Nevertheless, the widespread adoption of SDSs and their consistent long-term use still face challenges [17]. Despite notable advancements, SDSs frequently encounter difficulties with complex dialogues and varied speech patterns; thus, expecting them to fully grasp every human utterance remains impractical. These challenges often result in conversational breakdowns and a variety of negative effects. Not only do these breakdowns substantially diminish user satisfaction and damage trust, they also decrease the likelihood of continued use [60, 154]. Furthermore, they often prevent users from completing tasks, leading to task abandonment [17], negatively impacting user experience and damaging the overall effectiveness of SDS performance [60, 154, 144, 17]. In SDSs, communication breakdowns, primarily misunderstandings and non-understandings, pose substantial challenges [30]. Misunderstandings occur when the system incorrectly interprets user input, and are particularly challenging to detect and repair since the system believes it has understood the user correctly. Non-understandings, on the other hand, are recognised immediately by the system, which lacks any viable interpretation of the user's input [30] (See Figure 2.2). Given these challenges, particularly the prevalence of conversational breakdowns, some research has shifted towards developing effective repair strategies, which are crucial for enhancing SDS performance and user satisfaction [48].

Centred on the main dialogue actor in the breakdown-repairing process, research in repair strategies has been broadly divided into two areas: investigations into system-repair strategies and studies on user-repair strategies [52, 35]. System-repair strategies research explores how SDSs can be designed to manage conversations effectively, including mechanisms for grounding (i.e., establishing mutual understanding between system and user), interruptions, and corrections, and assesses their impact on user experience and SDS efficiency [204, 53].

In this scoping review, system-repair strategies refers to those strategies related to user interaction and experience rather than the technical approaches of how these systems execute repairs. Purely technical repairs, which involve back-end algorithmic adjustments, data processing techniques, or system performance optimisations that occur without direct user interaction, fall outside the scope of this review. In contrast, user-repair strategies research aims to comprehend how users interact with and repair conversational breakdowns in SDSs across various domains and contexts, often highlighting adaptations in communication style to improve interactions [156, 180, 128, 179].

In this review, the focus is on strategies adopted by both SDS's main dialogue actors (system and user) to correct conversational breakdowns in spoken SDS. Whereas Chapter2 focused on establishing the theoretical background and key concepts relevant to repair in SDSs, this chapter shifts the emphasis to a systematic scoping review of the empirical literature on how repair strategies are conceptualised and studied in practice. Although some references and themes recur across both chapters, this overlap reflects the complementary roles of conceptual grounding (Chapter 2) and structured evidence synthesis (Chapter3).

The development and effectiveness of these repair strategies have been investigated across multiple disciplines, such as conversational analysis and business. However, the diversity of methodologies and terminologies makes it difficult to conduct comparative analyses or to determine whether findings are truly consistent across studies. Importantly, the comparison and combination of different repair strategies has gained attention as researchers have discovered that employing multiple repair strategies is more effective than relying on a single repair strategy [133, 284]. Such diversity not only complicates the aggregation of findings but also prevents interdisciplinary collaboration, which negatively impacts the development of more effective repair strategies. The objective of this review is to present the first comprehensive scoping review of repair strategies in SDSs, with the aim of consolidating and analysing the various strategies employed by these systems and their users. The methodologies used for studying repair strategies in the SDS domain are analysed, and the characteristics of SDSs are defined. To achieve this, a diverse range of repair strategies developed and proposed for use in SDSs have been identified and categorized, drawing from a broad spectrum of literature. This includes both system-repair and user-repair strategies. This review contributes to Human-Computer Interaction (HCI) by providing frameworks for both system and user repair strategies, offering a standardised terminology and conceptual foundation that researchers can adopt to harmonise terminology across the literature.

This scoping review study is guided by an overarching question and three sub-questions:

- **RQ1: What are the state-of-the-art strategies of conversational breakdown when interacting with SDSs?**
  - **RQ1a**: What types of user-facing strategies are employed by SDSs to repair breakdowns?
  - **RQ1b**: What strategies or tactics do users apply to repair breakdowns in SDSs?
  - **RQ1c**: What research methods have been utilised to investigate and evaluate repair strategies in SDSs, and, in this context, what are the defining characteristics of SDSs?

The most relevant studies to this scoping review were conducted by Feng [77] and Benner et al. [26]. Feng developed a taxonomy to categorise system-repair strategies in CAs, focusing on both theoretical and practical aspects. Their taxonomy includes five dimensions: acknowledgement, reasons, explanation, repair, and delegation to humans. Benner et al. [26] conducted a systematic review of strategies for handling conversational breakdowns in both text- and voice-based agents. They categorised system-repair strategies into six categories: Confirmation, Information, Disclosure, Social, Solve, and Ask. While these studies significantly contribute to the field, this scoping review adopts a broader approach by systematically examining both system-repair and user-repair strategies in SDSs and providing a comprehensive overview. In contrast to the scope of Feng and Benner [77], which includes text- and voice-based interactions, this review focuses on voice-based interactions alone. Additionally, an overview of research methods and key characteristics of SDSs employed across relevant literature is provided, addressing aspects not covered by previous studies. It is important to note that a detailed evaluation of the effectiveness of these strategies on user experience or SDS performance is outside the scope of this review study.

3.2 Method

A scoping review enables us to gather a broad spectrum of literature, offer a comprehensive overview, and address the research questions. As Peterson et al. [202] and Munn et al [177] have suggested, this approach provides a broader perspective compared to delving into a detailed research agenda.

Scoping reviews are an increasingly popular and valid approach for synthesising and mapping research evidence [66, 14, 177]. This review follows both the PRISMA-ScR guidelines and the methodological framework proposed by Arksey and O'Malley [14] for scoping reviews. This framework has six iterative stages: 1) identifying the research question; 2) finding relevant studies; 3) selecting studies; 4) charting the data; 5) collecting, summarising, and reporting the results; and 6) an optional consultation exercise. Stage 6, regarding consultation work, is optional and outside the scope of this review.

3.2.1 Identifying relevant studies

3.2.1.1 Databases

The cross-disciplinary search included four electronic databases: ACM Digital Library, IEEE Xplore, Scopus, and Web of Science. The ACM Digital Library and IEEE Xplore were chosen for their comprehensive coverage of computing and technology papers, making them essential for research in the HCI field [52, 254]. To capture a broader range of scientific domains, Scopus and Web of Science were incorporated into the search strategy. These were selected due to their extensive inclusion of multidisciplinary research and their recognised use as sources for literature reviews in the HCI community [15, 257, 178]. Papers spanning multiple domains beyond computing were retrieved from Web of Science and Scopus, including psychology, (e.g., [149, 160]), linguistics (e.g., [134]), marketing and hospitality management (e.g., [117, 260]), as well as gerontology and cognitive science (e.g., [272, 163]). This enriched the scoping review with diverse perspectives on the study of breakdowns and repair strategies in HCI.

3.2.1.2 Search Query

The next step was a keyword search within the selected databases to identify papers. The literature uses numerous terms inconsistently to describe dialogue systems operating on different devices, such as 'personal digital assistant', 'conversational agent', 'voice chatbot' and 'conversational user interface' [229]. For the current review, the list of terms was derived directly from similar papers that performed systematic reviews on speech-based and conversational interfaces [52, 178, 26]. Two main concepts needed to be elucidated in this scoping review: SDSs and repair strategies (see Table 3.1)

Wildcards were used to simplify the search string tailored to each database's specific features. Employing these keywords, a comprehensive search across four databases was conducted, targeting all peer-reviewed scholarly articles and full conference papers published up to August 2023. The searches were conducted in August 2023. To capture the entire spectrum of relevant research, no lower date restrictions were applied. This allows for a thorough inclusion of foundational and contemporary studies. Keywords were used to search titles and abstracts, yielding 818 papers.

Table 3.1: General Query Structure for All Searches

Concepts	Terms
SDSs	[“spoken dialog system” OR “speech interface” OR “voice user interface” OR “voice system” OR “speech-based” OR “voice-based” OR “speech-mediated” OR “voice-mediated” OR “human computer dialog” OR “human machine dialog” OR “natural language dialog system” OR “natural language interface” OR “conversational interface” OR “conversational agent” OR “conversational system” OR “conversational dialog system” OR “automated dialog system” OR “interactive voice response system” OR “spoken human machine” OR “intelligent personal assistant”] AND
Repair strategies	[“repair strateg*” OR “mitigation strateg*” OR “error-handling strateg*” OR “dialogue confirmation strateg*” OR “handling breakdown*” OR “conversational breakdown*” OR “handling error*” OR “mitigating error*” OR “recognition error*” OR “system error*” OR “error recovery” OR “error handling” OR “Correction type*” OR “failure*” OR “correction*” OR “error detection” OR “error recovery” OR “miscommunication” OR “erroneous interpretation” OR “erroneous situation*” OR “correction mechanism*” OR “recovery mechanism*”]

These were imported into Endnote 20.6 as a single library, including basic metadata e.g. title, author, year, and conference/journal details. Duplicate papers were removed, resulting in 505 papers from the initial database search.

3.2.2 Study selection

In systematic reviews, inclusion and exclusion criteria based on a specific research question are developed to ensure consistency in decision-making [14]. For this scoping review, a series of inclusion and exclusion criteria were developed through multiple rounds of discussions among the co-authors [8].

Due to the wide variety and modalities of dialogue systems, the scope of this review was narrowed to voice interactions and excluded papers dealing with modalities such as text, gesture, eye-tracking and multi-modal interaction (where voice was not a primary modality). As this review focused on the user-interaction perspective, studies purely concerned with technical strategies for handling errors were also excluded. The finalised formal criteria are as follows:

3.2.2.1 Inclusion criteria

1. User studies examining strategies to repair or mitigate breakdowns in SDSs.
2. Research involving solely or primarily voice-based user interactions (e.g., voice assistants, smart speakers).
3. Studies involving at least one user and one SDS, focusing on user interaction rather than system performance or machine-machine interaction.
4. Original research in peer-reviewed journals and full-length conference papers.

3.2.2.2 Exclusion criteria

1. Papers not specifically investigating repair strategies in SDSs.
2. Studies on multi-modal interactions where voice is not the main modality.
3. Dissertations, review articles, conference abstracts.
4. Publications not in the English language

To ensure the relevance of studies to the research questions, inclusion and exclusion criteria were applied in a two-step process. Initially, the titles and abstracts of 505 papers obtained from a database search were screened using these criteria. This led to the identification of 128 papers meeting inclusion criteria. In cases where a paper's relevance was unclear from its title and abstract,

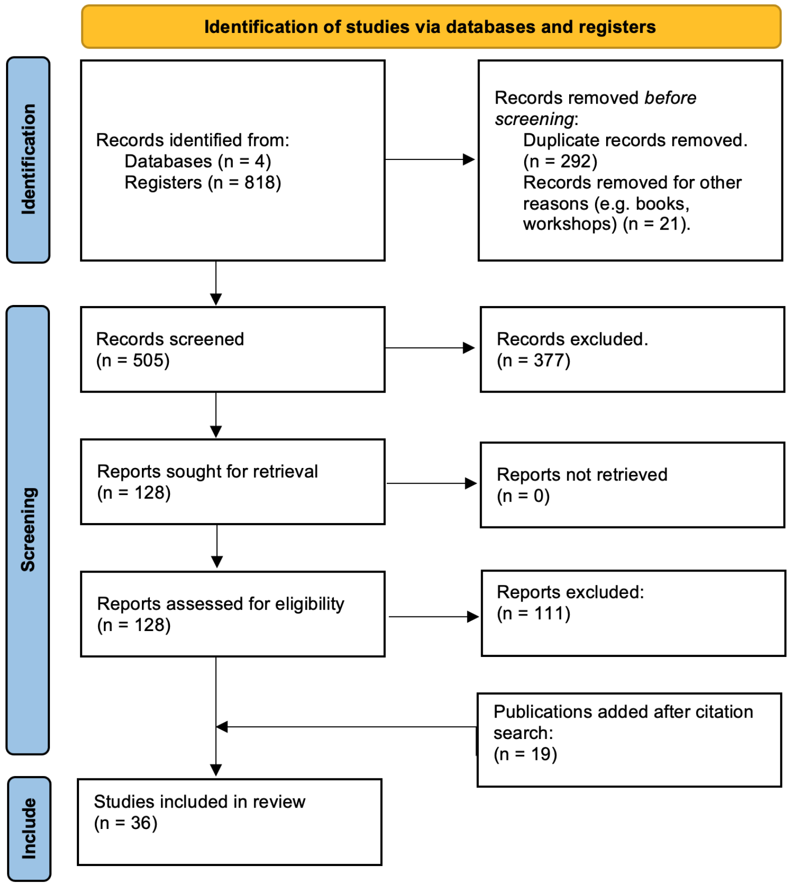


Figure 3.1: **PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flowchart.**

a full-text review was conducted to further assess eligibility. Ambiguities about paper eligibility were resolved through discussions among the co-authors [8], resulting in 17 papers considered eligible from the database search. Additionally, manual searches were employed to capture relevant studies that might not have included key search terms in their titles or abstracts [26, 14]. This involved both backward chaining (examining the bibliographies of identified studies) and forward chaining (using Google Scholar to check citations), together yielding an additional 19 papers. This combined method of electronic and manual searches identified a total of 36 papers (see Figure 3.1).

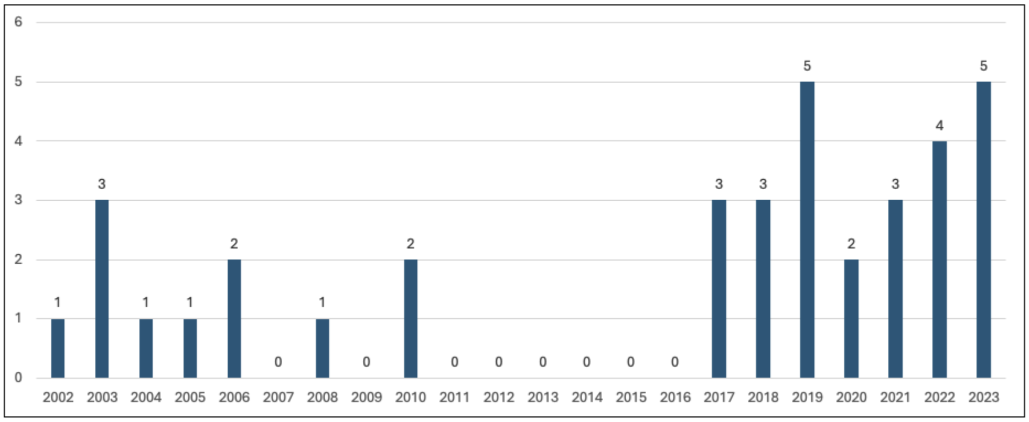


Figure 3.2: **Distribution of Relevant Literature According to Year**

3.3 Results

The findings of the scoping review are organised along the lines of the research questions into three sections.

3.3.1 The characteristics of included studies and SDSs

3.3.1.1 Publication Overview

The publication dates of the reviewed papers ranged from 2002 [231] to 2023 [13, 278, 53, 117]. No studies were found in the period 2010–2017. Approximately one or two studies were published per year before 2011, and this number increased to five in 2023, indicating an increase in awareness among researchers about the importance of handling breakdowns in SDSs (See Figure 3.2 for the number of papers by publication year). While the initial contributions in this field were more exploratory in nature, covering a range of spoken dialogue breakdowns, more recent papers have tended to be more specific, focusing on particular system-repair and user-repair strategies within more defined contexts and settings. For example, around 75% (8 out of 11) of the studies published before 2011 primarily focused on analysing audio corpora from real-world environments in order to explore system-repair and user-repair strategies in communication breakdowns in SDSs (e.g. [231]). Conversely, recent papers (e.g.[53]) delved into specific topics, such as investigating the impacts of four distinct humour styles (employed as system-repair strategies by voice assistants) on user frustration during communication breakdowns. Table 3.2 lists 32 unique publication venues identified from the reviewed studies.

Table 3.2: Publication Venues

Publication Type	Names	Frequency
Academic Journals	International Journal of Human-Computer Interaction	3
	Speech Communication	2
	In INTERSPEECH	2
	International Journal of Hospitality and Tourism Administration	1
	The Service Industries Journal	1
	Universal Access in the Information Society	1
	Recent trends in discourse and dialogue	1
	Frontiers in Psychology	1
	Association in Computer Science	1
	Elektronische Sprachsignalverarbeitung	1
	In Proceedings of SST	1
	Computational linguistics	1
	International Ergonomics Association	1
Conference Proceedings	CHI conference on human factors in computing systems	4
	Proceedings of the ACM on Human-Computer Interaction	1
	ACM conference on interaction design and children	1
	HCI International Conference	1
	Conversational Interruptions in Human-Agent Interactions	1
	ISCA tutorial and research workshop on error handling in spoken dialogue systems	1
	Conversational User Interfaces	1
	International Conference on Automotive User Interfaces and Interactive Vehicular Applications	1
	International Conference on Human-Agent Interaction	1
	Conference on Human Information Interaction and Retrieval	1
	IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)	1
	Conference of the International Speech Communication Association	1
	International Conference on Spoken Language Processing	1
	Conversational User Interfaces	1
	Conference on Conversational User Interfaces	1
	IEEE Workshop on Automatic Speech Recognition and Understanding	1

3.3.1.2 Methodological Overview

Given the diverse and multifaceted nature of the included studies, the author organised the findings by broad methodological approach and then by specific technique. This allowed us to illustrate not only the variety of methodological approaches but also how these interrelate and contribute to the field of SDSs. The papers were divided into three broader approaches: qualitative, quantitative and mixed-method. To ensure accurate categorisation, explicit labels (qualitative, quantitative, or mixed-method) were initially searched for within each paper. Where these labels were not specified, the research designs and methods described were analysed to determine the appropriate category. Studies employing in-depth interviews or observational analyses were classified as qualitative, those using statistical tests or numerical data analysis were classified as quantitative, and studies indicating substantial use of both approaches were classified as mixed-method. Within each approach, the specific research designs and methods of each study are discussed. This includes the data source of corpora (lab- and field-collected corpora), experimental designs (between- and within-subjects experiments) and specific research methods (e.g. interviews, questionnaires and Wizard of Oz method (WoZ)).

The most frequent approach undertaken in the studies in this review was mixed-methods (N=16), followed by qualitative (N=10) and quantitative (N=10) approaches. Among these studies, over half (N=21) conducted analyses of audio corpora, which were categorised into two main types: field- and lab-collected corpora. Field-collected corpora (N=11) are derived from real-world sources, while lab-collected corpora (N=10) are generated and analysed by the authors of the study itself. Most studies utilising field-collected corpora were published prior to 2008 and predominantly focused on telephone platforms, such as the DARPA Communicator Dialogue Travel Planning System [39, 90, 231, 124], the RoomLine conference reservation system [213], the Pizza Corpus ordering system [48], the Let's Go bus schedule information system [213] and the TOOT train information system [152]. Apart from corpus analysis, some of the reviewed studies employed experimental research designs, including between-subject experiments (N=10) and within-subject experiments (N=10).

A range of research techniques was employed for data collection in both qualitative and quantitative studies. These methods predominantly include interviews, questionnaires and Wizard of Oz techniques [163, 286, 149, 84, 128, 232]. A number of studies (N=7) favoured semi-structured interviews [149, 64, 286, 173, 284, 179, 128] while others chose structured interviews [13] or did not specify interview format [25, 232]. Questionnaires were also popular, with 13 studies employing them, predominantly using a 7-point Likert Scale [13, 232, 64, 84, 260, 117, 288, 53], though some used a 5-point scale [286, 127, 143, 128] and others varied in the type of scale used [72, 149, 160]. To visualise the prevalence of the most common techniques across the three broad methodological approaches, a heat map is provided in Figure 3.3. This heat map reveals that corpus analysis is the predominant technique in qualitative studies, while a combination of questionnaires and

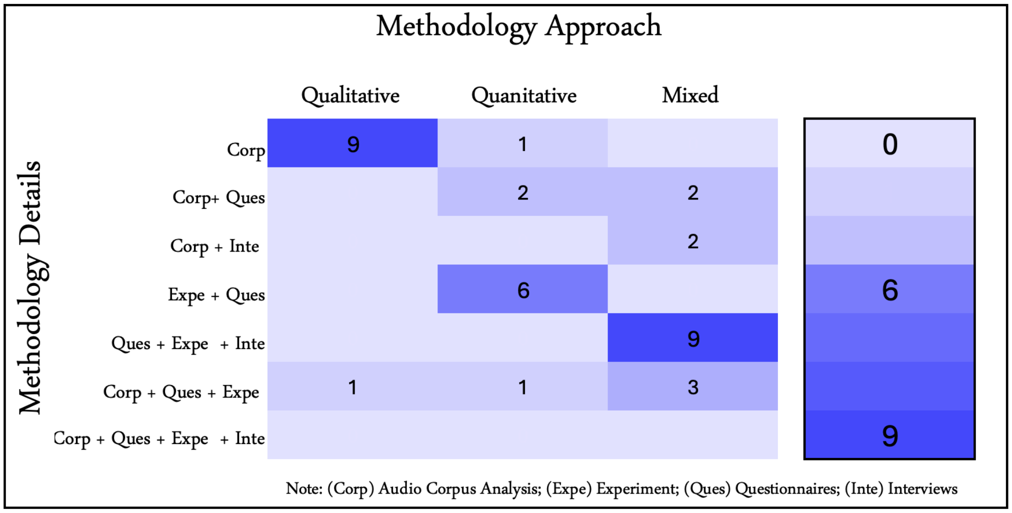


Figure 3.3: **Heat Map of Methodology Details for the Selected Studies**

interviews is typically used in experimental mixed-method approaches. This is followed by questionnaires as the typical technique for experimental quantitative studies. Moreover, the contextual and environmental settings of the reviewed studies varied, with a majority conducted in real-world environments (N=12) and laboratory settings (N=11), followed by controlled field settings (N=5), simulated environments (N=4) and online or crowdsourced settings (N=4). Approximately half of the SDSs (N=16) employed in the reviewed studies were mock-ups in which participants interacted with the SDSs through Wizard of Oz (WoZ) setups(e.g. [284, 53]). These were mostly used in mixed-methods studies (N=10), followed by quantitative studies (N=5) and one qualitative study (N=1).

Sample sizes ranged from as few as five participants in a qualitative study in a real-world setting [204] to as many as 850 participants in an online multi-language quantitative study [134]. Although most study participants were drawn from the general population, four studies specifically included participants with computer science backgrounds [288, 53, 72, 179]. In studies reporting gender information (N=30), there was a notable gender disparity, with males outnumbering females by approximately 20%. Finally, participants' ages varied considerably, from as young as three years old [46] to as old as 92 years old [192].

3.3.1.3 SDS Characteristics Overview

In this sub-section, the key characteristics of SDSs are reviewed to provide a comprehensive overview of their features and interactions. The literature review revealed a heterogeneous landscape of domains and platforms for SDSs. The author categorised SDSs into six distinct domains and seven platforms based on their application contexts, functionalities, and task types. Recurring themes and interaction types were identified across studies and organised into categories reflecting their primary functions and user interactions. Each domain corresponds to settings with specific interaction types and tasks. For example, the personal assistance domain includes SDS applications that aid tasks such as scheduling and reminders, featuring personalised daily interactions. These categories were refined iteratively to accurately reflect the diverse aspects of SDS applications observed in the included studies.

Each study in this review falls into one of the following domains: customer service, personal assistance, public assistance, gaming, automotive and smart home assistance. The platforms identified, including computer speakers, smart speakers, mobile devices, smart home devices, in-vehicle voice-user interfaces, telephones and robots, were determined via detailed analysis of systems utilised across the studies reviewed. Most studies were related to customer service (N=10) and computer speakers (N=10). Certain domains are aligned with particular platform types, indicating specialised applications. For example, the smart home assistance domain was associated with devices such as Amazon Alexa [204, 25, 160, 127, 53, 64], the INSPIRE system [272], and devices not specified [134]. Similarly, the automotive domain was uniquely paired with in-vehicle voice-user interfaces in simulated environments [13, 128], and field settings [278].

Some studies involved more complex combinations of platforms presented in other domains, indicating a broader range of functionalities and user interactions. Studies within the public assistance domain utilised smart speakers [131] or robots [149] in their SDSs. By contrast, studies within the customer service domain utilised telephones [30, 124, 231, 48, 90, 39] or computer speakers [117, 134]. Further emphasising the diversity of platforms, studies within each domain of the personal assistance and gaming categories featured the integration of at least three distinct types of platform. Specifically, in the personal assistance domain, individual studies focused on utilising either computer speakers [288, 232, 179, 180, 192], mobile devices [173] or smart speakers [143, 84]. In the gaming domain, studies employed either computer speakers [129, 286], mobile devices [46]or robots [72] in their SDS designs.

Other dimensions providing further insight into the context and interaction of SDSs included interactional sessions, language, interaction types, interface personification, motivation, and collaboration goals. The majority of studies (N=30) examined interactions within a single session in which participants engaged in one conversation with an SDS. A smaller number (N=5) explored multi-session interactions ranging from 1–5 weeks, predominantly in real-world environments with families, (e.g. [46, 160]). One study [284], which did not involve direct interaction with SDSs, investigated user classification of error messages and preferences for error responses using categorisation and sorting surveys. All but one study employed single-language SDSs, the exception being[134], which implemented multilingual SDSs across various regions to study differences in user responses to system breakdowns in different linguistic contexts. Most studies focused on single-individual interactions with SDS (N=32), while four studies explored interactions involving multiple people with SDSs [204, 46, 25, 160]. Regarding interface personification, which refers to the degree to which SDSs exhibit visual or physical human-like features [119], most studies (N=30) used disembodied SDSs. A few studies used embodied SDSs, employing animated systems [46, 286], avatars [192, 129, 260], and robots [72], mainly in the gaming domain, with one study each in the personal assistant [192] and customer service domains[260].

Regarding collaboration goals, SDSs were either goal-oriented (N=27), aiding users in achieving specific objectives, or non-goal-oriented (N=9), focusing on broader engagement without a defined end goal. Participant motivation for engaging with SDSs was broadly classified into two categories, depending on the task at hand: assistance/facilitation (N=25) and entertainment/engagement (N=11). Roles of SDSs were primarily as facilitators (N=32), with only three studies [286, 129, 72], examining SDSs as peers, all within the gaming domain. Details of methodologies and repair strategies are presented in the literature matrix (See Table 3.3).

Table 3.3: A Literature Matrix of Repair Strategies in SDSs

Authors/Year	Repair Strategies											Research Methodologies									
	SDSs						Users					Approach			Corpus		Experiments		Methods		
	Conf.	Info.	Social	Solve	Ask	Disc.	Clar.	Info.	Modu.	Syst.	Cont.	Qual.	Quan.	Mixed	Filed	Lap	Between	Within	Ques.	Interview	WoZ
Huang and Sénécal [2023])[117]			•										•				•		•		
Clausen et al. [2023][53]			•										•				•		•		•
Xu et al. [2023][278]		•	•										•					•	•		•
Meck et al. [2023][163]	•				•									•				•	•	•	•
Wang et al. [2023][260]			•										•				•		•		
Mavrina et al. [2022][160]							•		•				•			•			•		
Kisser and Siegert [2022][131]	•	•					•		•			•			•						
Zargham et al. [2022][286]	•				•									•			•		•	•	•
Motta and Quaresma [2021][173]				•			•	•		•	•			•				•		•	
Cuadra et al. [2021][64]	•													•			•		•	•	
Lin et al. [2021][149]	•			•	•									•			•		•	•	•
Myers et al. [2021][180]							•	•		•		•				•					
Yuan et al. [2020][284]	•		•											•			•		•	•	
Kim et al. [2020][129]							•	•						•		•		•	•		•
Ge et al. [2019][84]			•											•				•	•	•	•
Beneteau et al. [2019][25]	•						•		•					•		•				•	
Lee et al. [2019][143]		•			•								•					•	•		•
Kim et al. [2019][128]				•	•		•	•	•					•				•	•	•	•
Kiesel et al. [2019][127]	•		•		•								•			•			•		
Porcheronet al. [2018][204]							•	•			•	•				•					
Cheng et al. [2018][46]							•				•	•				•					
Myers et al. [2018][179]							•	•	•	•	•			•		•				•	
Engelhardt et al. [2017][72]	•													•	•				•		•
Opfermann and Pitsch [2017][192]							•	•				•			•			•			•
Kraljevski and Hirschfeld [2017][134]									•	•			•			•		•			•
Wolters et al. [2010][272]				•										•		•			•		•
Zgorzelski et al. [2010][288]	•			•	•								•				•		•		•
Bohus and Rudnicky [2008][30]	•			•	•									•	•		•		•		•
Litman et al. [2006][152]							•	•	•			•			•						
Raux et al. [2006][213]	•			•	•								•		•						
Bulyko et al. [2005][39]			•		•		•			•				•	•			•	•		
Choularto et al., (2004)[48]							•	•				•			•						
Gabriel Skantze [2003][232]	•				•									•			•		•	•	•
Goldberg et al. [2003][90]	•		•		•		•			•	•	•			•						
Kazemza et al. [2003][124]							•	•		•	•	•			•						
Shin et al. [2002][231]	•			•	•		•	•		•	•	•			•						
Sum (N=36)	15	3	9	8	13	0	18	11	8	8	7	10	10	16	11	10	10	10	21	11	16

Note:(Conf) Confirmation; (Info) Information; (Disc)Disclosure; (Clar) Clarification; (Modu) Modulation; (Syst) System-centric; (Cont) Contextual& Affective; (Qual) Qualitative; (Quan) Quantitative; (Ques) Questionnaire

3.3.2 System-Repair Strategies

To categorise repair strategies in SDSs, system and user responses to SDS breakdowns reported in the included studies were thoroughly examined to identify potential categories. Where applicable, existing categories from frameworks presented in previous research (e.g., [26, 30, 179, 173]) were adapted to maintain consistency and address any observed gaps. Importantly, the criteria for the categorisations were based on the function and objective of each response within SDS interactions. This approach led to the merging of categories where similar objectives were pursued by the repair strategies. For example, under the category of 'Modulation strategies' which aim to resolve communication breakdowns by altering vocal characteristics, three primary user-repair strategies were merged: adjusting speech rate, enunciating clearly, and modifying speech volume. Each strategy shares the common goal of enhancing speech intelligibility and effectiveness in response to a breakdown.

Following this methodological framework, system-repair strategies for SDSs were classified using a framework derived from the systematic analysis by [26], which examined recovery strategies in CAs. This framework was particularly relevant for the review as it offered well-defined categories for systematically categorising repair strategies in both text and voice interactions, which is crucial for the focus on voice interactions in SDSs. Using this framework, system-repair strategies were categorised into six distinct categories: Confirmation, Information, Solve, Social, Ask, and Disclosure.

3.3.2.1 Confirmation

In this review, Confirmation strategies in SDSs are not designed to completely repair a failed conversation but are instead intended to acknowledge or ignore errors. Regarding error acknowledgement, SDSs utilize both explicit and implicit strategies for verification or rejection. Explicit error acknowledgement involves the SDS admitting its misunderstanding with phrases like "I don't understand," which prompts users to guide the recovery process. In contrast, implicit error acknowledgement involves the SDS either staying silent or responding in ways that reflect the miscommunication. Notably, explicit acknowledgements often serve as precursors to further repair strategies, such as requesting a rephrasing or clarification of the user's utterance, details of which are explored later in this section.

The error ignoration strategy, on the other hand, sees SDSs overlooking errors and continuing along a predefined conversational path. This strategy may involve posing new questions [30, 13, 288, 232] or taking steps towards goal completion [127, 286, 131, 72, 232], even in the face of misunderstandings. For example, [72] found that employing the ignore repair strategy, where a robot overlooks its failure and proceeds with a simple 'OK', resulted in enhanced perceptions of perceived intelligence and animacy (as in an animated system) compared to alternative strategies such as apologising and problem-solving.

3.3.2.2 Information

The Information strategy transcends basic confirmation, as SDSs strive not only to confirm but also to elucidate the situation by providing potentially useful messages or feedback about the error. This recovery strategy appeared in only three of the reviewed studies [131, 278, 143]. For example, Xu et al. [278] designed a combined repair strategy that integrated apology, explanation, and promise to mitigate the adverse effects of automated driving system failures. Similarly, Lee et al. [143] employed what they termed an 'elaborated feedback' strategy to develop common ground (e.g., "I don't understand what you said. It is noisy around here."), which positively influenced user acceptance and usability of VUIs.

3.3.2.3 Social

The Social recovery strategy in SDSs integrates human-like behaviours to address and repair breakdowns in communication. Grounded CASA paradigm, these strategies embody a range of human-like qualities through various responses, including apologies and compensatory strategies [182]. For instance,[84] explored user preferences between two different strategies, apology ('I'm sorry') and humour ('My IQ is still recharging'), across distinct failure contexts. The results indicated a pronounced preference for smart speakers that offered apologies in both instances. On the other hand, in the same study, humour as a repair strategy proved counterproductive, especially when smart speakers misunderstood user requests.

3.3.2.4 Solve

The Solve strategy aims to actively resolve breakdowns by offering concrete solutions. This strategy is goal-oriented and distinct from that of the Information strategy, as it surpasses the mere elucidation of breakdown causes. Instead, it encompasses the delivery of supplementary information and guidance tailored to help users resolve breakdowns effectively. SDSs can augment their support by providing users with a range of potential solutions in the form of options. Additionally, SDSs can introduce pre-defined speech structures or templates designed to assist users in their recovery efforts. For example, in instances where users use language or expressions beyond the SDS's processing capabilities, the SDS presents a set of understandable utterances to ensure mutual understanding between user and system.

Based on a study of conference room reservations, [30] assessed the effectiveness of 10 strategies in addressing errors of non-understanding. Among these, "Full Help" and "TerseYouCanSay" strategies where the system guides the user on what to say next, emerged as the most effective. When the system offers help, which includes sample responses related to the conversation (e.g. SDS:

'Have you had lunch yet? You can answer whether or not you have eaten.'), users can discover more effective ways (from the system's viewpoint) to express their intent and explore further dialogue options [149].

3.3.2.5 Ask

The Ask strategy in SDSs transfers the repair responsibility to the user via three main techniques. Initially, the SDS may repeat the inquiry to give the user another chance to articulate their request [149, 143, 30, 90, 39, 128, 288, 213]. Alternatively, the SDS could ask the user to rephrase their utterance [213, 30, 90, 39]. Lastly, if the first attempt was unclear or incomplete, the SDS might request additional input to formulate an adequate response [13, 213, 286, 30, 128, 232, 72]. For instance, [39] examined the NIST 2000 Communication Evaluation, which included nine mixed-initiative telephone dialogue systems. They found that employing repetitions as a system-repair strategy resulted in significantly higher rates of frustration compared to rephrasing.

3.3.2.6 Disclosure

Similar to the Information strategy, the disclosure strategy is designed to educate users on how to navigate potential communication issues. The key distinction is that disclosure aims not to directly address or repair breakdowns but to set realistic user expectations by requiring the SDS to identify itself as a computational entity and to openly state its capabilities and limitations. This approach helps manage user expectations by clarifying what the system can and cannot do, thereby potentially preventing some breakdowns before they occur. In this review, the dataset reveals an absence of studies focusing on the broader implications of such disclosure strategies in SDSs, echoing the findings of previous reviews and emphasising the need for more comprehensive research in this domain [26]. For a depiction of system repair strategies following a dialogue breakdown, see Figure 3.4.

3.3.3 User-Repair Strategies

Based on an aggregation of evidence from the included studies, the author proposes a framework of user-repair strategies for interactions with SDSs. The author developed this framework by synthesising and adapting categories from the frameworks presented in the reviewed studies, to address the need for a more comprehensive framework. While prior studies have examined specific forms of user repair in isolation, this review is the first to consolidate these behaviours into a unified, user-centred framework of repair strategies in SDS interactions. Due to the complexities and unexpectedness of user reactions to SDS breakdowns, user-repair strategies can be categorised into five broad strategies: Clarification, Information adjustment, Modulation, System-centric and Contextual and Affective Responses.

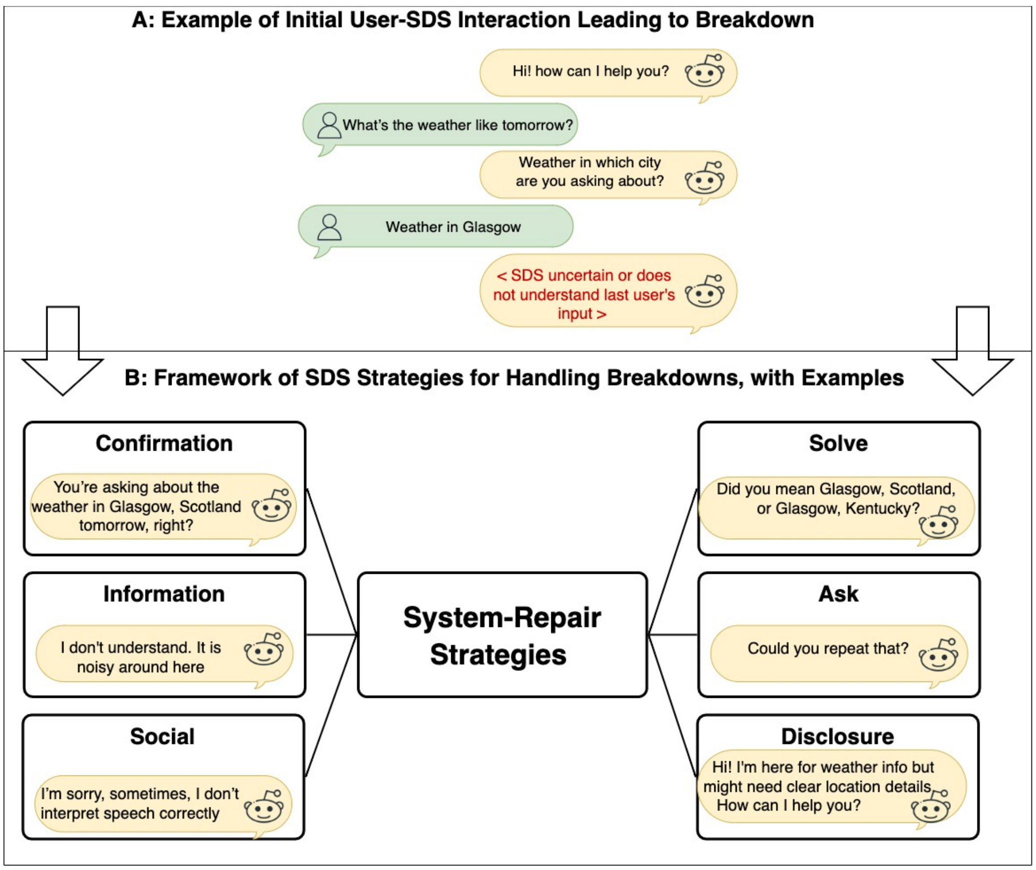


Figure 3.4: **This is a flowchart depicting an initial interaction between a user and SDS that results in a communication breakdown (A). It is followed by a comprehensive framework of system-repair strategies (B). The framework includes examples of each strategy in action, demonstrating how the SDS can repair communication breakdowns.**

3.3.3.1 Clarification

Clarification strategies play an important role in user-repair in SDSs, focusing on ensuring clarity and correctness of the intended message. These strategies are the most widely employed by users, with repetition and reformulation being the primary forms. Repetition is a direct and common strategy, especially when users feel that the SDS has misunderstood their input. Users often repeat their previous utterances, hoping for better system comprehension (e.g. [231, 39, 25]). For example, [25] investigated family interactions with Alexa and found that children frequently use repetition as their initial attempt to repair communication breakdowns.

Reformulation involves a range of techniques to enhance communication when initial attempts are unsuccessful. Central to this strategy is rephrasing. Users alter the original utterance to make it more comprehensible to the SDS while maintaining their original intent [25]. This can include simplifying sentences, using synonyms or restructuring complex requests into simpler formats [25]. Users often rephrase by replacing verbs or nouns or by modifying sentence structure [129, 200]. Four of the reviewed studies investigated user-repair strategies specifically focused on lexical adjustments (when users modify vocabulary), semantic adjustments (when users alter the meaning of the original statement) and syntactic adjustments (when users restructure sentences; e.g. [160, 118]).

3.3.3.2 Information Adjustment

Information adjustment strategies modify the quantity or precision of the information provided. This category encompasses three strategies: addition, subtraction and correction. Addition occurs when users provide additional details to repair a breakdown, particularly when the initial utterance lacks sufficient information for the SDS to comprehend the user's intent accurately [173, 180]. For example, a study on the sequence patterns of users interacting with VUIs [180] found that users often add more details to their utterances to clarify their intent when they encounter a breakdown with a new task. Subtraction strategies involve users simplifying their utterances by eliminating confusing or unnecessary details. Similarly, Motta and Quaresma [173], in a study of how different types of voice assistants (e.g., Siri and Google Assistant) relate to user-repair strategies, found that users adjusted their utterances by adding or subtracting elements during breakdowns, reflecting exploratory behaviour.

Correction strategies occur when users directly repair specific inaccuracies interpreted by the system. This user-repair strategy is particularly relevant when the SDS misinterprets part of the user input, leading to unsatisfactory responses (e.g., [30, 231]). For instance, if a user requests 'Play relaxing music', but the SDS plays rock music instead, the user may correct this by stating, 'No, play relaxing music.' Unlike addition, which provides extra context, or subtraction, which simplifies the input, correction specifically targets inaccuracies in the SDS's understanding. Some studies (e.g. [124, 231]) show that correction often manifests as a contradiction, wherein the user interrupts or 'barges in' to correct and guide the SDS's incorrect process. For example, Kazemzadeh et al. [124] analysed the acoustic features of user-repair strategies and found that contradiction, a form of correction user strategy, was associated with higher pitch and frequency compared to other responses, such as repetition.

3.3.3.3 Modulation

Modulation strategies, which aim to resolve communication breakdowns by altering vocal characteristics, encompass three primary user-repair strategies: adjusting speech rate, enunciating clearly and modifying speech volume. Recent studies, particularly those produced in the past five years, have shown increased interest in exploring modulation strategies across various domains and settings, primarily through the analysis of lab-collected corpora (e.g., [179, 129, 160]). These studies found that in correction and non-correction dialogue acts, the prosodic elements of user utterances typically shift towards hyper-articulated speech patterns. For example, [129] investigated frequent breakdowns and user reactions to them in voice-based dialogue interfaces. This study revealed that attempts by many participants to amend errors by repeating phrases more slowly or with clearer enunciations often resulted in higher rates of error recognition by the SDS.

3.3.3.4 System-Centric

System-centric strategies are tactics employed by users that reflect their understanding of the capabilities and limitations of the SDS. In response to breakdowns, users often emphasise specific keywords that activate certain functions within the SDS. Utilising existing telephone dialogue systems, Bulyko et al. [39] explored how users react to re-prompting – when the SDS repeatedly asks for the same information. They found that commands such as 'Back up', 'Start over' and 'Scratch that' are processed more efficiently by the SDS than would be the case by simply repeating phrases. Another system-centric strategy is interface switching. When an SDS supports alternative input modes, users may opt to switch to typing [173] or using a touchscreen [179] to overcome communication breakdowns.

3.3.3.5 Contextual and Affective Responses

Contextual and Affective strategies encompass a broad range of repair strategies that extend beyond the four previously described categories and comprise a diverse and complex class of user behaviours. Within this category, three primary user-repair strategies for SDS breakdowns can be identified: contextual adjustment, emotional expression and decision to disengage. Contextual adjustment strategies involve alterations to the user's approach or surroundings to enhance interaction with the SDS and repair breakdowns. For example, some studies (e.g., [46, 204]) highlighted how users proactively adjust their environment, such as a child repositioning a tablet following parental advice or participants minimising ambient noise to aid voice recognition – actions that reflect an instinctive adaptation to improve communication with the system.

Emotional expression strategies encompass users' verbal and non-verbal reactions to SDS breakdowns, often reflecting frustration [179]. This can range from swearing, vocal interjections like 'ah' and laughter, to non-verbal cues such as eye-rolling [124, 179, 192]. Finally, decisions to disengage are characterised by users stepping back from interaction with SDSs, which may involve quitting, expressing a desire to stop or ending an interaction [231, 179, 173, 118]. These behaviours signal a user's intention to cease efforts to overcome breakdowns, which is frequently associated with confusion and frustration, leading to task abandonment. [179]. The literature underscores the importance of these strategies as they mark the moments when users choose to stop interacting with the task. For a depiction of user-repair strategies following dialogue breakdown, see Figure 3.5.

While the proposed user-repair framework distinguishes between five categories of user repair strategies, these categories should not be understood as mutually exclusive. In practice, users often combine or sequence multiple strategies when attempting to resolve communication breakdowns [133, 284]. For example, adjustments to informational content may co-occur with vocal modulation, and strategies may shift dynamically across interaction turns. The framework is therefore intended as an analytical lens for organising reported behaviours in the literature, rather than as a rigid taxonomy of independent actions.

3.4 Discussion and Future Directions

3.4.1 System and User Repair Strategies in SDSs

3.4.1.1 Frameworks of repair strategies

This review provided an overview of how SDSs and their users repair communication breakdowns. From the 36 included papers, frameworks for both SDS and user repair strategies are presented. For system-repair strategies, the framework categorises system-repair strategies into six categories: Confirmation, Information, Social, Solve, Ask and Disclosure. This comprehensive categorisation directly addresses RQ1a by identifying the specific types of strategies SDSs employ to repair conversation breakdowns. The majority of system-repair studies focused on Confirmation, Ask, and Social strategies. However, the review identified only a limited number of studies (N=3) focusing on Information repair strategies. Importantly, there is a complete absence of studies investigating the disclosure strategy as a means to repair communication breakdowns within SDS. The disclosure strategy, which involves the SDS openly acknowledging its limitations and capabilities to users, could be crucial for enhancing user interaction by setting realistic expectations. Evidence suggests that a conversational agent's self-disclosure as a computer artefact can efficiently set user expectations by acknowledging its non-human intelligence [225].

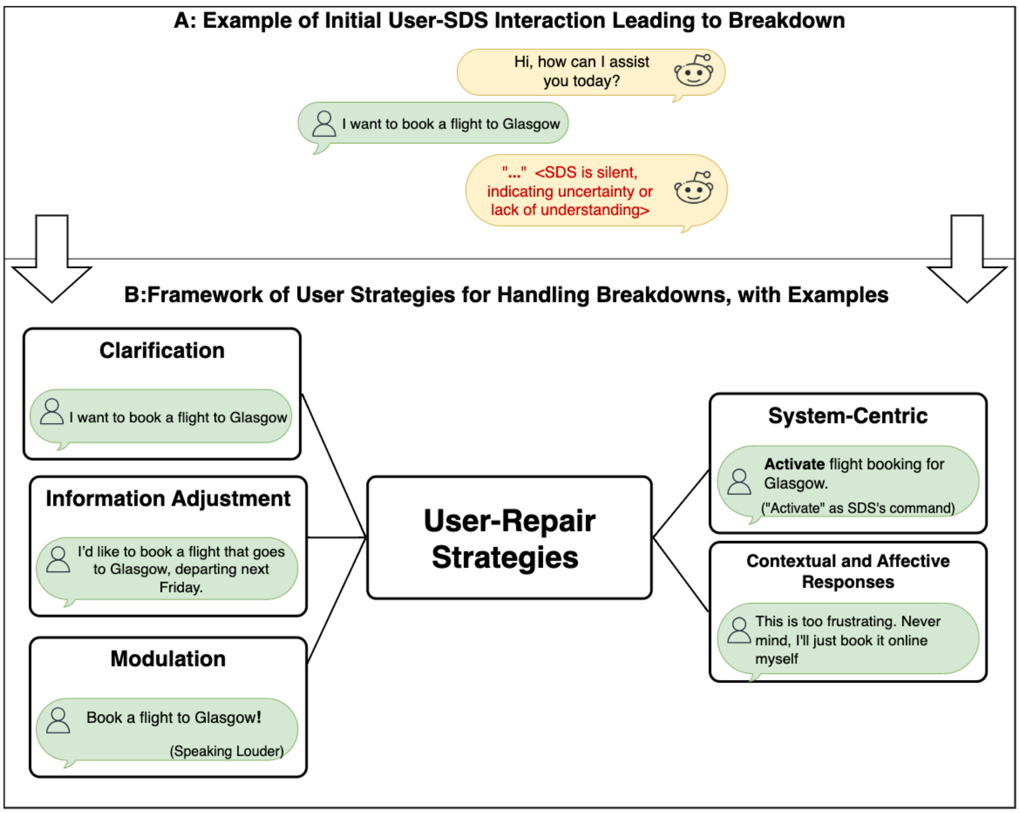


Figure 3.5: **This flowchart highlights user-initiated repair strategies in response to SDS communication breakdowns (A). It is followed by a comprehensive framework of use-repair strategies (B). The framework includes examples of each strategy in action, demonstrating how the user can repair communication breakdowns.**

Similarly, this review answers RQ1b through the development of a framework for user-repair strategies, divided into five distinct categories: Clarification, Information, Modulation, System-centric, and Contextual and Affective Responses. This framework constitutes a key contribution of the review by providing the first comprehensive synthesis of user-repair strategies across SDS studies. This framework highlights the active role that users play in the repair process by employing various strategies to overcome communication breakdowns. In the review of studies on user-repair strategies, the author observed a balanced distribution across all strategy types. However, while Clarification and Information strategies were more common, there was a notable gap in the exploration of Modulation, System-centric, and Contextual and Affective strategies. Expanding research to cover these under-explored strategies will provide a deeper understanding of strategies for addressing breakdowns in SDSs, providing a road map for future studies.

Furthermore, the literature in the review used inconsistent terminologies for repair strategies. For example, the 'error ignoration' system-repair strategy, classified under the confirmation category, involves an SDS ignoring a breakdown and continuing along a predetermined path. This strategy is named differently in the reviewed studies, though they refer to the same concept and mechanism of repair methods such as 'ignore' [72], 'task-related question' [232], and 'anticipatory error handling' [286]. Thus, for future work, adopting the frameworks proposed in this review will promote a standardised terminology and conceptual foundation within the research community and provide uniform references to repair strategies.

3.4.1.2 Interaction of Repair Strategies

The reviewed studies have demonstrated that researchers often employ a combination of system-repair strategies to repair complex communication breakdowns and improve user experience. Research shows the effectiveness of system-repair strategies is stronger when combined with other strategies as a multiple-repair strategy [133, 284]. For instance, Xu et al. [278, p: 5] employed a multi-repair strategy combining Social and Information repair strategies. The system used an integrated prompt: "I am very sorry [apology]. There is a fault and a take-over is needed due to a possible out-of-memory error [explanation]. We will ensure this does not recur [promise].". Similarly, Goldberg et al. [90] found that an apology as a social system-repair strategy led to lower word error rates and lower frustration when combined with rephrasing as a confirmation strategy. This combination of system-repair strategies highlights the complexity of their interplay. Moreover, the integration of LLMs such as ChatGPT-voice and Alexa-LLMs offers fertile ground for exploration. LLMs, known for their robust NLP capabilities, could significantly refine the adaptability and effectiveness of both system and user repair strategies [262, 279]. Future research could explore how these models can be customized to enhance interaction dynamics within SDSs, particularly in terms of real-time responsiveness and contextual understanding. This suggests the potential value of future studies considering the frameworks identified in this scoping review to identify and articulate these complex combinations more precisely.

Regarding user-repair strategies, certain strategies present more challenges to SDSs than others, with users frequently adopting clarification strategies, such as repetition [70, 231] and modulation strategies, such as changes in voice pitch or volume [160, 81, 109]. However, these strategies are often the least successful at repairing breakdowns and can negatively impact the SDS's ability to recognise voice commands accurately [90, 81, 70]. Importantly, the selection of user-repair strategies is not random but is significantly influenced by the system's repair strategies [173, 152, 204, 205]. This indicates a complex interaction between user behaviour and system feedback, where system prompts directly shape the user's choice of repair strategy. This dynamic interaction has been explored in only a few studies (N=8) within this review (e.g., [173]). For future research, it will be essential to investigate the development of SDSs that guide users away from less successful user-repair strategies and towards more effective ones. In conducting this review, the aim was to map the landscape of strategies employed, rather than to evaluate their effectiveness on user experience or SDS performance.However, during the review, notable variations were observed in how similar strategies were evaluated across studies, underscoring the diversity of research methods and contexts. For example, while some studies (e.g., [260, 278, 284]) suggest that the effectiveness of apologies employed as social repair strategies may be limited, others (e.g., [72, 84, 90]) report significant positive impact on users' perceptions of SDS. This divergence hints at the complex ways in which users interpret apologies from SDS, potentially as a standard response to failures [284]. Such observations reveal the nuanced and context-dependent effectiveness of repair strategies [278], suggesting fertile ground for future research. Further examination via systematic review or meta-analyses, could provide crucial insights into conditions influencing strategy effectiveness.

3.4.2 Research Methods and Defining Characteristics of SDSs

An overview of research methods employed and key characteristics of SDSs in the reviewed papers is provided here, addressing RQ1c. As the papers showcase a diversity of research methods, the author classifies methodologies into three broad approaches: qualitative, quantitative, and mixed-methods. Each employs distinct techniques and data collection methods. The findings indicate that a notable number of qualitative studies (N=9) exclusively utilised corpus analysis, demonstrating a well-established foundation of research employing this method. However, it is important to note that the majority of the studies analysing field-data corpora from real-world settings were published prior to 2009. Their findings are grounded in older corpora, often focused on telephone-based platforms. For instance, Litman et al. [152] utilised the TOOT train information corpus from 1999. Given the significant advancement of SDSs, which are revolutionising the dynamics of user-system interaction, contemporary research may yield distinctly different results [185]. The evolving capabilities of these systems from real-world settings suggests the need for renewed investigations to better understand current repair strategies in SDS.

The review also reveals a trend towards combining research techniques, which underscores the complexity of SDS research. Nevertheless, apart from corpus analysis, the findings highlight the absence of other qualitatively-driven method designs such as focus groups, case studies, or ethnography. Such methods could enable an in-depth understanding of user interaction with SDSs in real-world contexts [171]. These methodological patterns also intersect with broader contextual factors, particularly cultural background and interaction language, which remain under-explored in existing SDS research and are discussed in the following subsection.

3.4.3 Cultural Background and Interaction Language

Beyond methodological considerations, the findings of this scoping review highlight cultural background and interaction language as critical yet under-explored dimensions in the study of repair strategies in SDSs. In particular, most existing work focuses on English-language interactions conducted by native speakers in Western contexts (e.g., [179, 46, 204]), leaving limited understanding of how repair strategies are perceived and enacted when users interact in a non-native language. This narrow focus highlights a gap in understanding regarding the potential influence of cultural and linguistic diversity on the acceptance and evaluation of SDSs. For instance, Rau et al. [212] illustrated that people prefer robots with communication styles familiar to their culture. Similarly, [134] found that speakers' responses to communication breakdowns varied significantly by language, with 55.3% of Italian speakers but only 9.6% of Turkish speakers using a special keyword as a system-centric repair strategy. Recently, Galbraith [82] investigated the impact of interaction language, specifically examining user reactions to the use of "huh?" as a system-repair strategy. Findings revealed that English speakers generally found this approach somewhat acceptable whereas Spanish speakers found it unacceptable. Future research should thus be more inclusive of diverse cultural factors (e.g., [114]), offering a broader and more global perspective on the design and effectiveness of repair strategies in SDSs.

In this review, it was also observed that most studies do not explicitly ground their research in established theories despite the complexity of system–user interactions. These interactions often mirror human face-to-face interactions, as demonstrated by models based on Human-Human Interaction (HHI) theories [135]. Notable theories employed by such models include Common Ground Theory [163], focusing on shared knowledge and assumptions; Justice Theory [260], examining fairness in interactions; and Humour Theory [53], exploring the role of humour. Additionally, Politeness Theory provides a comprehensive framework for exploring and understanding interactions between humans and SDSs, offering insights into managing social repair strategies such as apologies, warmth, and humour to mitigate actions or statements that challenge an individual's dignity or self-esteem [84, 39, 117]. Exploring underutilised aspects of these theories, such as expressing solidarity, exerting power, or employing indirectness to soften messages [28, 168, 236], could reveal novel ways for SDSs to more effectively repair communication breakdowns.

There is substantial potential to advance research in SDSs by applying these theories in more innovative ways. For instance, while Common Ground Theory has traditionally emphasised the importance of confirmation and questioning [163, 17], future studies could investigate how variations in explicitness of confirmations influence user satisfaction and effectiveness across different cultural contexts [104]. Politeness Theory [38] and Hall's Theory [103] could be extended to explore how cultural variations in perceptions of politeness directly shape the design and effectiveness of repair strategies in SDS responses. Such approaches could greatly enhance the interactivity and responsiveness of SDSs, leading to richer, more adaptive user experiences. Taken together, these gaps in the literature motivate the empirical focus of the following chapters 5 and 6. As a starting point for future research, the following research questions are proposed:

- How do user characteristics, such as gender, cultural background, and language proficiency influence perceptions of SDS performance in error scenarios across various contexts?
- How can multi-theoretical principles such as those used in HHI impact the effectiveness of repair strategies in SDS and user experiences?

3.5 Conclusion

The objective of this review was to tackle the existing challenges associated with repair strategies in SDSs, specifically addressing the absence of comprehensive conceptual frameworks and standardised terminologies. To this end, the author conducted an extensive scoping review, aiming to consolidate and analyse the repair strategies employed by both SDSs and their users. The author developed comprehensive frameworks for system and user repair strategies in the SDS domain, which provide a foundational basis for future research to explore less investigated areas. In addition to consolidating existing repair strategy frameworks, this review identifies cultural background and interaction language as key yet under-explored dimensions in SDS repair. It highlights important gaps in current knowledge that directly motivate the empirical investigations presented in the subsequent chapters.

3.6 Chapter Summary

While the previous chapter provided the foundational background and theoretical context for this thesis, this chapter presented a systematic scoping review of repair strategies in SDSs within HCI research. The review resulted in two key frameworks: one for categorizing system repair strategies and another for user repair strategies. These frameworks not only organize existing knowledge but also highlight research gaps, particularly in unexplored methodologies and contexts of repair strategies, thereby informing the design of the subsequent empirical studies. In particular, the system repair framework serves as the basis for the experimental and qualitative studies that follow. The next chapter outlines the methodology used in these empirical investigations. Building on these identified gaps, the subsequent empirical chapters investigate how cultural background and interaction language shape user preferences for repair strategies in SDSs.

Part II

Two Empirical Studies on User-Centred Evaluation of Repair Strategies in Spoken Dialogue Systems

Chapter 4

Methodology for User-Centred Empirical Studies

This chapter describes the methodological approach adopted across the two empirical studies investigating user preferences for repair strategies presented in this PhD thesis. While Study 1 (Chapter 5) examines the effect of cultural background using a cross-cultural quantitative approach, Study 2 (Chapter 6) investigates the effect of interaction language and individual-level factors, as well as exploring user perspectives on different repair strategies, using a within-culture mixed-methods approach. Section 4.2 outlines the system-repair strategy framework employed across both empirical studies, developed based on the preceding scoping review in Chapter 3. Section 4.3 provides an overview of the research design, including a detailed description of the goal-oriented task, simulated task scenario, and participant groups.

Section 4.4 introduces the consistent experimental design, based on a pairwise comparison method, used to investigate user preferences for five repair strategies (addressing RQ2 and RQ3a). Section 4.5 describes how individual-level factors were measured, including computer self-efficacy, language proficiency, and prior experience with technology and DVAs (addressing RQ3b). Section 4.6 presents the design of the semi-structured interviews conducted to explore user perspectives on repair strategies and communication breakdowns (addressing RQ4). Finally, Section 4.8 details the preparatory procedures for the quantitative phase, including pilot testing and translation processes.

4.1 Methodological Rationale

While the primary purpose of this chapter is to describe the empirical methods employed in this thesis, it is also important to situate these methods within the broader methodological landscape of research on repair strategies in SDSs. Drawing on the findings of the scoping review presented in Chapter 3, this section reflects on the methodological approaches commonly adopted in prior work, identifies their limitations, and explains how the methodological choices made in this thesis were informed by these observations.

While hypothesis-driven designs are common in experimental research, the present thesis adopts a research-question–driven (exploratory) approach, which is well established within HCI research. This choice is motivated by the limited and fragmented empirical evidence concerning user preferences for repair strategies in SDSs, particularly across cultural and linguistic contexts. In such under-theorised domains, formulating strong a priori hypotheses risks over-specification and may result in confirmatory evaluations that provide limited theoretical or practical insight [92]. Accordingly, the empirical studies presented in Chapters 5 and 6 are structured around clearly defined research questions, enabling systematic exploration of user preferences through predefined experimental designs and statistical models, rather than premature hypothesis testing.

The scoping review revealed that much of the existing research on conversational repair in SDSs has predominantly focused on system-centred approaches, such as corpus analysis (e.g., [90]), task success metrics (e.g.,[232]), and recognition accuracy (e.g., [134]). While these approaches have yielded valuable insights into system performance and technical capabilities, they are less well suited to capturing user preferences, subjective evaluations, and culturally grounded interactional expectations, which are central to understanding user experience during conversational breakdowns.

Moreover, the review indicated that relatively few studies adopt comparative experimental paradigms designed to systematically elicit user preferences across controlled conditions. Studies that explicitly examine preferences for repair strategies often rely on Wizard of Oz methodologies (e.g., [84, 128, 53, 163]) or categorisation-based surveys (e.g., [284]). While these approaches provide important exploratory insights, they offer limited resolution for assessing fine-grained preference differences between competing repair strategies.

In response to these methodological limitations, this thesis adopts a pairwise comparison methodology to examine user preferences for repair strategies. Pairwise comparison has been shown to reduce scale bias and cognitive ambiguity compared to traditional rating-based methods, such as Likert scales [191, 3], while enabling more precise judgments between alternative interactional options. The use of pairwise comparison is particularly well suited to cross-cultural and multilingual research contexts [67, 191], where response styles and scale interpretation may vary across participant groups.

At the same time, this approach entails certain limitations. Preferences elicited through paired presentation are inherently relative to the alternatives presented, meaning that a strategy selected within a given pair represents the most preferred option in that comparison rather than an unequivocally optimal strategy overall. Alternative approaches for measuring user preference, such as absolute rating scales, ranking tasks, or best–worst scaling, were therefore considered. However, prior research indicates that these methods can introduce increased cognitive load, scale-use bias, and cross-cultural response variability (e.g., [191, 67]).

For these reasons, pairwise comparison has been widely adopted as a robust approach for eliciting comparative judgments (e.g. [17, 141, 80, 16]), as it reduces ambiguity and places lower cognitive demands on participants by requiring only relative decisions between alternatives [201, 67]. Further details of the pairwise comparison methodology and its implementation in this thesis are provided in Section 4.4.

While pairwise comparison enables systematic elicitation of user preferences, it does not by itself explain the underlying reasons or interpretations that shape these preferences. To address this limitation, this thesis adopts a mixed-methods approach (Chapter 6) by complementing quantitative preference data with qualitative interviews. The qualitative component enables exploration of how users interpret repair strategies, how interaction language shape these interpretations, and why particular strategies are perceived as more appropriate or acceptable in specific contexts. Further details of the qualitative interviews and its implementation in this thesis are provided in Section 4.6.

In addition, the scoping review (Chapter 3) identified a bias toward English-language interactions and native-speaker populations in prior work, leaving limited understanding of how repair strategies are perceived when users interact in a non-native language or come from different cultural backgrounds. The methodological design of this thesis directly addresses this gap by systematically varying interaction language and cultural grouping, while maintaining a consistent experimental structure to support meaningful and comparable analyses across studies.

4.2 Repair Strategy Framework

In the previous scoping review (Chapter 3), the author introduced two comprehensive frameworks developed through a systematic scoping review: one categorizes system-repair strategies into six distinct types, and the other categorizes user-repair strategies into five types. Since this thesis focuses on how user characteristics influence preferences and perspectives regarding different repair strategies employed by SDSs during breakdowns, the system-repair strategy framework was adopted for the two empirical studies. Specifically, five repair strategies were selected: Confirmation, Information, Social, Solve, and Ask (see Strategy Framework 3.4). These strategies were chosen because they (1) exhibit distinct communicative attribute and (2) focus on the immediate system responses following a conversational breakdown.

Although the original framework also included a Disclosure category, it was excluded from this thesis. The Disclosure strategy is intended to set user expectations by having the system acknowledge its identity, limitations, or capabilities. It is typically employed proactively or in a general onboarding context, rather than as a direct response to specific communication breakdowns. As such, it does not align with the scope of this thesis, which centres on evaluating user preferences for reactive repair strategies i.e., those initiated by the system after miscommunication. Additionally, the current body of literature lacks empirical evaluations of Disclosure strategies in such breakdown contexts, reinforcing the category's limited applicability to comparative user preference studies such as those presented in this PhD thesis.

Each repair strategy embeds a main attribute corresponding to its specific function and communication style during breakdowns. While some overlaps may exist in how these strategies address breakdowns, this thesis focuses on measuring their distinct attribute, which represent the specific functions and communication styles offered by each strategy. Previous research in text-based CAs [27] highlights how functional characteristics, such as formality, empathy, and humour shape user preferences and experiences across different applications. Similarly, in this thesis, the main attributes of repair strategies (such as confirmation, explanation of the causes of breakdowns, and social apologies) define system responses during conversational breakdowns in voice-based interactions. These attributes illustrate the different approaches each strategy uses to repair non-understandings and promote effective interaction. Table 4.1 provides descriptions and main attributes for the five repair strategies investigated in this thesis. The identification and distinction of the main attribute of each repair strategy were first validated through a pilot study, which is described in detail in Preparatory Procedures section (see Section 4.8.1).

Table 4.1: **Descriptions and Attributes of Repair Strategies**

Repair Strategy	Main attribute	Description	Employed Example
Social	Apology	Politely apologizes for the non-understanding	***I am sorry***, *I could not understand you. Could you* ***please*** *say that again?*
Confirmation	Seeking Confirmation	Guesses the user's input and seeks confirmation	*It sounds like you want to travel to Luton.* ***Is that correct?***
Information	Breakdown Explanation	Explains why it could not understand	*I could not understand you* ***because it is noisy here****. Could you say that again?*
Solve	Providing Options	Offers a list of possible options based on partial understanding	*I could not understand you.* ***Do you mean Luton, Sunderland, Liverpool, or non of these?***
Ask	Requesting Repetition	Directly asks the user to repeat the input without additional context or elaboration	*I could not understand you.* ***Could you say that again?***

4.3 Research Design Overview

This thesis adopts a two-study empirical research design to investigate how user characteristics influence preferences and perspectives regarding repair strategies in SDSs. Study 1 (Chapter 5) employs a cross-cultural quantitative approach to examine the effect of cultural background on user preferences. Study 2 (Chapter 6) adopts a within-culture mixed-methods approach to explore the effects of interaction language and individual-level factors, such as computer self-efficacy, language proficiency, and prior experience with SDSs, while also gaining deeper insights into user perspectives through qualitative interviews.

Both studies were built around a consistent experimental design, which strengthens the comparability and integration of results across studies [219]. Specifically, both studies used a simulated goal-oriented task that created controlled communication breakdowns between the user and the system. This prompted the system to respond using one of five pre-defined repair strategies. Participants then evaluated the system's repair responses through a pairwise comparison method by selecting their preferred option in each scenario. By applying the same experimental structure across both studies, this design enables a comprehensive understanding of user preferences: capturing both cross-cultural variations (Study 1) and within-group variations related to interaction language and individual traits (Study 2). This consistency also ensures methodological coherence, making it possible to interpret preferences in light of broader sociocultural and contextual communication variables.

Although the simulated SDS interactions used in both studies were voice-only, participants were told that the responses were from a DVA named *Echo*, such as Amazon Alexa or Google Assistant. This decision was intentional as DVAs are a familiar and widely used example of SDSs [122]. Using DVAs as a reference point helped participants to better engage with the task, understand the interaction context, and evaluate the repair strategies meaningfully. Referring to a familiar system increased ecological validity and ensured that participants could imagine realistic interactions [33, 34, 24]. Although this approach reflects the nature of many real-world SDS applications, participants were not instructed to refer to a specific commercial DVA such as Amazon Alexa or Siri.

In both studies, all experimental tasks and survey instruments were administered online using the Qualtrics[1] platform. In Study 2, qualitative data were also collected through semi-structured interviews conducted via Microsoft Teams to further explore users' reasoning, perspectives, and expectations regarding system repair strategies and communication breakdowns.

[1] https://www.qualtrics.com/

4.3.1 Goal-Oriented Task

The choice of a goal-oriented flight-booking task was a deliberate methodological decision aimed at examining repair strategies under conditions where accuracy and task success are critical. Compared to non-goal-oriented or social interactions, transactional tasks place greater emphasis on grounding, confirmation, and error recovery, making communication breakdowns more consequential for users. Flight booking was selected as a representative medium-complexity task that is familiar to participants.This task context therefore provides an appropriate and controlled setting for evaluating user preferences for different repair strategies. This rationale is further supported by prior work on transactional SDS interactions and repair frequency, discussed below. SDSs can be broadly categorized based on the type of interaction they support: (1) goal-oriented tasks and (2) non-goal-oriented tasks (also referred to as chit-chat communications) [44]. While non-goal-oriented tasks aim to sustain casual conversation and provide entertainment, goal-oriented tasks involve structured interactions in which users seek to accomplish a specific objective. Popular commercial systems such as Apple's Siri and Amazon's Alexa are considered goal-oriented, as users interact with them to fulfil information-seeking or task-completion goals [17].

In this thesis, a goal-oriented task is employed to evaluate participants' preferences for repair strategies in SDSs. As shown in Figure 4.1, the complexity of SDS tasks has progressively increased, from low-complexity tasks involving basic information retrieval (e.g., package tracking or bus schedules) to high-complexity tasks such as customer care and technical support that require problem solving and more dynamic interaction [223]. Transactional tasks, such as managing bank accounts or booking flights, are positioned at a medium level of complexity, as they involve the exchange of structured information to achieve a specific user goal [161, 223]. In the empirical studies conducted in this thesis, participants were asked to complete a flight booking task, which is a representative example of a medium-complexity transactional interaction.

The decision to use a goal-oriented task rather than a non-goal-oriented task is intentional. In transactional tasks, accuracy and clarity are critical, and communication breakdowns have a direct impact on task success. Therefore, it is essential that exchanged information be confirmed and grounded to prevent miscommunication [161]. Furthermore, previous research shows that repair strategies are more frequent in goal-oriented dialogues than in non-goal-oriented, social dialogues [56]. In social dialogues, the primary aim is to sustain engagement, and repair is often less focused on resolving misunderstandings and more on continuing the conversation [17]. Therefore, in goal-oriented tasks where accuracy is prioritized, the benefits of employing repair strategies outweigh the potential drawback of increased transaction time [161].

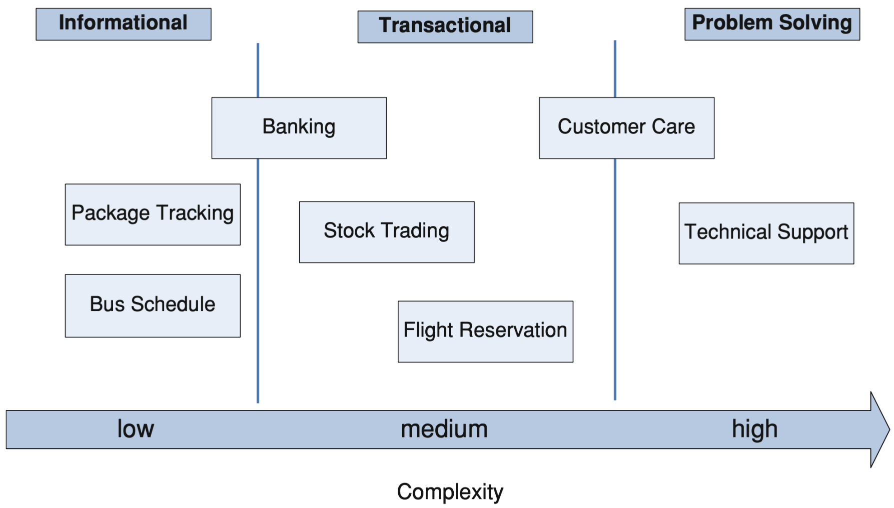


Figure 4.1: **Increasing complexity of spoken dialogue system (SDS) applications, ranging from low-complexity informational tasks to high-complexity problem-solving tasks (Pieraccini and Huerta [223]). The flight reservation task employed in this thesis is categorized as a medium-complexity transactional task, involving structured information exchange to achieve a specific user goal.**

4.3.2 Simulated Work Task Scenario

The goal-oriented task was embedded within the framework of the Simulated Work Task Scenario (SWTS) to enhance realism and maintain experimental control, as recommended by [33, 34]. Borlund [33, 34] describes SWTS as "a short textual description that presents realistic information," providing a meaningful frame of reference for making relevance judgments. Simulated task scenarios offer participants a more relatable and authentic context, thereby increasing the likelihood of external and ecological validity [24].

In the experimental studies, the SWTS served as a reliable background story to engage participants and guide their responses. The implemented SWTS fulfilled two core functions in this thesis:

- **Simulated work task situation:** Participants were placed in a realistic situation where they needed to book a flight for the following day due to important plans, using a voice-based digital assistant named Echo. This setup created a clear and urgent information need, as the task was goal-oriented and required identification of the correct travel destination in order to proceed. Additional contextual details were included, such as Echo's usual role in helping with travel arrangements and the current issue where Echo fails to understand the city name, This leads to a communication breakdown and triggers repair strategies.

- **Indicative request:** After the information need was established, the scenario led to a conversational breakdown when Echo failed to understand the user's intended destination ("London"). Participants were then asked to compare two voice-based system responses. This acted as an indicative request, prompting participants to judge which system response was more appropriate or effective for resolving the breakdown in this context.

In line with Borlund's framework, the scenario clearly presents the source of the information need (flight booking), the environment (interaction with a SDS), the problem (failure to understand the destination), and the objective (to evaluate system responses). As recommended by Borlund, the SWTS was also pilot tested prior to the main study to ensure realism and reliability. Further details of the pilot study can be found in Section 4.8.1. The following graphic shows the SWTS that was presented to participants:

> Simulated Work Task Scenario
>
> **Text-based Scenario**
> **Scenario:**
> You need to book a flight for tomorrow because you have important plans. Echo is your digital voice assistant such as Siri or Alexa. Echo usually helps you with travel arrangements. However, during the conversation, Echo is unable to understand the city you want to fly to, causing a conversational breakdown.
> **Dialogue:**
> **Echo**: "Hello, I am Echo, your digital voice assistant. How can I assist you today?".
> **You**: "I need to book a flight for tomorrow."
> **Echo**: "Understood. Which city are you flying to?"
> **You**: "I want to fly to London."
> *(At this point, **Echo** doesn't understand your input and initiates one of the following responses.)*

4.3.3 Type of Communication Breakdown

Breakdowns in SDSs are generally categorised into two types: misunderstanding and non-understanding [30]. A misunderstanding occurs when the system incorrectly interprets the user's input, often leading to inappropriate actions or responses. In contrast, non-understanding refers to the system's inability to recognize the user's input, prompting the initiation of repair strategies [30] (see Figure 2.2 for comparative dialogue examples).

In the empirical studies, the focus is specifically on a subtype of non-understanding breakdown known as partial non-understanding [234, 232, 233]. Partial non-understanding occurs when the SDS correctly identifies the user's general intent (e.g., intent: booking a flight) but fails to recognize a specific entity required to complete the task (e.g., entity-destination: "London"). In the experimental scenario, the SDS is intentionally designed to simulate a situation in which the confidence level of its speech recognition module for a specific entity ("London") falls below an acceptable threshold [198, 29]. Consequently, rather than proceeding with uncertain or incorrect information, the system initiates a repair strategy immediately. This approach aligns with the principles of Grounding Theory, which emphasizes the importance of immediate and collaborative actions to achieve mutual understanding during dialogue interactions [49, 51]. In accordance with this theory, the system responds promptly to the breakdown by employing one of five predefined repair strategies, aiming to restore mutual understanding and facilitate the completion of the goal-oriented task.

To control for potential biases in evaluating repair strategies, the correct entity ("London") is deliberately excluded from the system's suggestions in repair strategies that involve presenting alternative options (e.g., Solve and Confirmation strategies). Instead, the system offers plausible but incorrect alternatives, such as "Luton", "Sunderland", or "Liverpool". This deliberate design choice reduces the risk of bias (e.g, item priming effects), where users might favour a particular strategy simply because it includes the correct answer ("London") [203]. This controlled manipulation not only reflects realistic interactional failures commonly observed in dialogue systems [234, 232, 233, 17], but also ensures that participants evaluate repair strategies based on their inherent form and function rather than the accuracy of their content alone.

Across the two empirical studies, the goal-oriented task, simulated scenario, and communication breakdown were consistently implemented. This consistency provided a controlled yet ecologically valid foundation for systematically examining user preferences for various repair strategies.

4.3.4 Participant Overview

This thesis involved three participant groups derived from two cultural backgrounds (British and Saudi) and two interaction languages (English and Arabic) to investigate how user characteristics influence preferences for repair strategies in SDSs. These groupings enabled two types of comparative analysis focusing on: 1) cultural background, comparing high-context and low-context cultures, and 2) interaction language comparing native and non-native language use. The participant groups included in this thesis are:

- **British participants (BL1)**, interacting with the system in their native language (English).
- **Saudi participants (SL1)** interacting with the system in their native language (Arabic).
- **Saudi participants (SL2)**, interacting with the system in their non-native language (English)

In Study 1 (Chapter 5), British participants (BL1) represented a low-context culture, while Saudi participants (SL1) represented a high-context culture, with both groups interacting in their native language, contributing to RQ 2. In Study 2 (Chapter 6), Saudi participants evaluated system repair responses presented either in their native language (Arabic; SL1) or in a non-native language (English; SL2), contributing to RQ 3a. In both empirical studies, the Saudi native group (SL1) served as a common baseline for across both studies (see Figure 4.2).

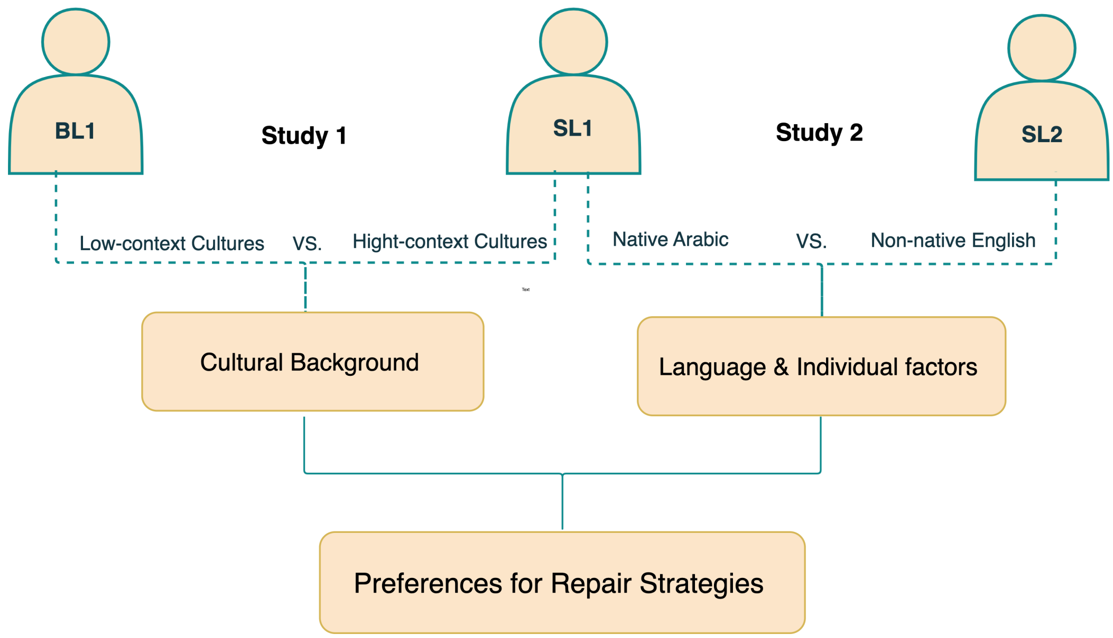


Figure 4.2: **The three participant groups: British English speakers (BL1), Saudi Arabic speakers (SL1), and Saudi English speakers (SL2). The studies examine how cultural background (Study 1) and interaction language (Study 2) influence preferences for repair strategies in SDSs. Dashed lines indicate shared group contributions across both variables.**

To address RQ3b in Study 2 (Chapter 6), Saudi participants (SL1 and SL2) were also included to quantitatively examine the influence of individual-level factors, such as prior experience, computer self-efficacy, and demographic characteristics, on repair strategy preferences. For participants interacting in a non-native language (SL2), self-assessed language proficiency was also considered. The experimental design, including the task scenario, procedures, and survey, remained consistent across both empirical studies, with only the participant groups varying. This design allowed for a structured comparison by isolating the effects of cultural background and interaction language use while maintaining experimental consistency.

Additionally, addressing RQ4 in Study 2 (Chapter 6), the focus shifted from the quantitative investigation of user preferences to a qualitative exploration of user perspectives on communication breakdowns and repair strategies. A subset of Saudi participants took part in semi-structured interviews to examine how they perceive and evaluate different system responses during breakdown scenarios, in line with the study's exploratory and culturally grounded nature.

Participants in the British group (BL1) were recruited from the University of Strathclyde in the UK, while Saudi participants (SL1 and SL2) were recruited from Prince Sattam Bin Abdulaziz University[2], using various recruitment strategies including emails, posters, flyers, and snowball sampling. Detailed demographic information for each user group is presented in the methods section of the corresponding study chapter.

To be eligible to participate in the studies, individuals were required to be 18 years of age or older and to provide informed consent. Participants also had to self-identify their nationality as either British or Saudi. British participants were required to indicate that their native language is English, while Saudi participants were required to indicate that their native language is Arabic. Only those who met these criteria passed the initial screening and were included in the studies. These criteria ensured cultural and linguistic alignment within each group, which was essential for the comparative cross-cultural design of the studies.

An a priori power analysis was conducted using G*Power 3.1 [75] to estimate the minimum required sample size for the repeated-measures design in this PhD thesis. Based on Cohen's [54] and Faul's guidelines [75], parameters were set to a medium effect size (f = 0.25), alpha error probability ($\alpha = 0.005$), statistical power $(1 - \beta) = 0.95$, two groups, and ten repeated measurements. The analysis indicated a minimum $N \approx 22$ per group. The actual group sample in the empirical studies ($N \approx 49$ per group) exceeded this minimum to allow for potential exclusions and increase estimate precision.

All participants were compensated for their time with a £5 Starbucks eGift card, except for those in the interview study, who received a £10 eGift card. Ethical approval for all empitical studies reported in this PhD thesis was obtained from the Ethics committee of the Department of Computer and Information Sciences (approval number: 2781; see Appendix B).

4.4 Consistent Experimental Design Across Studies (RQ2 & RQ3a)

To collect participant preferences for repair strategies, online pairwise comparison experiments were conducted using the Qualtrics platform. Addressing RQ2 and RQ3a, a comparative research design within a mixed factorial framework was adopted to investigate user preferences for repair strategies in SDSs. Each study employed a consistent 2×5 factorial design, involving two independent variables

[2]Prince Sattam Bin Abdulaziz University is located in Saudi Arabia

and five distinct repair strategies. Specifically, Study 1 (RQ2; Chapter 5) examined the effect of cultural background (high-context vs. low-context cultures) as a between-subjects factor, while Study 2 (RQ3a; Chapter 6) explored the effect of interaction language (native language [L1] vs. non-native language [L2]) as the between-subjects factor. Both studies treated the five repair strategies as a within-subjects factor, assessing participants' preferences through pairwise comparisons.

A key consideration in experimental research is ensuring internal validity, which refers to the degree to which observed effects can be attributed to the manipulated variables rather than to uncontrolled confounding variables. As noted in Experimental Methods in Human-Computer Interaction, "the most obvious threat to internal validity comes from confounding variables" [41]. Cairns [41] outlines two general approaches to mitigating confounds: randomization and experimental control. In randomization, participants are randomly allocated to conditions, which reduces the likelihood of systematic differences between groups. In contrast, an experimental control involves directly managing the environment or participant selection to eliminate potential sources of variation (e.g., the same testing environment, participant screening) that could confound results.

This principle guided the design of both empirical studies in this thesis. A consistent experimental design was applied across both studies, including identical task scenarios, simulated communication breakdowns, pairwise comparison format, and experimental procedures. This consistency ensured that participants across both studies encountered comparable system interactions, allowing for meaningful comparison of their preferences for repair strategies.

To examine the effect of cultural background (Study 1), interaction language was controlled: all participants interacted with the system in their native language (Arabic for high-context culture; English for low-context culture), ensuring that interaction language differences did not confound the cultural comparison. Conversely, in Study 2, which focused on the effect of interaction language and individual-level factors, the participants' cultural background was controlled by sampling from a single cultural group (Saudi Arabia), and participants were exposed to one of two language conditions (native vs. non-native) through random assignment. The experiments were administered in two language versions: Arabic for SL1 participants, and English for both BL1 and SL2 participants. Thus, this methodological design incorporates both experimental control and randomization to reduce confounds and strengthen the validity of findings across studies. More details on the translation process is provided later in the Preparatory Procedures section of this thesis (see Section 4.8.2).

Each experimental task began with a text-based scenario describing a goal-oriented task (booking a flight), followed by binary pairwise comparisons of voice-based repair strategies. This design minimizes extraneous influences, such as confusion arising from simultaneous evaluation of multiple strategies [67]. Participants first read a text-based scenario describing a conversational breakdown, followed by two voice responses, each representing a different repair strategy (see Figure 4.3).

Each participant evaluated all five repair strategies through a total of 10 pairwise comparisons, calculated using the combination formula: $n(n-1)/2 = 5(5-1)/2 = 10$. This constitutes a complete paired-comparison design, in which all possible pairs of strategies are presented exactly once. As a result, each repair strategy appeared an equal number of times. With five repair strategies, each strategy was compared once with each of the remaining four strategies, resulting in four evaluations per strategy. This equal exposure ensured that subsequent preference metrics were not influenced by unequal presentation frequency across strategies.

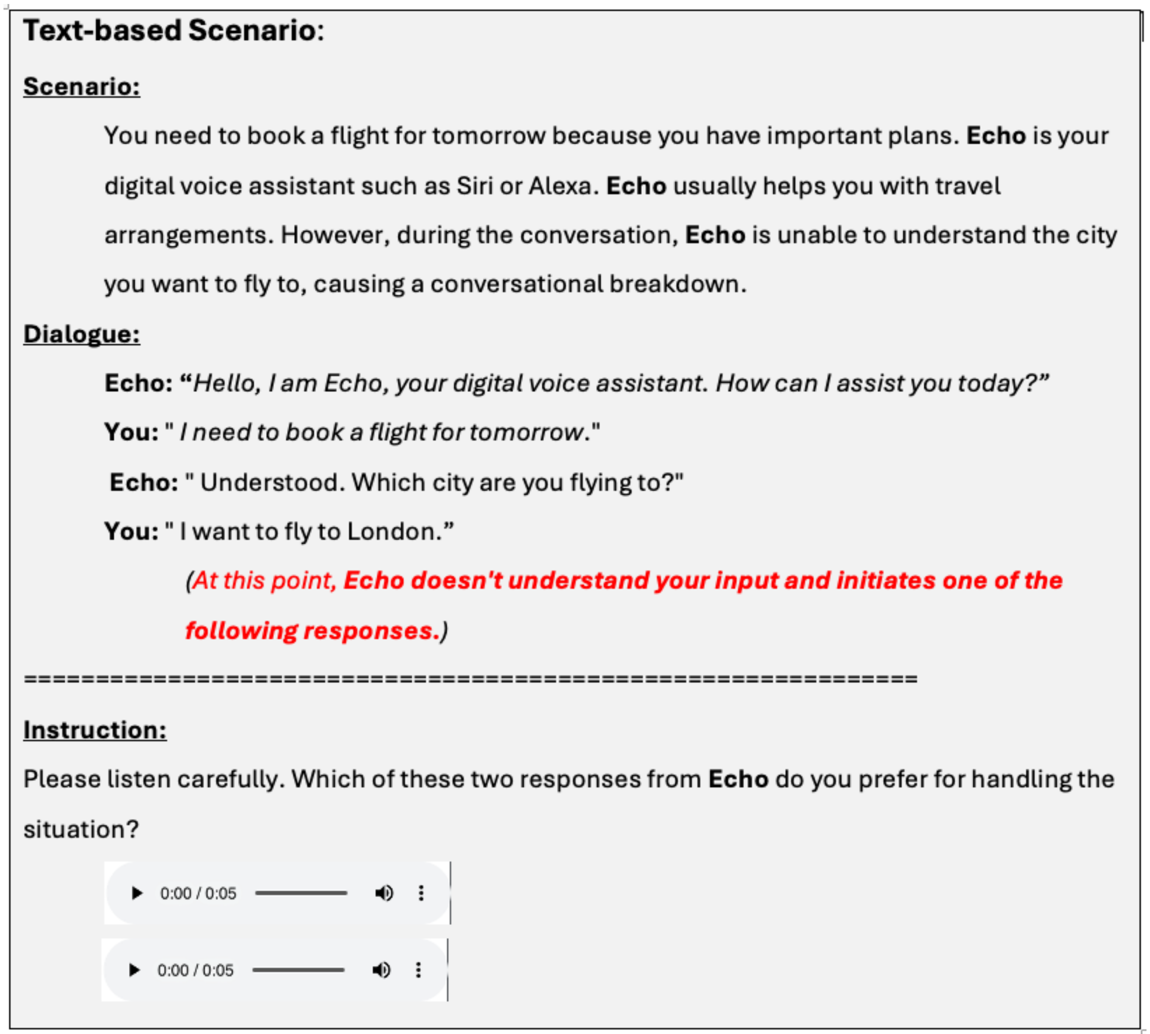
The 10 paired comparisons of the five repair strategies are geometrically represented by Figure 4.4. All comparisons were presented to participants in a random order, and within each comparison, the order of the two repair strategies was also randomized using Qualtrics' built-in randomization tools. This approach was used to minimize potential order bias and effect. Voices were synthesized using Amazon Polly[3], a text-to-speech tool, to generate natural-sounding female speech in both Arabic and English. These voices were selected for several reasons. First, Amazon Polly provides high-quality, human-like intonation and supports both target languages, ensuring consistency in speech delivery across experimental conditions. Second, using Amazon Polly allowed the study to avoid direct associations with commercial voice assistants such as Alexa or Siri, which could have introduced bias due to participants' prior experiences or expectations.

In addition, the choice of female, human-like voices was informed by findings in the literature. Although user voice preferences in SDSs vary depending on factors such as the conversational task [184, 187] and the perceived gender of the assistant [157, 197], multiple studies show that female [199, 237] and human-like voices [184, 278] are generally perceived as more trustworthy, warm, persuasive, and understandable compared to male or artificial-sounding voices. These contribute to more engaging and naturalistic user interactions when evaluating repair strategy preferences in this thesis.

Pairwise comparison experiments are widely used in HCI [17, 141, 80, 16, 159]. The most common application of the pairwise comparison method is for assessing participants' judgments, personal ratings, and generally all forms of preference testing [67, 47]. The use of pairwise comparisons provides more realistic results when compared with Likert scales [191, 3] and is better suited to accurately detecting linguistic and cultural differences [191]. Furthermore, pairwise comparison eliminates the need for linguistic and cultural specific scenarios, taking advantage of straightforward judgments [67, 191].

In the experimental scenario and survey materials, participants were presented with the concept of DVA (e.g., Alexa or Siri) as a familiar and relatable example of SDS. This approach was used to improve participant engagement and ensure clarity during the task, while maintaining consistency with the overarching focus on SDSs in the thesis.

[3] https://aws.amazon.com/polly/



Text-based Scenario:

Scenario:

You need to book a flight for tomorrow because you have important plans. **Echo** is your digital voice assistant such as Siri or Alexa. **Echo** usually helps you with travel arrangements. However, during the conversation, **Echo** is unable to understand the city you want to fly to, causing a conversational breakdown.

Dialogue:

Echo: "*Hello, I am Echo, your digital voice assistant. How can I assist you today?"*

You: " *I need to book a flight for tomorrow*."

Echo: " Understood. Which city are you flying to?"

You: " I want to fly to London."

*(At this point, **Echo doesn't understand your input and initiates one of the following responses.**)*

==

Instruction:

Please listen carefully. Which of these two responses from **Echo** do you prefer for handling the situation?

0:00 / 0:05

0:00 / 0:05

Figure 4.3: **This is the text-based scenario participants encountered in the empirical studies. In this scenario, the DVA (Echo) fails to understand the user's flight destination, leading to a conversational breakdown. The DVA then initiates a repair strategy. Participants listened to two voice responses, each representing a different repair strategy (see Table 4.1), and selected their preferred response.**

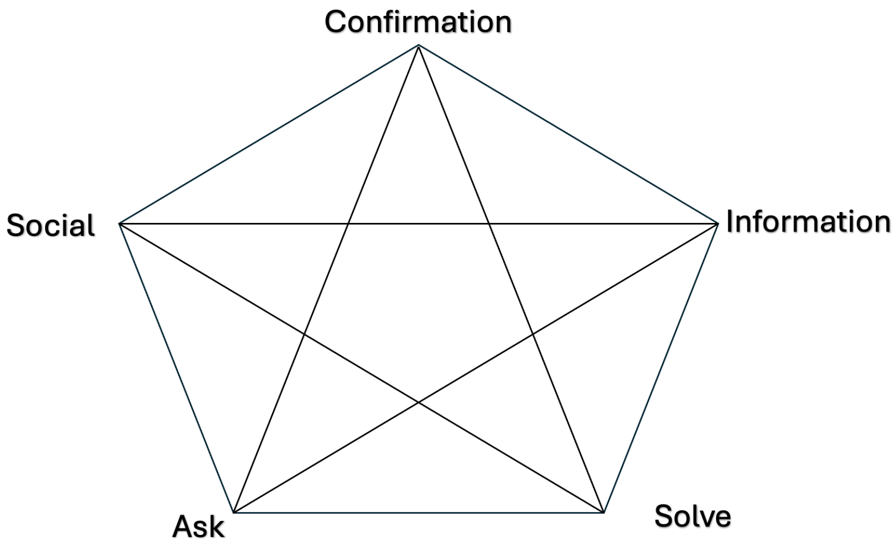


Figure 4.4: **Geometric Representation of Paired Comparisons Among Repair Strategies: Each vertex represents one of the five repair strategies. The connecting lines illustrate all possible pairwise comparisons (10 in total) between the strategies, which were used in the experimental designs.**

4.5 Measurement of Individual-Level Factors (RQ3b)

In addition to investigating the effects of cultural background (RQ2, Study 1, Chapter 5) and interaction language (RQ3a, Study 2, Chapter 6), through a controlled experimental design, this thesis also explores how individual-level factors influence user preferences for repair strategies in SDSs. This section addresses RQ3b (Study 2, Chapter 6) which focuses on the role of user, specific characteristics, namely: computer self-efficacy, language proficiency, and prior experience with technology and DVAs, in shaping these preferences. These individual-level factors were assessed using self-report questionnaires embedded within the experimental tasks in Study 2, which were completed by Saudi participants (SL1 and SL2). By examining these individual differences within a consistent experimental framework and among participants from the same cultural background (Saudi Arabia), this analysis offers a more deep understanding of user preferences. It captures within-group variation shaped by individual traits, helping to isolate their influence on interaction with SDSs.

4.5.1 Computer Self-efficacy

To measure participants' computer self-efficacy, the Computer Self-Efficacy questionnaire developed by Gupta and Bostrom [100] was adopted. Gupta and Bostrom [100] outlined four distinct types of computer self-efficacy, emphasizing the importance of both technology type and task complexity. They argued that computer self-efficacy depends on “both the technology type (general or specific) and the task type (simple or complex),” resulting in a 2×2 matrix (Figure 4.5, p. 75). The four types of computer self-efficacy are:

1. Specific Technology- Simple Task Self-efficacy (SS-SE)
2. Specific Technology- Complex Task Self-efficacy (SC-SE)
3. General Technology- Simple Task Self- efficacy (GS-SE)
4. General Technology- Complex Task Self-efficacy (GC-SE)

Technology Type		***Task knowledge Type***	
		Simple	Complex
	Specific	***Referred to as SS-SE*** *Judgment about the knowledge needed to use a specific software application (e.g., Excel) to accomplish a known task (e.g., create a graph)*	***Referred to as SC-SE*** *Judgment about the knowledge needed to use a specific software application (e.g., Excel) to accomplish a self-conceptualized business/work related task*
	General	***Referred to as GS-SE*** *Judgment about the knowledge needed to use a new/unknown software application to accomplish a known task (e.g., create a graph)*	***Referred to as GC-SE*** *Judgment about the knowledge needed to use a new/unknown software application to accomplish a self-conceptualized business/work related task*

Figure 4.5: Computer Self-Efficacy Constructs, devloped by Gupta and Bostrom [100]

Addressing RQ3b, computer self-efficacy was measured using the Specific Technology–Complex Task Self-Efficacy (SC-SE) scale, as proposed by Gupta and Bostrom [100]. This scale was selected for two main reasons: (1) the focus of this thesis is on interaction with DVAs, which constitutes a specific type of technology, distinct from general technology use; and (2) complex tasks, rather than simple or highly specific skills, better reflect users’ perceived ability and confidence to perform imagined goal-oriented interactions such as booking flight. The scale uses a 7-point Likert agreement format, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree). The full adapted scale are summarised in Table 4.2.

4.5.2 Prior Experience with Technology and SDSs

Prior studies highlight how experience with technology and digital assistants substantially shapes preferences concerning repair strategies in digital interactions [288, 17]. For example, individuals

Table 4.2: Adapted Items for Measuring Computer Self-Efficacy and Experiences with Technology and DVAs

Independent Variable	Adapted Items	Likert Agreement Format
Computer Self-Efficacy [100]	I have mastered the use of Digital Voice Assistants (DVAs): I cannot yet use Digital Voice Assistants (DVAs) as effectively as I would like: I am able to perform tasks well using Digital Voice Assistants (DVAs): It is not yet possible for me to use Digital Voice Assistants (DVAs) at the level I prefer: I think my ability to use Digital Voice Assistants (DVAs) can be improved substantially:	7-point Likert Scale
Experience with Technology [17]	I consider myself an advanced technology user: I am eager to try new technologies:	5-point Likert Scale
Experience with SDSs [17]	I am familiar with Digital Voice Assistances (DVAs): I use Digital Voice Assistances (DVAs) frequently:	5-point Likert Scale

more familiar with chatbots often prefer Confirmation repair strategies as these are frequently present in text-based chatbots [17].

Self-reported prior experience with technology and prior experience with DVAs were assessed using two items each, rated on a 5-point Likert agreement scale (1 = Strongly Disagree to 5 = Strongly Agree). These items were adapted from Ashktorab et al. [17], whose work focused on the influence of prior experience on preferences for repair strategies in chatbot-based interactions. This alignment is particularly relevant, as the empirical studies similarly investigate how individual factors shape repair strategy preferences, but in voice-based interactions rather than text-based settings. The full adapted scale is summarised in Table 4.2. Higher scores on each item indicate a greater level of the respective factor (i.e., more experience with technology or DVAs).

4.5.3 Language Proficiency

This subsection addresses the language proficiency of Saudi participants who interacted with the DVAs (*Echo*) using their second language (**SL2**). Language proficiency, defined as a user's skill level and fluency in a particular language, can influence how users interact with and adapt to SDSs [126, 210, 275]. Prior research shows that higher language proficiency is associated with better understanding and more favourable user experiences when interacting with such systems [126, 210].

To evaluate the English proficiency of **SL2** participants, the Common European Framework of Reference for Languages (CEFR) [189] was used. CEFR is widely used in HCI research for self-assessing foreign language proficiency (e.g., [241, 74, 123, 210]). CEFR defines six levels for each of the four language skills: listening, reading, speaking, and writing[4]. These levels range from basic (A1, A2), to intermediate (B1, B2), to proficient user (C1, C2) [189].

For language proficiency, the focus was specifically on the listening skill, as the primary task involved evaluating voice-based repair strategies. Although participants also read a text-based scenario, it was their ability to interpret spoken system responses that was critical to the experimental goal. Table 4.3 presents the six CEFR listening levels used by **SL2** participants to self-assess their

[4] https://europass.europa.eu/en/common-european-framework-reference-language-skills

Table 4.3: **Common European Framework of Reference for Languages**

Level	Descriptions
A1	I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.
A2	I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.
B1	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.
B2	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.
C1	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. I can understand television programmes and films without too much effort.
C2	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.

English language proficiency.

4.6 Qualitative Interview Design (RQ4)

Unlike the previous quantitative research questions, RQ4 was addressed by employing a qualitative design using semi-structured interviews with Saudi participants to explore their perspectives on communication breakdowns and repair strategies. This culturally grounded and exploratory method aimed to complement the quantitative findings by providing deeper insights into how users interpret DVAs behaviour. While the previous research questions (RQ2 & RQ3) quantitatively investigated user preferences for repair strategies in response to communication breakdowns, RQ4 qualitatively explored participants' broader perspectives on these strategies, their interactions with DVAs, and their previous experiences. This complementary approach enabled a more nuanced understanding of user interpretations and real-world coping behaviours when interacting with DVAs.

According to Greene et al. [93], mixed-methods research can serve five primary purposes:

triangulation (seeking convergence and corroboration of results), complementarity (providing elaboration, illustration, and clarification of findings), development (using results from one method to inform another), initiation (uncovering paradoxes or contradictions), and expansion (extending the scope of inquiry). Within this mixed-methods approach in Study 2 (Chapter 6), the qualitative analysis is positioned as a complementary qualitative component, intended to elaborate and enrich the findings from the quantitative components.

Following their participation in the comparison experiment, Saudi participants were invited to take part in individual semi-structured interviews conducted in their native language (Arabic). Conducting interviews in participants' first language ensured the authenticity of responses and allowed accurate representation of their perspectives [267, 239]. Semi-structured interviews are the most common approach for collecting qualitative data due to their balance of structure and flexibility [107], enabling researchers to effectively explore participants' opinions and experiences [139]. As a native Arabic speaker, I conducted the interviews in Arabic, which not only facilitated open communication but also served as a powerful route to building trust and acceptance [267]. This approach demonstrated a willingness to engage with participants in their own cultural and linguistic terms through shared language [217, 11, 267].

Two bilingual (Arabic/English) researchers, including the author, independently conducted the qualitative analysis using an inductive thematic analysis approach [37]. This analytical method was selected due to its flexibility and suitability for identifying patterns and themes emerging directly from participant narratives. The analysis remained grounded in the original Arabic transcripts throughout the entire process, ensuring that interpretations remained culturally and linguistically appropriate. Maintaining the data analysis in the original language is essential, as it preserves the nuances, expressions, and culturally embedded meanings that can be lost in early-stage translations, thus enhancing the fidelity and validity of qualitative interpretations [256, 239].

Santos et al. [218] emphasizes that the timing of translation in cross-language qualitative research should be guided by the objectives of translation. Translation prior to data collection is appropriate when translating instruments of data collection, such as interview guides, questionnaires, and surveys, particularly when these instruments are designed to avoid colloquial language. In this thesis, translation prior to data collection was applied primarily to the preparation of quantitative materials, such as experimental instructions and survey instruments (see Section 4.8.2 for detailed translation processes).

However, for the qualitative interview component, translation occurred predominantly during data analysis, when categories, codes, and concepts generated from analysis in the original language (Arabic) were selectively translated into English to support collaborative and interpretive discussions within the supervisory research team. Maintaining the analysis in the original Arabic transcripts ensured cultural and linguistic fidelity, while translating selected codes and quotations facilitated collaborative interpretation and improved analytical transparency across the multilingual research team [256, 239, 218]. Additionally, translation was also undertaken at the dissemination stage, where final themes, sub-themes, and illustrative quotations from the analysis were translated into English for reporting and publication purposes, aligning with recommended practice when presenting findings from cross-language qualitative research [256, 218].

Interviews were conducted remotely via Microsoft Teams, using its audio recording functionality. The author manually transcribed and anonymized interview recordings and in preparation for thematic analysis [37]. Interview analysis was conducted using MAXQDA 22.80 software[5] as it supports Arabic language (see Appendix B.4.6 for the interview questions) .

4.7 Data Processing and Analysis

4.7.1 Preference Score Calculation

To prepare the data for analysis, participant responses from the 10 pairwise comparisons were converted into individual preference scores for each repair strategy. Each pairwise comparison required participants to choose one of two presented repair strategies. For each comparison, the selected strategy was coded as 1, while the non-selected strategy was coded as 0.

A complete paired-comparison design was used, in which all five repair strategies were compared once with each of the remaining strategies. As a result, each repair strategy appeared exactly four times per participant. Consequently, for each participant, preference scores for each strategy ranged from 0 (the strategy was never selected across its four comparisons) to 4 (the strategy was selected in all four comparisons in which it appeared).

This process resulted in five numerical preference scores per participant, one for each repair strategy. Because all strategies were presented an equal number of times, these scores provide a directly comparable measure of relative preference across strategies. The preference scores were subsequently used to generate descriptive statistics (e.g., mean, median) and to conduct inferential analyses (e.g., Friedman tests and Wilcoxon Signed-Rank tests), as reported in the empirical study results. For between-group comparisons, analyses were conducted at the preference-selection (trial) level using Mann–Whitney U tests, reflecting individual choice behaviour across repeated pairwise comparisons rather than aggregated participant scores. Table 4.4 presents an example of participant data to illustrate the preference score calculation.

To provide a clear overview of the analytical tests used across the two empirical studies, Table 4.5 outlines the main methods of analysis, their scope, and their specific purpose. This summary highlights how both quantitative and qualitative approaches were used to examine participants' preferences for repair strategies. While RQ2 and RQ3a relied on non-parametric tests to compare preferences within and between groups to investigate the effects of cultural background and inter-

[5]https://www.maxqda.com/

action language, RQ3b incorporated binary logistic regression to investigate how individual-level factors predicted strategy preference. RQ4 employed inductive thematic analysis to explore participants' perspectives in greater depth via the interview data. Table 4.5 provides an overview of the analytical procedures and tests used across the results sections in the empirical studies.

Table 4.4: Example of Preference Score Calculation for a Single Participant

Comparison	**Strategy A**	**Strategy B**	**Chosen**	**Strategy Score Update**
1	Social	Ask	Social	Social +1
2	Confirmation	Solve	Solve	Solve +1
.....	.....	.....	.....	.....

Table 4.5: Overview of Analytical Methods and Their Purpose Across Studies

RQ	**Test**	**Scope**	**Purpose**
RQ2 & RQ3a	Descriptive Statistics	Overall	Count how often each strategy was selected; explore general trends
	Friedman Test	Within-group	Test if there are overall significant differences in preferences and provides mean ranks
	Wilcoxon Signed-Ranks Test	Within-group	Compare pairwise differences between strategies within each group
	Mann–Whitney U Test	Between-group	Compare preference rankings across groups to identify significant differences
RQ3b	Binary Logistic Regression	Individual factors	Identify user characteristics predict strategy preference
RQ4	Inductive thematic analysis	Interview data	Identify themes related to user perspective of repair strategies

4.8 Preparatory Procedures

Before conducting the main two empirical studies, several preparatory steps were undertaken to ensure the validity and linguistic relevance of the research materials and methodology: 1) a pilot study to verify the distinctions between repair strategies and assess the realism of the employed scenarios, and 2) translation of the original English materials into Arabic following standardized guidelines.

4.8.1 Pilot Study

4.8.1.1 Aim

The primary aim of the pilot study was to ensure that each repair strategy was unique and distinguishable based on its main embedded attribute, despite potential overlaps with other strategies. For instance, apologies as politeness should be associated with the Social category, while the Information category should focus on providing explanations for the causes of conversational breakdowns (see Table 4.1). Additionally, this pilot study also aimed to validate that the task scenario (booking a flight) adapted for the experimental studies, realistically mimicked real-world interactions with DVAs. This aligns with the SWTS approach suggested by Borlund et al. [33, 34]. As Borlund recommends, it is important to tailor SWTS to the target population to ensure realism and experimental control [33] (see Section 4.3.2). Therefore, the aims of this pilot study were:

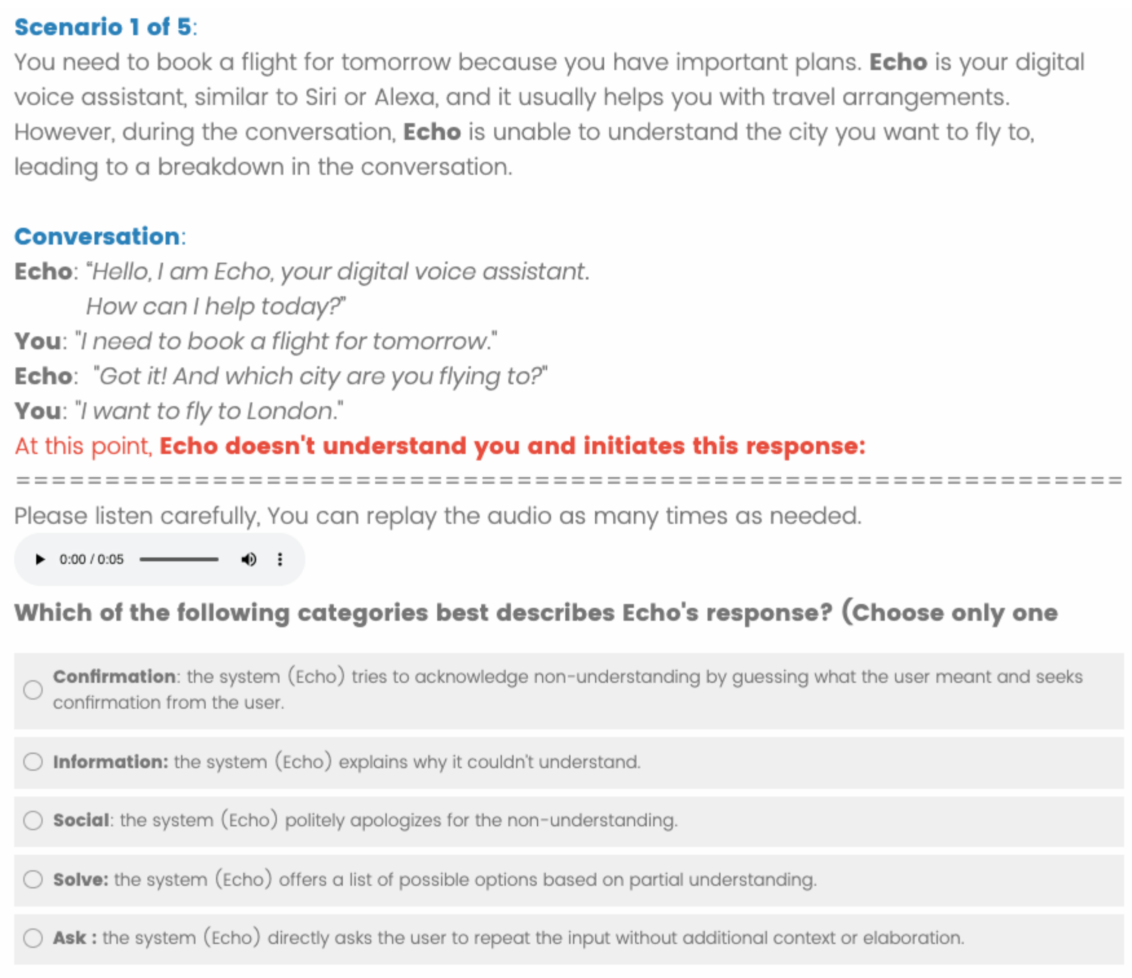
- **Validate Category Distinction:** To verify that repair strategies are clearly distinguishable based on their main attributes.
- **Assess Realism of the Scenario:** To evaluate whether the task scenario used in this thesis (the flight booking scenario) feels realistic and engaging for participants.

4.8.1.2 Method

To verify that each repair strategy aligns with its main attributes, an online experimental study was conducted in which participants were presented with a text-based scenario, followed by a voice response demonstrating one of the five repair strategies. After each scenario, participants were shown five predefined categories of the repair strategies, each describing its key attribute. They were then asked to select the category that best matched the repair strategy they heard. Figure 4.6 shows the task scenario along with the predefined categories of the five repair strategies. This direct attribute identification approach assesses how well participants can recognize the main attribute of each repair strategy and associate it with its corresponding category. A similar approach was employed by Yuan et al. [284], where participants were presented with error responses and asked to categorize them.

In this pilot study, there are five repair strategies, each representing a specific category, aiming to ensure that each repair strategy is distinguishable based on its key attributes and associated category. The text-based scenarios and voice responses (repair strategies) were identical to those subsequently used in the main experimental studies in this thesis. While the text-based scenario and voice responses were the same in both the pilot study and the main experimental studies, the key difference lies in the task design. In the pilot study, the scenario was followed by a randomly selected voice response representing one of the five repair strategies, after which participants were asked to select the predefined category that best described the response. In contrast, the main experimental studies presented each scenario with two voice responses (i.e., a pair of repair strategies), and participants were asked to indicate which of the two they preferred for handling the situation.

Participants were then asked to complete a realism check questionnaire adapted from [63] to assess the realism of the adapted simulated task scenario. Crescenzi et al. [63] suggest that a realism check questionnaire is a relatively low-effort method for evaluating the realism of SWTS. Participants completed the exit questionnaire that included two closed-ended items and one open-ended question. The closed-ended items were rated using a 7-point Likert scale, ranging from 1 = strongly disagree to 7 = strongly agree:



Scenario 1 of 5:
You need to book a flight for tomorrow because you have important plans. **Echo** is your digital voice assistant, similar to Siri or Alexa, and it usually helps you with travel arrangements. However, during the conversation, **Echo** is unable to understand the city you want to fly to, leading to a breakdown in the conversation.

Conversation:
Echo: *"Hello, I am Echo, your digital voice assistant.*
How can I help today?"
You: *"I need to book a flight for tomorrow."*
Echo: *"Got it! And which city are you flying to?"*
You: *"I want to fly to London."*
At this point, **Echo doesn't understand you and initiates this response:**
==
Please listen carefully, You can replay the audio as many times as needed.
0:00 / 0:05

Which of the following categories best describes Echo's response? (Choose only one

- **Confirmation**: the system (Echo) tries to acknowledge non-understanding by guessing what the user meant and seeks confirmation from the user.
- **Information:** the system (Echo) explains why it couldn't understand.
- **Social**: the system (Echo) politely apologizes for the non-understanding.
- **Solve:** the system (Echo) offers a list of possible options based on partial understanding.
- **Ask :** the system (Echo) directly asks the user to repeat the input without additional context or elaboration.

Figure 4.6: **Task scenario and predefined categories of the five repair strategies used in the pilot study.**

- I can imagine myself interacting with a digital voice assistant like the one in this study.
- I can imagine myself in the scenario (booking a flight through a digital voice assistant).
- Were there any moments in the scenario where the system's responses (Echo) felt unrealistic? If so, please describe. (open-ended)

4.8.1.3 Results

Using a convenience sampling method, 19 participants were recruited (males: 12, females: 7, average age: 30 years, SD = 7.53). This range of sample size is sufficient when a pilot study is used to identify and resolve issues related to implementing an intervention, such as verifying clarity and realism [108]. This pilot study took each participant typically 8 minutes to complete. Participants received a £5 Starbucks eGift Card. Ethical approval for this pilot study was obtained from Ethics Committee in the Department of Computer and Information Sciences (ethical approval number:2769; see Appendix A).

Piloting showed that participants correctly identified repair strategies, with accuracy exceeding 85%. This suggests the five repair strategies are distinguishable by their unique attributes. Additionally, participants' perspectives of the realism of the study scenario were evaluated using a one-sample t-test for the realism questionnaire. Results indicate a mean realism score ($M = 5.89, SD = 1.09$) significantly higher than the neutral midpoint ($t(18) = 7.59, p < .001$), with a large effect size ($Cohen'sd = 1.74$). Furthermore, 84.2% of participants responded "No" to whether the system's responses felt unrealistic, providing further support that the scenario was perceived as realistic. Insights from this pilot were used to enhance the design and layout of the main studies.

4.8.2 Translation Process

The questionnaires and experimental research materials were originally written in English and subsequently translated into Arabic. To ensure a high-quality translation, this process was guided by the International Test Commission (ITC) Guidelines for Translating and Adapting Tests [94]. These guidelines aim to promote validity and comparability in cross-linguistic and cross-cultural research. The ITC guidelines are widely recognized and commonly applied in cross-cultural studies (e.g., [208, 243]) and in HCI research (e.g., [68, 227, 226]) to ensure that translated materials preserve the original meaning and context. The translation aimed to ensure linguistic clarity and contextual relevance in the target language (Arabic).

All research materials were initially developed in English and then translated into Arabic. This included the survey instruments used in the empirical studies, such as participant instructions, questionnaire items, and system-generated voice prompts, as well as the semi-structured interview guide used in mixed-method Study 2. Following Santos [218], this stage aligns with translation prior to data collection, where the primary objective is to prepare data collection instruments in the participants' native language, free from colloquialisms or culturally ambiguous phrasing. This timing ensured that materials were accessible, linguistically appropriate, and culturally sensitive for Arabic-speaking participants, thereby enhancing the validity of both the quantitative and qualitative components of the studies.

Following ITC recommendations, a rigorous multi-step translation process was undertaken. This included double forward translation, reconciliation, and back-translation. First, two independent translators translated the English research materials into Arabic. The author, as a native Arabic speaker, collaborated with the translators to compare versions, resolve discrepancies, and ensure clarity and accuracy. A third bilingual translator then performed a back-translation into English. The back-translated version was compared with the original to identify any shifts in meaning, and necessary adjustments were made.

All translators were Saudi professional academics with doctoral degrees in English or translation studies and had a deep understanding of both linguistic and cultural contexts. Throughout the process, the author maintained active involvement in reviewing and validating the translated materials to ensure alignment with the research goals and the ITC's quality criteria. Finally, the Arabic version was reviewed for readability and clarity by two native Arabic speakers who were not involved in the initial translation process. They read through the materials and provided feedback to ensure the language was natural, clear, and appropriate for a general audience.

4.9 Chapter Summary

Building on the system repair framework developed in the previous scoping review chapter, this chapter presented the methodological foundations for the two empirical studies conducted for this thesis. It detailed the research designs used to investigate the effects of cultural background through a quantitative approach and interaction language through a mixed-methods approach. Common methodological components across both studies were introduced, including participant overview, experimental procedures, and preparatory steps such as pilot testing and translation. These consistent elements ensure comparability and coherence across the user studies. The next chapter presents the first empirical study, which focuses on the effect of cultural background on user preferences for repair strategies.

Chapter 5

Study 1: Effect of Cultural Background on Repair Strategy Preferences

5.1 Introduction

SDSs are increasingly embedded in daily human-machine interactions, offering intuitive, hands-free, and efficient communication [40, 43]. Despite these benefits, SDSs frequently experience conversational breakdowns that prevent users from completing tasks, leading to frustration, task abandonment [17], and diminished user satisfaction and trust [60, 154]. To mitigate these issues, SDSs employ communication repair strategies that aim to recover from misunderstandings and restore the flow of interaction. While research in ASR and NLP has advanced repair techniques (e.g. [190, 45]), understanding how diverse users respond to these strategies remains a critical gap.

Given the global use of SDSs across culturally and linguistically diverse populations, it is essential to examine how users' cultural backgrounds influence their preferences for repair strategies during conversational breakdowns. Prior studies in HCI and cross-cultural communication have shown that cultural background significantly shapes interaction styles and expectations (e.g., directness, elaboration, politeness) [186, 276, 228, 167], yet its role in shaping user preferences for SDS repair strategies is unexplored. As SDSs are increasingly deployed in multicultural environments, designing culturally responsive systems becomes a necessity, not just for functionality and utility, but also for supporting users' emotional well-being, rather than merely an optional feature [268, 269].

As HCI becomes increasingly conversational, users now engage with SDSs in ways that resemble interactions with human conversational partners rather than purely technical tools. This shift aligns with CASA paradigm, which suggests that people instinctively apply human-like social norms and expectations to machines [183, 181]. As a result, users bring their culturally shaped communication styles and norms into their interactions with SDSs, influencing how they perceive and respond to the system's behaviour, particularly during communication breakdowns.

To interpret these cultural influences, this study draws on well-established frameworks in intercultural communication. Hall's high- and low-context communication theory [103] and Hofstede's cultural dimensions [114] provide a valuable foundation for understanding how preferences for directness, elaboration, uncertainty tolerance, and social politeness vary across cultures. These cultural traits are particularly relevant to how users evaluate repair strategies in which SDSs attempt to recover from misunderstandings or non-understandings during interactions. A more detailed overview of these theoretical models is provided in Chapter 2, Section 2.4.

This chapter presents a comparative quantitative study that investigates how cultural background affects user preferences for communication repair strategies during conversational breakdowns with DVAs. The study specifically contrasts two cultural groups: British participants (BL1), representing a low-context culture, and Saudi participants (SL1), representing a high-context culture. Through this cross-cultural comparison, the study aims to reveal how cultural background shape preference of system-initiated repair strategies. The following research questions guide this investigation:

- **RQ2**: **How does users' cultural background influence their preference for communication repair strategies in DVAs?**
  - **RQ2a:** Which repair strategies are preferred in low-context cultures when a conversational breakdown occurs with DVAs?
  - **RQ2b:** Which repair strategies are preferred in high-context cultures when a conversational breakdown occurs with DVAs?

To address these questions, this chapter is structured as follows. Section 5.2 outlines the methodology, including the repair strategy framework, participant demographics, task scenario, and paired comparison experiment design. Section 5.3 then presents the results, comprising both within- and between-group comparisons. Section 5.4 discusses the findings through the lens of cultural communication theories, and Section 5.5 concludes with theoretical reflections.

5.2 Method

To explore how cultural background influences user preferences for repair strategies in SDSs, the author conducted a quantitative experimental study comparing two participant groups from different cultural backgrounds using a pairwise comparison method (addressing RQ2). Participants' preferences were measured by asking them to choose between pairs of system responses (repair strategies) following a simulated interaction breakdown. This study employed quantitative data analysis to compare preference trends between groups. This section provides details on the participant groups involved, the simulated task scenario, the comparative study design, and the overall experimental setup.

5.2.1 Repair Strategies Framework

This study examines user preferences for different system-repair strategies within SDSs, specifically focusing on the effect of cultural background (high-context vs. low-context cultures). To ensure methodological consistency across the empirical studies in this thesis, the system-repair strategies framework previously established in the scoping review was adopted (Chapter 3, Section 3.3.2). This framework categorizes system-initiated repair strategies into six distinct types, each defined by its communicative purpose and the specific function of the system's response following a conversational breakdown (see Framework 3.4).

For this study, five repair strategies were selected from the framework: Confirmation, Information, Social, Solve, and Ask. These particular strategies were chosen due to their 1) clearly distinct communicative attributes and 2) their direct role in addressing immediate conversational breakdowns in DVA interactions. The sixth category, Disclosure, was intentionally excluded from this thesis because it primarily conveys broader background information about the system's limitations rather than directly repairing conversational breakdowns. By focusing on these five categories, this study specifically aims to understand how cultural background influences user preferences for repair strategies during communication breakdowns.

Each of these repair strategies possesses a unique primary attribute that corresponds to its functional objective during miscommunication events. Although there may be some overlap in the strategies' approach to breakdown resolution, this study highlights the individual and distinguishing features of each strategy. Emphasizing these distinct attributes provides clarity in measuring user preferences, thereby facilitating a deeper understanding of how cultural context impacts users' preferences for these repair responses. A detailed summary of these attributes and their specific descriptions is provided in Table 4.1 in the Methodology Chapter 4. Further information on the full repair strategies framework used across the thesis can be found in Section 4.2.

5.2.2 Participants

Guided by cultural theory [103], this study involved two participant groups from different cultural backgrounds: (1) British speakers (BL1) interacting with the system in their native language (English), representing a low-context culture, and (2) Saudi speakers (SL1) interacting with the system in their native language (Arabic), representing a high-context culture. Participants were categorized into these groups based on their self-identified nationality and the native language they used when interacting with the system. These two groups were selected to enable a direct comparison between high- and low-context cultures, thereby examining how cultural background shapes user preferences for repair strategies in SDSs. By ensuring both groups interacted with the system in their native languages, the study aimed to isolate the effects of cultural background while minimizing the influence of language proficiency differences.

A total of 101 participants participated in this study: 52 British participants (30 females, 19 males, 1 non-binary, 2 prefer not to say; mean age = 24.55 years, SD = 6.67) recruited primarily from the University of Strathclyde in the UK, and 49 Saudi participants (26 females, 23 males; mean age = 31.95 years, SD = 9.57) recruited primarily from Prince Sattam University. Participants were recruited through university email lists, posters, flyers, and snowball sampling techniques. Each participant received a Starbucks eGift voucher valued at £5 (approximately 25 SR) as compensation for their participation. Ethical approval for this study was granted by the Ethics Committee in the Department of Computer and Information Sciences at the University of Strathclyde (approval number: 2781; see Appendix B)

5.2.3 Simulated Task Scenario

This study employed a goal-oriented task embedded within the framework of the SWTS to evaluate user preferences for repair strategies in SDSs, as recommended by Borlund [33, 34]. In SWTS, scenarios are drawn from real-life situations to help participants relate more easily (see Section 4.3.2 for more details). In this study, participants were asked to imagine that they urgently needed to book a flight for the following day using a DVA named *Echo*. The assistant, which is typically used for travel arrangements, served as the system through which the task was conducted. DVAs are a common form of SDSs, with well-known commercial examples including Apple Siri and Amazon Alexa [122]. A DVA was used here to provide a familiar and realistic interface, helping participants better relate to the task and ensuring clarity in understanding the scenario.

The scenario was designed to simulate a realistic and structured interaction, in line with the characteristics of goal-oriented tasks, where the primary objective is task completion through accurate and efficient communication. Goal-oriented tasks are known for their transactional nature, requiring precise information exchange [161]. In this case, the system (Echo) was expected to recognize the participant's travel destination ("London") in order to proceed with the conversation and complete the flight booking task.

A deliberate communication breakdown was introduced to the scenario to reflect a common issue in SDSs, called partial non-understanding [234, 232, 233]. In this instance, the system successfully recognized the general user intent (i.e., booking a flight) but failed to identify the specific destination entity ("London"). This type of breakdown is a subtype of non-understanding, where the system's speech recognition confidence for the entity falls below a usable threshold [198, 29]. As a result, the system did not proceed with uncertain information and instead initiated a repair strategy. This approach aligns with the principles of Grounding Theory, which emphasizes the importance of immediate and collaborative actions to achieve mutual understanding during dialogue interactions [49, 51]. In alignment with this theory, the system promptly responded to the breakdown by employing one of five predefined repair strategies, aiming to restore mutual understanding and facilitate the completion of the goal-oriented task. The full text-based scenario, including the dialogue and task instruction presented to participants, is illustrated in Figure 4.3, Methodology Chapter 4.

To reduce potential biases in evaluating repair strategies (such as item priming effects [203]), the correct entity ("London") was excluded from all system suggestions in repair strategies involving alternatives. For instance, the system (Echo) might propose alternative destinations such as 'Luton', 'Sunderland', or 'Liverpool', while omitting the correct option. This design ensured that participants' judgments were based on the form and function of the repair strategies, rather than the presence of the correct answer. The same scenario and breakdown condition were used across all experimental studies (Study 1 and 2) in this thesis to ensure consistency and comparability. The realism of the employed SWTS was validated through a pilot study. Further details about the pilot study can be found in the Preparatory rocedures section of this thesis (see Section 4.8.1).

5.2.4 Paired Comparison Experiment

To collect participants' preferences for repair strategies, an online pairwise comparison experiment was conducted using Qualtrics[1]. Pairwise comparison is a widely used method in Human-Computer Interaction (HCI) for assessing judgments, preferences, and personal ratings [17, 141, 80, 16, 159, 67, 47]. Compared to Likert scales, it offers more realistic results and is better suited for detecting linguistic and cultural differences [191, 3]. It also avoids the need for culturally specific scenarios by relying on straightforward binary judgments [67, 191].

In the experiment, participants began by reading a short text-based scenario that described a goal-oriented task involving a typical interaction with a DVA. The scenario simulated a realistic situation in which the assistant experienced a communication breakdown, specifically, a failure to understand the user's travel destination (London). To support participant understanding and engagement, the concept of a DVA (e.g., Amazon Alexa or Apple Siri) was introduced at the beginning of the experiment as a familiar and relatable example of a SDS. This approach helped ensure clarity throughout the task and maintained alignment with the broader focus on SDSs in this thesis.

[1]https://www.qualtrics.com/

Following the text-based scenario, participants listened to two voice responses from a DVA named *Echo*, each representing a different repair strategy. Participants were then asked to select the strategy they preferred for resolving the breakdown. Each participant completed 10 binary pairwise comparisons, covering all possible pairings among the five repair strategies. The dependent variable in this study was repair strategy preference, operationalised as the outcome of each binary comparison (i.e., which strategy was selected). Because each repair strategy appeared in four comparisons per participant, each participant contributed four preference decisions per strategy.

Pre-recorded voices were synthesized using Amazon Polly[2], a text-to-speech tool, to generate natural-sounding female speech in both Arabic and English. For Arabic-speaking Saudi participants (SL1), the scenario and system responses were translated from English to Arabic following International Test Commission (ITC) Guidelines for Translating and Adapting Tests [94]. Details on the translation process are provided in the Preparatory Procedures Section of this thesis (see Section 4.8.2).

To minimize potential order bias [150], both the sequence of comparisons and the order of the two strategies within each pair were randomized using Qualtrics' built-in randomization features. Participants were explicitly informed that they would: (1) read a scenario where they are booking a flight using a DVA, and (2) hear two voice responses from Echo, and should listen carefully and choose the one they preferred. These clear instructions were designed to improve focus and reduce task confusion.

Importantly, the task scenario and experimental procedure were kept consistent across all experimental studies in this thesis. This consistency ensured that participants across different conditions and groups experienced comparable interactions, enabling meaningful and fair comparisons. Further details about the consistent experimental design and procedures can be found in Section 4.4 of Chapter 4 Methodology.

5.2.5 Study Design

To address RQ2, a comparative research design using a mixed factorial framework was employed to examine how cultural background influences user preferences for repair strategies in SDSs. Guided by Hall's cultural theory [103], this study involved two different cultures: a low-context culture (BL1) and a high-context culture (SL1).

The experiment followed a 2×5 factorial design, with cultural background (low-context vs. high-context) as a between-subjects factor, and repair strategy (five types) as a within-subjects factor. This mixed design allowed us to compare how each group ranked different system responses and to observe any significant differences in preferences across cultural groups. To implement this design, this study followed a structured series of stages, beginning with participant briefing and screening,

[2] https://aws.amazon.com/polly/

followed by demographic surveys, and then 10 pairwise comparison tasks. The overall procedure is illustrated in Figure 5.1, which visually outlines the experimental stages that participants completed in this study.

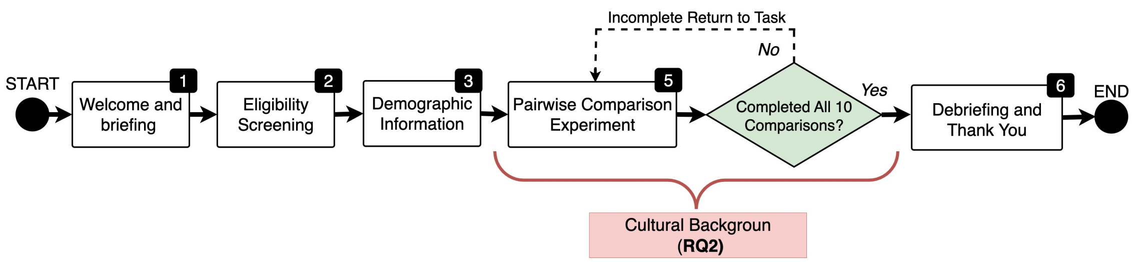


Figure 5.1: Overview of the experimental procedure completed by participants in Study 1.

5.3 Results

This section presents the analysis of participants' preferences for the five distinct repair strategies in DVAs, with a focus on the impact of cultural background (addressing RQ 2). Specifically, the analysis examines whether there are significant differences in repair strategy preferences within and between two cultural groups: low-context cultures (BL1) and high-context cultures (SL1).

All analyses were conducted using IBM SPSS Statistics 29.0. As the data did not meet the assumptions of normality, non-parametric tests were employed [59]. These included the Friedman test and Wilcoxon-Singed Ranks test for within-group comparisons and the Mann-Whitney U test for between-group comparisons. To account for multiple comparisons, a Bonferroni correction was applied where appropriate, adjusting the significance threshold to $p < 0.005(0.05/10)$ based on the total number of pairwise tests conducted [62, 265, 59]. All statistical tests were conducted using two-tailed significance levels. The results are presented in the following three subsections: overall descriptive statistics, within-group comparisons, and between-group comparisons.

5.3.1 Overall Descriptive Statistics

As described in Section 4.7.1, preference scores were calculated based on how often each repair strategy was selected during pairwise comparisons, with scores ranging from 0 (never selected) to 4 (selected in all comparisons in which the strategy appeared). Table 5.1 presents descriptive statistics for the five repair strategies, showing the mean and standard deviation of the number of times each strategy was selected by participants from the BL1 and SL1 groups. These results provide an initial overview of participant preferences prior to conducting inferential statistical analysis.

Among SL1 participants, the Social strategy received the highest mean preference score (M = 2.47, SD = 1.08), followed closely by the Ask strategy (M = 2.35, SD = 0.99). In contrast, the Solve and Information strategies were less frequently selected (M = 1.61 and M = 1.78, respectively). On the other hand, for BL1 participants, the Ask strategy was the most preferred (M = 2.67, SD = 1.06), followed by the Confirmation strategy (M = 2.27, SD = 1.21), while the Information strategy received the lowest mean score (M = 1.13, SD = 1.12).

Table 5.1: **Mean preference scores (and standard deviations) for each repair strategy, separately for BL1 and SL1 groups.**

Strategy	**BL1** **Mean (SD)**	**SL1** **Mean (SD)**
Social	2.25 (1.21)	2.46 (1.08)
Confirmation	2.26 (1.20)	1.79 (1.45)
Information	1.13 (1.12)	1.77 (1.10)
Solve	1.67 (1.29)	1.61 (1.22)
Ask	2.67 (1.06)	2.34 (0.99)

5.3.2 Within-Group Comparisons

5.3.2.1 Overall Within-Group Differences: Friedman Test

This subsection presents the results of within-group analyses for both BL1 and SL1 participants. To examine whether participants demonstrated statistically significant differences in their preferences among the five repair strategies, within-group comparisons were conducted separately for each cultural group. Given the non-parametric nature of the data, the Friedman test was used to assess overall differences in preference rankings within each group [59].

The Friedman test results revealed significant differences in the mean ranks of the five repair strategies for both groups: BL1 participants ($\chi(4) = 34.04, p < .001$) and SL1 participants ($\chi(4) = 14.00, p = .007$). These results indicate that within each group, participants demonstrated meaningful variation in their preferences for different repair strategies. Figure 5.2 presents the mean ranks and ranking orders for each strategy, separated by group. Among BL1 participants, the Ask strategy had the highest mean rank (3.69). In contrast, SL1 participants preferred the Social strategy (3.53) most.

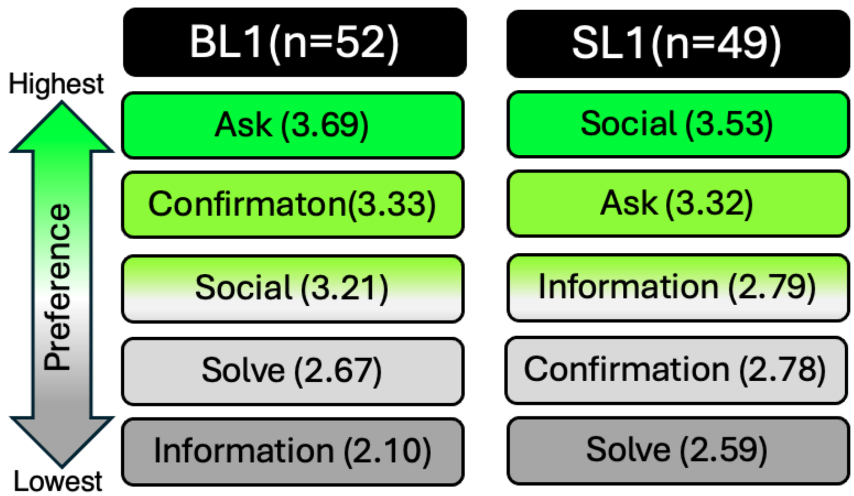


Figure 5.2: : **Preferences for repair strategies: British participants (BL1), and Saudi participants (SL1) based on Friedman Mean Rank Tests. The figure displays the mean ranks for each repair strategy, where higher values indicate greater preference. Strategies are ordered from highest (Green) to lowest (Gray) preference; vertical spacing is illustrative only.**

5.3.2.2 Pairwise Preference Comparisons: Wilcoxon Signed-Ranks Test

To further explore specific pairwise differences in strategy preferences within each cultural group, the Wilcoxon Signed-Ranks Test was conducted. This analysis identifies precisely which pairs of repair strategies differed significantly from one another within each group. The results, summarized in Table 5.2, reveal several statistically significant differences, with significance thresholds adjusted using the Bonferroni correction method ($p < 0.005$).

Among SL1 participants, the Social strategy was significantly preferred over the Solve strategy ($Z = -3.04, r = 0.30$) and the Information strategy ($Z = -3.06, r = 0.30$). In contrast, BL1 participants exhibited a distinctly low preference for the Information strategy, ranking it significantly lower than the Ask ($Z = -4.94, r = 0.49$), Social ($Z = -3.99, r = 0.40$), and Confirmation ($Z = -3.57, r = 0.36$) strategies. Additionally, BL1 participants showed a significant preference for the Ask strategy compared with the Solve strategy ($Z = -3.17, r = 0.32$).

Overall, the Wilcoxon Signed-Rank pairwise comparisons (Table 5.2) indicate multiple statistically significant differences in repair strategy preferences within each cultural group. These results demonstrate that participants in both groups exhibited structured and differentiated preference patterns among the five repair strategies, providing a basis for subsequent between-group comparisons. Specifically, SL1 participants showed a significant preference for the Social strategy over the

Table 5.2: **Adjusted Significance Levels from Wilcoxon Signed-Rank Tests for Each Pairwise Comparison Using Bonferroni Correction** ($p < 0.005$); **'**' Indicates Significant Values.**

Paired Comparisons of Strategies	**All Participants p-value**	**SL1 p-value**	**BL1 p-value**
Social vs. Confirmation	0.144	0.043	0.968
Social vs. Information	0.001**	0.002**	0.000**
Social vs. Solve	0.001**	0.002**	0.055
Social vs. Ask	0.266	0.574	0.051
Confirmation vs. Information	0.006	0.854	0.000**
Confirmation vs. Solve	0.025	0.591	0.007
Confirmation vs. Ask	0.012	0.059	0.086
Information vs. Solve	0.356	0.558	0.052
Information vs. Ask	0.000**	0.010	0.000**
Ask vs. Solve	0.000**	0.010	0.002**

Information and Solve strategies ($p = .002$). Conversely, BL1 participants exhibited a pronounced aversion to the Information strategy, ranking it significantly lower compared to Ask ($p < .000$), Social ($p < .000$), and Confirmation ($p < .000$) strategies. Moreover, the BL1 group showed a significant preference for the Ask strategy compared with the Solve strategy ($p = .002$). These findings highlight consistent intra-group preferences and provide a foundation for exploring cross-cultural comparisons in subsequent analyses.

5.3.3 Between-Group Comparisons: Mann–Whitney Tests

To compare participants' repair strategy preferences across cultural groups (BL1 vs. SL1), the dataset was restructured into a long-format representation at the comparison level. Each row captured a single pairwise preference judgment, with every participant completing 10 comparisons. Since each repair strategy appeared in four comparisons per participant, this restructuring produced 404 preference judgments across all 101 participants in total, representing 2,020 strategy-level decisions. From these judgments, mean rank scores were computed for each strategy per group, allowing for an aggregated measure of how often a strategy was preferred relative to others.

Group comparisons were conducted at the comparison level to maximise sensitivity to preference patterns across strategies. This approach is commonly used in pairwise comparison designs (e.g., [17, 16]), where individual preference judgements constitute the fundamental observational unit. While participant-level aggregation was used to derive descriptive mean rank scores, inferential analyses were performed on the trial-level preference data to reflect the underlying choice behaviour across repeated comparisons.

Using the Mann–Whitney U test, comparisons were conducted to assess whether preference scores for each strategy differed significantly between British (BL1) and Saudi (SL1) participants (Bonferroni-adjusted, $p - value < 0.005$) [59]. As shown in Figure 5.3, only one strategy showed a statistically significant between-group difference: the Information strategy. Saudi participants demonstrated a statistically significant stronger preference for the Information strategy compared to British participants ($U = 17118$, $n_{BL1} = 208$, $n_{SL1} = 196$, $p = .001$), who ranked it consistently lower. This result indicates a statistically significant difference in preferences for the Information strategy between the two cultural groups.

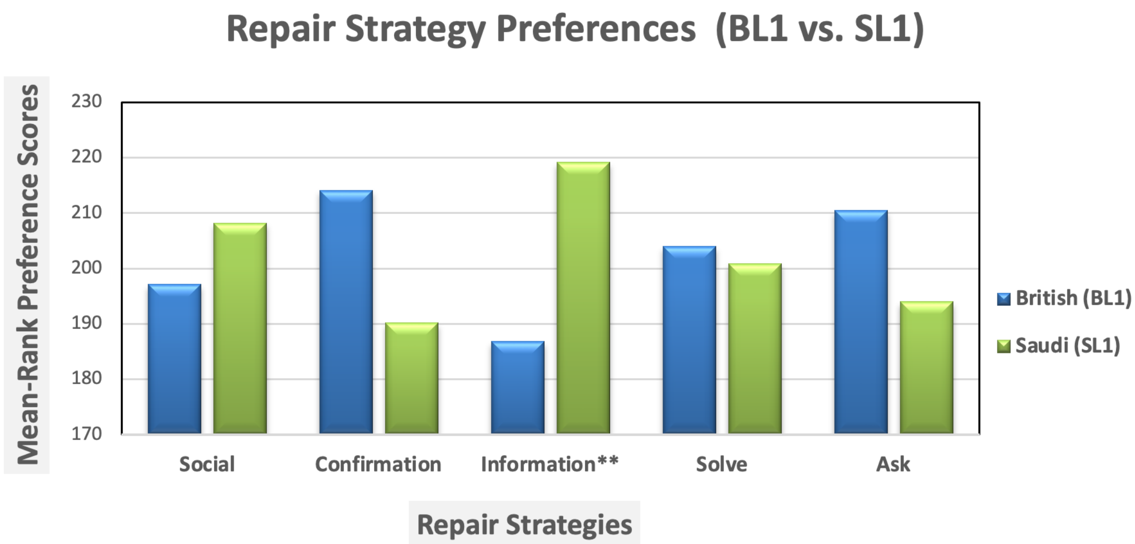


Figure 5.3: **Bar chart comparing the mean preference scores of five repair strategies between British (BL1) and Saudi (SL1) participants. Scores reflect the cumulative ranking of each strategy across all pairwise comparison trials (N = 404), where higher values indicate greater preference. Strategies marked with (**) showed statistically significant differences in preference between cultural groups (Mann–Whitney U test, Bonferroni-adjusted** $p < 0.005$).

5.4 Discussion

This section discusses the key findings of the study, beginning with an overview of the within- and between-group results (Section 5.4.1). It then interprets the statistically significant findings through the lens of cultural communication theories and prior HCI literature (Section 5.4.2). Finally, Section 5.4.3 considers why no statistically significant cultural differences were observed for the remaining repair strategies.

5.4.1 Overview of Key Findings

This study examined whether and how cultural background influences user preferences for repair strategies in SDSs. Using a pairwise comparison method, participants from low-context (British, BL1) and high-context (Saudi, SL1) cultural groups evaluated five repair strategies within a simulated task-based interaction. Preferences were analysed using non-parametric within-group and between-group comparisons to identify structured preference patterns and potential cultural effects.

The results revealed clear and statistically significant preference structures within each cultural group. Within-group analyses showed that BL1 participants significantly preferred the Ask strategy, followed by the Social and Confirmation strategies, over the Information strategy, whereas SL1 participants significantly preferred the Social strategy over the Solve and Information strategies. However, when preferences were compared directly between cultural groups, only the Information strategy showed a statistically significant difference. Specifically, SL1 participants demonstrated a statistically stronger preference for Information repair strategy than BL1 participants. Overall, these findings suggest that cultural background influences preferences for explanatory repair strategies, rather than shaping preferences across all repair types.

5.4.2 Cultural Background and Preference for Information Repair Strategies (RQ2)

This study examined how cultural background influences user preferences for repair strategies in SDSs (RQ2). The between-group comparison (Mann–Whitney U test) indicated that cultural background was associated with a statistically significant difference for one strategy only: the Information strategy. Specifically, Saudi participants (SL1) showed a significantly stronger preference for the Information strategy than British participants (BL1) ($p < .001$). This finding suggests that cultural background influences preferences in a selective manner, with the cultural effect emerging primarily for explanatory repair.

The Information strategy in this study was represented by the prompt "*I could not understand you because it is noisy here.*". Unlike strategies that primarily request repetition or confirmation, this response provides an explicit explanation for the breakdown (i.e., "it is noisy here"), adding contextual detail beyond acknowledging that an error occurred. Hall's context theory offers one explanation for why this form of repair may be more valued in high-context cultures such as Saudi Arabia [103]. High-context communication relies more heavily on contextual cues and shared understanding, and it often places greater value on richer background information to support interpretation. In this sense, the Information strategy may be perceived as a more complete and cooperative repair move because it supplies contextual grounding for the system's failure, rather than leaving the cause of the breakdown implicit. By contrast, low-context communication styles are typically characterised by greater emphasis on brevity and explicit task-oriented requests [104], which may contribute to the lower relative preference for an explanatory repair among BL1 participants.

A complementary explanation can be drawn from Hofstede's uncertainty avoidance dimension [114]. Cultures higher in uncertainty avoidance tend to show lower tolerance for ambiguity and place greater value on communication that reduces uncertainty through clarification and justification. Within SDS breakdowns, an explanation of the cause (e.g., background noise) can reduce uncertainty about why the system failed to understand the input and can provide reassurance that the problem is situational rather than user-related. The present findings are consistent with the interpretation that SL1 participants may place greater value on this type of clarification during breakdowns, whereas BL1 participants may prefer more minimal repairs that move quickly to repetition or confirmation.

Prior cross-cultural HCI research provides converging evidence that preferences for elaboration and information density can vary systematically across cultural contexts. For example, Liljenberg et al. [148] used eye-tracking to compare user interaction with culturally targeted website designs and reported that UK users preferred lower information density, whereas Chinese users preferred higher information density. Similarly, Kyriakoullis and Zaphiris [140] found that Chinese tourism e-commerce websites contained more textual information and navigational links than Australian websites, reflecting differences in how much contextual information is provided and valued. Focusing more directly on conversational settings, Miehle et al. [167] reported cross-cultural differences in preferences for elaborate versus direct system responses, with German participants (low-context) preferring less elaborate responses and Spanish participants (high-context) preferring more elaborate responses. Taken together, these findings indicate that cultural background influences preferences for repair strategies primarily when the repair response provides explanation and additional contextual information, with the observed cultural effect in this study emerging specifically for the Information repair strategy.

5.4.3 Understanding the Absence of Cultural Differences in Other Repair Strategies

In contrast to the Information strategy, the between-group comparisons indicated no statistically significant cultural differences in preferences for the Ask, Social, Confirmation, or Solve strategies after Bonferroni correction ($p < 0.005$). These null results suggest that, within the constraints of the present design and task context, cultural background did not reliably differentiate preferences for these strategies within the constraints of the current design and task scenario. Below, several design- and context-based considerations are provided that may interpret the absence of detectable cultural effects for these strategies.

First, the experimental instantiation of some strategies may have reduced the likelihood of observing cultural differences. In particular, the Confirmation and Solve strategies intentionally included incorrect content (i.e., an incorrect suggestion or an option set that did not contain the correct destination). This design choice was used to simulate a realistic breakdown and to avoid inflating preferences by presenting a correct answer (see Section 4.3.3). However, it may also have reduced the perceived usefulness and trustworthiness of these strategies, thereby compressing preference differences and shaping selections in the pairwise comparison context. Accordingly, the preference scores should be interpreted as relative judgments under the specific strategy implementations used in this study. Future work could evaluate alternative versions in which Confirmation provides a correct suggestion and/or Solve includes the correct option, to better disentangle the effect of strategy type from the accuracy of the suggested content.

Second, the absence of cultural differences for the Ask strategy may reflect the task-oriented nature of the scenario. Although the Ask strategy was most preferred by BL1 participants, it was also rated highly by SL1 participants. In goal-oriented contexts such as booking a flight, efficiency and clarity may be prioritised, and users may converge on repair strategies that support rapid recovery from breakdowns. Prior work on conversational agents suggests that users often adapt their behaviour and expectations once system limitations are apparent, for example by simplifying their language, repeating commands, or making requests more directly [154]. This adaptive behaviour may reduce cross-cultural variation for highly functional repair strategies such as a direct request to repeat.

Third, the null effects for Social repair strategies are consistent with mixed findings in the literature on apology-based repair. While apologies have been shown to reduce frustration and increase perceived usability in some contexts [284, 5, 251], other studies report that apology-based strategies can be rated lower on likeability or perceived intelligence [72]. Effects may also depend on situational and user factors such as mood or demographics [5, 260]. In this study standardised prompts and a single task scenario may have limited sensitivity to these contextual moderators, potentially contributing to the absence of detectable cultural differences for Social strategies. Finally, prior research suggests that relying on a single repair strategy may be insufficient in practice, and that multi-strategy or adaptive repair approaches (e.g., combining an apology with an explanation, or tailoring responses to context) can support stronger recovery and sustained engagement [39, 73, 26, 35]. While the present study compared strategies in isolation to identify relative preferences, future work could examine culturally adaptive sequences of repair strategies and test whether cultural differences become more apparent when strategies are combined, tailored, or deployed dynamically across different breakdown types and task contexts [73, 26]. Several of the considerations discussed above also reflect methodological limitations of the present study, including the specific instantiation of repair strategies and the task context. Broader limitations and directions for future research are discussed in detail in Conclusion Chapter 8.

5.5 Conclusion

Using a pairwise comparison experiment, this chapter quantitatively examined how cultural background influences user preferences for repair strategies in SDSs, comparing participants from high-context (Saudi) and low-context (British) cultures. Across the five repair strategies investigated, only the Information strategy (where the system explicitly explains the cause of a communication breakdown) demonstrated a statistically significant difference in preference between the two groups ($U = 17118$, $n_{BL1} = 208$, $n_{SL1} = 196$, $p = .001$). Saudi participants showed a stronger preference for this explanatory repair strategy than British participants, indicating that cultural background can influence preferences for system responses that provide additional contextual information during breakdowns.

In contrast, no statistically significant between-group differences were observed for the Ask, Social, Solve, or Confirmation strategies after correction for multiple comparisons. These findings suggest that cultural background does not uniformly affect preferences for all repair strategies, and that cultural influence in this study was selective rather than pervasive. Taken together, the results highlight the particular importance of explanatory repair strategies in cross-cultural SDS interactions and point to the need for future research to examine how task context, strategy combinations, and adaptive repair mechanisms may further shape user preferences across cultural settings.

5.6 Chapter Summary

While the previous chapter outlined the overall methodology for the empirical studies, this chapter investigated the effect of cultural background on user preferences for repair strategies in SDSs using a cross-cultural design. Although only the Information strategy showed a statistically significant difference between cultural groups, other patterns emerged that suggest underlying influences from within-culture and individual-level factors. These findings indicate that cultural background alone may not fully explain user preferences. The next chapter shifts the focus to a within-culture mixed-methods investigation of interaction language, exploring how using a native versus non-native language, alongside individual-level characteristics, shapes preferences for repair strategies when interacting with a SDS.

Chapter 6

Study 2: Effect of Interaction Language on Repair Strategy Preferences

6.1 Introduction

While the previous Chapter 5 examined how cultural background influences user preferences for repair strategies in SDSs, the current chapter shifts focus to the influence of within-culture variability, particularly among users from the same cultural background (Saudi Arabia). Specifically, this study investigates how interaction language operationalised as the language used by the system in its repair responses (native vs. non-native) and individual-level user characteristics (such as self-efficacy, prior experience, and language proficiency) influence preferences for communication repair strategies in SDSs. To develop a deeper understanding, this study also incorporates a qualitative component that explores users' perspectives and their experiences during interactions with SDSs, especially during communication breakdowns and repair phases. To develop a deeper understanding, this study also incorporates a qualitative component that explores users' perspectives and their experiences during interactions with SDSs, especially during communication breakdowns and repair phases.

As users increasingly interact with SDSs across different languages and devices, it becomes crucial to understand not only group-level cultural effects but also how users within the same cultural group vary based on their interaction context and individual characteristics. As the prevalence of DVAs (such as Apple's Siri and Google Assistant) continues to expand globally, they serve an increasingly diverse user base by supporting multiple languages [99]. However, the usability and effectiveness of these technologies depends not only on advances in ASR and NLP but also on their ability to accommodate diverse linguistic and communicative needs [281, 282].

A key challenge for ensuring the accessibility of DVAs is their adaptation across different languages, particularly for both native and non-native interactions. While HCI research has investigated general user experiences of native and non-native speakers interacting with DVAs [275, 273, 209], less attention has been paid to how DVAs handle communication breakdowns when they fail to understand or misinterpret user input. The strategies used to repair these breakdowns are critical for sustaining smooth and effective interactions [48, 17, 163]. While much research on DVA adaptation has focused on improving speech recognition accuracy [124, 120, 35], little attention has been paid to how repair strategies should be tailored to account for differences between native and non-native speakers.

This study contributes to the body of research on SDSs by examining how interaction language influences user preferences for repair strategies when interacting with DVAs. To address this gap, this study aims to identify the most suitable repair strategies for goal-oriented tasks in DVAs. Specifically, this study adopts a mixed-methods design, combining both quantitative and qualitative approaches to gain a more comprehensive understanding. The quantitative component investigates how interaction language (native vs. non-native) and individual-level factors (self-efficacy, prior experience with DVAs, and second-language proficiency) influence user preferences for repair strategies. The qualitative component, based on semi-structured interviews, explores users' personal experiences with DVAs and their perspectives on different repair strategies and communication breakdowns. This integration of methods enables a more holistic understanding of user-system interaction dynamics [93].

This study is guided by two overarching research questions, combining both quantitative and qualitative aims:

- **RQ3: What is the influence of interaction language and individual-level user characteristics on user preferences for repair strategies in SDSs?** This question is addressed using quantitative analysis and is broken down into the following sub-questions:
  - **RQ3a**: Which repair strategies are preferred by users when interacting with a DVA in their native versus non-native language during goal-oriented tasks?
  - **RQ3b**: How do individual-level factors, such as computer self-efficacy, prior experience with DVAs, and language proficiency, influence user preferences for repair strategies?
- **RQ4:** What are users' perspectives on different repair strategies and communication breakdowns in DVAs?

The remainder of this chapter is structured as follows: Section 6.2 details the methodology for both the quantitative and qualitative components; Section 6.3 presents the comparative results and the thematic analysis of the qualitative interview data; Section 6.4 discusses the integrated findings. Finally, Section 6.5 summarizes the key findings and prepares the ground for the final Discussion chapter.

6.2 Method

To explore how variation in user characteristics within a shared cultural context influences user preferences for repair strategies in DVAs, a mixed-methods approach was adopted. The influence of interaction language was quantitatively examined through an experimental study that compared two participant groups from the same cultural background, each evaluating system repair responses presented in either their native or non-native language (addressing RQ3a). In addition, individual-level factors, such as computer self-efficacy, prior experience with DVAs, and language proficiency, were explored using a survey-based method (addressing RQ3b). Finally, users' perspectives on different repair strategies and communication breakdowns in DVAs were qualitatively investigated through semi-structured interviews (addressing RQ4). This method section provides details of the participant groups involved, simulated task scenario, Paired comparison experiment, measurement of individual-level factors, and the design of the qualitative interview component.

6.2.1 Repair Strategies Framework

This study focuses on user preferences for different system-repair strategies in SDSs, specifically in the context of interaction language (native vs. non-native). Consistent with the previous study on cultural influences (Chapter 5), the author adopted the system-repair strategies framework developed in the scoping review, Chapter 3 [9]. This framework categorizes system-repair strategies into six distinct types, each defined by its communicative function and response objective during interactions with SDSs (see System-repair Strategy Framework 3.4).

For the current study, five repair strategies were selected from this framework: Confirmation, Information, Social, Solve, and Ask. These strategies were chosen for two main reasons: (1) they represent clearly differentiated system behaviours, and (2) they reflect the types of immediate responses that a DVA would likely produce after encountering a conversational breakdown. The sixth category, Disclosure, was excluded because it typically involves providing background information about system limitations rather than directly addressing the conversational breakdown. By focusing on the remaining five, this study aims to investigate how interaction language and individual-level factors influence user preferences for different system repair approaches in real-time non-understanding scenarios.

Each repair strategy is characterized by a primary attribute that defines its function during a breakdown. While some conceptual overlaps between strategies may occur, this study emphasizes the distinctive communicative goals that each strategy represents. These distinctions are central to understanding how different system responses contribute to repairing non-understandings and facilitating effective interaction. The specific attributes and descriptions of the five selected repair strategies are summarized in Table 4.1 in the Methodology Chapter 4, with more detailed information available in Section 4.2 on the full repair strategy framework.

6.2.2 Participants

This study involved only Saudi participants, allowing for a focused investigation of within-group variation. Participants were categorized into two groups based on the language used by the system (DVA) when presenting repair responses: Saudi speakers (SL1) who evaluated system repair strategies presented in their native language (Arabic) and Saudi speakers (SL2) who evaluated repair strategies presented in their non-native language (English). The SL1 participants were the same group included in the previous cultural background study Chapter 5, serving as a baseline group for cross-study comparisons in this thesis (see Scenario Figure 4.2).

A total of 99 Saudi participants (56 females, 43 males; average age = 32.7 years, SD = 9.21) were recruited from Prince Sattam Bin Abdulaziz University[1] through emails, posters, flyers, and snowball sampling. Of these, 49 participants interacted with the system in Arabic (SL1), and 50 in English (SL2). Each participant received a £5 Starbucks eGift Card as compensation. Ethical approval for this study was granted by the Ethics Committee in the Department of Computer and Information Sciences at the University of Strathclyde (approval number: 2781; see Appendix B)

All participants completed pairwise comparison tasks to indicate their preferences for different system-repair strategies, under one of the two language conditions (addressing RQ3). In addition, participants from both groups (SL1 and SL2) completed a survey measuring individual-level factors, including computer self-efficacy, prior experience with technology and DVAs, and demographic characteristics (addressing RQ4). For the 50 participants (SL2), who interacted with the DVA using their second language, English language proficiency was self-assessed based on the Common European Framework of Reference (CEFR) [189], categorizing participants into three overall levels: Basic (N = 7), Intermediate (N = 19), and Proficient (N = 24).

The two groups (SL1 and SL2) were found to be comparable in terms of individual-level characteristics. Independent-samples t-tests were conducted to assess the comparability of the SL1 and SL2 groups across individual-level characteristics, including age, computer self-efficacy, prior experience with technology, and prior experience with DVAs. The results indicated no significant differences between the two groups on any of these measures (all $p > .05$; see Appendix C.1 for full test statistics).

Overall, the sample rated themselves moderately on computer self-efficacy (mean = 4.43, SD =

[1]Prince Sattam Bin Abdulaziz University is located in Saudi Arabia.

0.94) using a 7-point Likert scale (1 = strongly disagree; 7 = strongly agree). They also reported moderate prior experience with general technology (mean = 4.14, SD = 0.81) and DVAs (mean = 3.57, SD = 0.93) on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The similarity in scores between the SL1 and SL2 groups strengthens the comparative analysis by reducing the likelihood that observed differences are due to underlying group differences in self-efficacy or experiences. Table 6.1 presents detailed demographic information, including interaction language and the individual-level factors assessed in this study.

Table 6.1: **Participant Demographics and Individual Factors by Interaction Language**

Individual Factors	**Native (SL1) Mean (SD)**	**Non-native (SL2) Mean (SD)**
Age	31.95 (9.57)	33.54 (8.87)
Computer Self-Efficacy (7-point Likert scale)	4.59 (0.97)	4.48 (0.85)
Experience with Technology (5-point Likert scale)	4.22 (0.85)	4.07 (0.77)
Experience with DVA (5-point Likert scale)	3.69 (0.91)	3.46 (0.96)
Language Proficiency (3 Levels)	NA	2.34 (0.71)

Higher scores indicate a higher level of the measured quantity.

Additionally, a follow-up semi-structured interview was conducted to explore participants' perspectives of communication breakdowns and repair strategies (addressing RQ4). After completing the interaction task, participants were invited to take part in an interview. A total of 15 Saudi participants (F = 4, M = 11; mean age = 31 years, SD = 6.4) participated. These interviewees were drawn from across both SL1 and SL2 groups, based on those who volunteered to take part. Participants received a £10 Starbucks eGift Card for their time.

Participants followed a structured sequence of stages throughout the study. The process began with participant briefing and screening, followed by a demographic and individual differences survey. This survey was identical for all participants, except that it included an additional self-assessment of English language proficiency for non-native (SL2) participants. Next, participants completed a series of pairwise comparison tasks involving different repair strategies. Finally, a subset of participants was invited to participate in a semi-structured interview. The overall procedure is illustrated in Figure 6.1, which visually outlines the survey-based and experimental stages, as well as the interview component, all aligned with the study's research questions.

6.2.3 Simulated Task Scenario

The same goal-oriented scenario described in Methodology Chapter 4 (Section 4.3.1) was used in this study to examine how the interaction language (native vs. non-native) influences user preferences for system repair strategies. The consistent use of this task enables reliable comparisons across

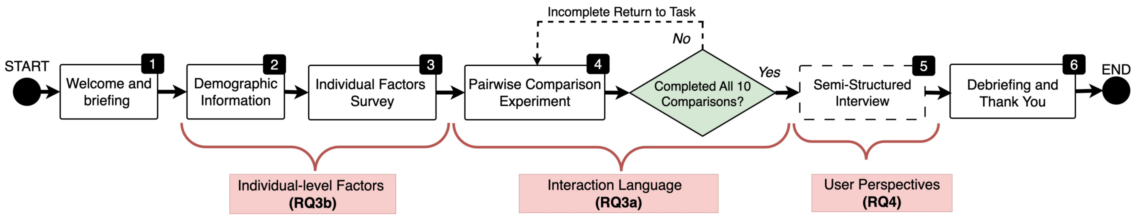


Figure 6.1: Overview of the survey-based and experimental procedures completed by participants in Study 2

studies while maintaining control over the communication context.

The author used a goal-oriented task, specifically designed within the framework of SWTS as recommended by Borlund and Schneider [34] and further discussed in [33], to assess user preferences for communication repair strategies. The scenario, booking a flight, simulates a realistic interaction with a DVA, named here as *Echo*. The scenario involves the same type of conversational breakdown, categorized as a partial non-understanding breakdown [234, 232, 233]. This occurs when DVAs successfully identify the user's general intent (e.g., booking a flight) but partially fail to recognize a specific entity (e.g., the destination: 'London'). It is assumed that the speech recognition confidence threshold for recognizing the entity falls below the acceptable level, resulting in a rejection non-understanding of the user's utterance [198, 29]. In response, DVAs immediately initiate repair strategies to address the breakdown. This approach aligns with the principles of Grounding Theory, which emphasizes the importance of immediate and collaborative actions to achieve mutual understanding in dialogue [49, 51].

Furthermore, in scenarios involving strategies such as Solve and Confirmation, the correct entity (e.g. 'London') is deliberately excluded from the system's suggestions. For instance, the system (*Echo*) might propose alternative destinations such as 'Luton,' 'Sunderland,' or 'Liverpool,' while omitting the correct option, to reduce potential biases in evaluating repair strategies, such as item priming effects [203]. This design mirrors potential real-world use cases and ensures that participants focus on assessing the main attributes of the system's repair strategies rather than favouring strategies simply because they provide the correct answer (entity: 'London').

6.2.4 Paired Comparison Experiment

To address RQ3, a comparative research design using a mixed factorial framework was employed to examine how interaction language influences user preferences for repair strategies in DVAs. The pairwise comparison experiment followed a 2×5 factorial design, with interaction language (native vs. non-native) as the between-subjects factor and repair strategy (five types) as the within-subjects factor. This mixed design enabled us to compare how each group ranked different system responses and to observe any significant differences in preferences between linguistic groups.

To collect participants' preferences for repair strategies, an online pairwise comparison experiment was conducted using Qualtrics[2]. The experiment included a text-based scenario, followed by binary paired comparisons of voice-based repair strategies. This design aimed to minimize extraneous influences, such as participant confusion when evaluating multiple repair strategies simultaneously [67]. Participants first read scenarios depicting conversational breakdowns, and then randomly listened to two different voice responses, each representing a distinct repair strategy. They were then asked to indicate their preference for one of the two strategies in handling the situation (see Figure 4.3).

Each participant evaluated all five repair strategies through a total of 10 pairwise comparisons, constituting a complete paired-comparison design in which all possible strategy pairs were presented exactly once. As a result, each repair strategy appeared an equal number of times across the experiment (four comparisons per strategy per participant). In each comparison, participants selected the single repair strategy they preferred; this choice was recorded as a binary outcome, with the selected strategy coded as 1 (preferred) and the non-selected strategy coded as 0. Preference scores for each strategy were then calculated by summing the number of times it was selected across its four pairwise comparisons, resulting in a per-participant preference score ranging from 0 (never selected) to 4 (selected in all comparisons). Further details about the experimental design are provided in Section 4.4, and the translation procedures from English to Arabic are described in Section 4.8.2 of the Methodology Chapter (4).

6.2.5 Individual-level Factors

To address RQ4, the influence of the following individual-level factors on preferences for repair strategies was also examined: computer self-efficacy, prior experience with technology and voice assistants and language proficiency. Research has demonstrated that these individual factors affect people's preferences and behaviours [17, 259, 210, 42]. The individual-level factors examined in this study and their measurement were introduced in detail in the Methodology Chapter 4 (see Section 4.5); the following subsections provide an overview of each factor as it relates to this study.

Computer Self-efficacy The concept of self-efficacy, as defined by Bandura, refers to individuals' beliefs in their capacity to execute the behaviours necessary to achieve specific performance goals [22]. It is a critical psychological construct and a robust predictor of behaviour change [23, 21, 22]. Compeau and Higgins [57, p.192] describe computer self-efficacy as "a judgment of one's capability to use a computer". Higher computer self-efficacy is associated with enhanced performance, faster adoption of new technologies, and improved ease of use and user experience [89, 138, 42].

[2] https://www.qualtrics.com/

Furthermore, Gupta and Bostrom [100] outline four types of computer self-efficacy measure, emphasizing the importance of both the technology domain and task specificity. For the study on DVAs, the scale designed for specific technologies and general tasks proposed by Gupta and Bostrom [100] was adopted. This scale was selected for two main reasons: (1) the focus of this study is on interactions with DVAs, which represent a distinct class of technology; and (2) unspecific tasks, rather than narrowly defined skills, better reflect users' perceived ability and confidence to engage in imagined goal-oriented interactions. The adapted scale consisted of five items that assess participants' self-efficacy, each rated on a 7-point Likert agreement scale (1 = Strongly Disagree to 7 = Strongly Agree), where a higher scores indicate greater self-reported confidence (see Table 4.2).

Experience with Technology and DVAs Prior studies highlight how experience with technology and digital assistants substantially shapes preferences concerning repair strategies in digital interactions [288, 17]. For example, individuals more familiar with chatbots often prefer Confirmation repair strategies, as these are frequently presented in text-based chatbots [17].

In this study, self-reported prior experience with technology and prior experience with DVAs were assessed using two items each, rated on a 5-point Likert agreement scale (1 = Strongly Disagree to 5 = Strongly Agree). These items were adapted from Ashktorab et al. [17], whose work focused on the influence of prior experience on preferences for repair strategies in chatbot-based interactions (see Table 4.2). This alignment is particularly relevant, as the current study similarly investigates how individual factors shape repair strategy preferences, albeit in voice-based interactions rather than text-based settings. Higher scores on each item indicate a greater level of the respective factor (i.e., more experience with technology or DVAs).

Language proficiency Language proficiency i.e. a user's skill level and fluency in a particular language, influences interactions and adaptation with DVAs [126, 210, 275]. Research shows that higher proficiency correlates with finding DVAs easier to understand and more pleasant to use [126, 210]. To evaluate language proficiency among SL2 participants, the Common European Framework of Reference for Languages (CEFR) was used, as participants interacted in their second language, English [189]. CEFR is commonly applied in HCI to self-assess foreign language proficiency (e.g. [241, 74, 123, 210]) and categorizes proficiency into six levels from low proficiency (A1) to high (C2) [3]. These were grouped into three broader categories: basic (A1, A2), intermediate (B1, B2), and proficient user (C1, C2) [189] (see Table 4.3).

6.2.6 Qualitative Interview Design

After completing their interaction with the DVA, participants were invited to take part in a semi-structured interview. 15 online semi-structured interviews were conducted in participants' native

[3] https://europass.europa.eu/en/common-european-framework-reference-language-skills

language (Arabic) to preserve the authenticity of responses and ensure accurate representation of their perspectives [267, 239]. The author, as a native Arabic speaker, conducted the interviews via Microsoft Teams using its audio recording functionality. Each session lasted between 32 and 46 minutes. Prior to analysis, the interviews were manually transcribed and all identifiable information was anonymized.

Two bilingual researchers (Arabic–English), including the author, independently analysed the data using inductive thematic analysis [37]. The analysis remained grounded in the original Arabic transcripts throughout. This ensured that interpretations were culturally and linguistically appropriate. Maintaining analysis in the source language ensures greater cultural and linguistic fidelity in qualitative interpretation [256, 239]. During the process, code names, initial codes, and relevant quotations were translated into English to support collaborative discussions with the co-authors [9], following established practices for translation in qualitative research [218, 2]. Following a collaborative and interpretive approach [37, 188], coding discrepancies were resolved through iterative discussions, with final codes and themes developed by consensus. In collaboration with a translator, I translated final themes, sub-themes, and quotations presented in this study.

Qualitative analysis was conducted using MAXQDA 22.80[4] software. In the Methodology Chapter 4, more detailed information about the qualitative interview design and conducted analysis can be found in Section 4.6.

6.3 Results

This section presents the results of the study in two parts. First, it reports the quantitative analysis of participants' preferences for five different repair strategies in SDSs, focusing on: (a) the effect of interaction language (addressing RQ3a), and (b) the influence of individual-level factors (addressing RQ3b). Specifically, it examines whether there are significant differences in repair strategy preferences between native (SL1) and non-native (SL2) system-language conditions, and how participant characteristics, such as self-efficacy, prior experience, and language proficiency, affect these preferences.

Second, this section presents the qualitative findings that explore participants' perspectives on repair strategies and their experiences with communication breakdowns in SDSs (addressing RQ4). All analyses were conducted using IBM SPSS Statistics 29.0. As the data did not meet the assumptions of normality, non-parametric tests were employed [59]. To account for multiple comparisons, a Bonferroni correction was applied where appropriate, adjusting the significance threshold to $p < 0.005 (0.05/10)$ based on the total number of pairwise tests conducted [62, 265, 59]. All statistical tests were conducted using two-tailed significance levels.

[4]https://www.maxqda.com/

Table 6.2: **Mean preference scores (and standard deviations) for each repair strategy, separately for SL1 and SL2 groups.**

Strategy	SL1 Mean (SD)	SL2 Mean (SD)
Social	2.46 (1.08)	2.34 (1.18)
Confirmation	1.79 (1.45)	2.50 (1.40)
Information	1.77 (1.10)	1.16 (1.09)
Solve	1.61 (1.22)	1.74 (1.20)
Ask	2.34 (0.99)	2.26 (1.12)

6.3.1 Effect of Interaction Language (RQ3a)

6.3.1.1 Overall Descriptive Statistics

As outlined in Section 4.7.1, preference scores were determined by counting how frequently each repair strategy was selected across pairwise comparisons, resulting in scores ranging from 0 to 4. Table 6.2 provides the descriptive statistics for the five repair strategies, showing the mean and standard deviation of the number of times each strategy was selected by participants in the SL1 and SL2 groups. These results present an initial overview of participant preferences prior to conducting inferential statistical analyses.

Among SL1 participants, the Social strategy received the highest mean preference score (M = 2.47, SD = 1.08), followed closely by Ask (M = 2.35, SD = 0.99). In contrast, Confirmation, Information and Solve were less frequently selected (M = 1.79, 1.77 and 1.61, respectively). On the other hand, SL2 participants showed the highest preference for the Confirmation strategy (M = 2.50, SD = 1.40), followed by Social (M = 2.34, SD = 1.18), while Information received the lowest mean score (M = 1.16, SD = 1.09).

Several repair strategies demonstrated notable differences in preference between native (SL1) and non-native (SL2) groups. The Confirmation strategy was selected more frequently by SL2 participants (M = 2.50) than by SL1 participants (M = 1.79). Conversely, the Information strategy was preferred more by SL1 participants (M = 1.77) than by SL2 participants (M = 1.16). Both groups showed consistently lower preference for the Solve strategy, where the system provided alternative options to the user's intent.

6.3.1.2 Overall Within-Group Differences: Friedman Test

This subsection presents the results of within-group analyses for both SL1 and SL2 participants. To examine whether participants showed statistically significant differences in their preferences among the five repair strategies, within-group comparisons were conducted separately for each interaction group. Given the non-parametric nature of the data, the Friedman test was used to assess overall differences in preference rankings within each group [59].

The Friedman test results revealed significant differences in the mean ranks of the five repair strategies for both groups: SL1 participants ($\chi(4) = 14.00, p = .007$) and SL2 participants ($\chi(4) = 26.99, p =< .001$). These results indicate that, within each group, participants demonstrated meaningful variation in their preferences for different repair strategies. Figure 6.2 presents the mean ranks for each subgroup rankings for SL1 and SL2 participants.

Notably, SL1 participants showed the strongest preference for the Social strategy, highlighting the value placed on apologies and social rapport. While not the top choice, the Information strategy was more favoured by the SL1 group than by the SL2 group. Conversely, the SL2 group showed a clear preference for the Confirmation strategy, possibly reflecting their need for clarity and assurance when communicating in a non-native language. Preferences for Social and Ask strategies were also strong in the SL2 group but slightly less pronounced compared with the SL1 group. These differences highlight the impact of interaction language on repair strategy preferences, with non-native users distinctly favouring strategies that confirm understanding.

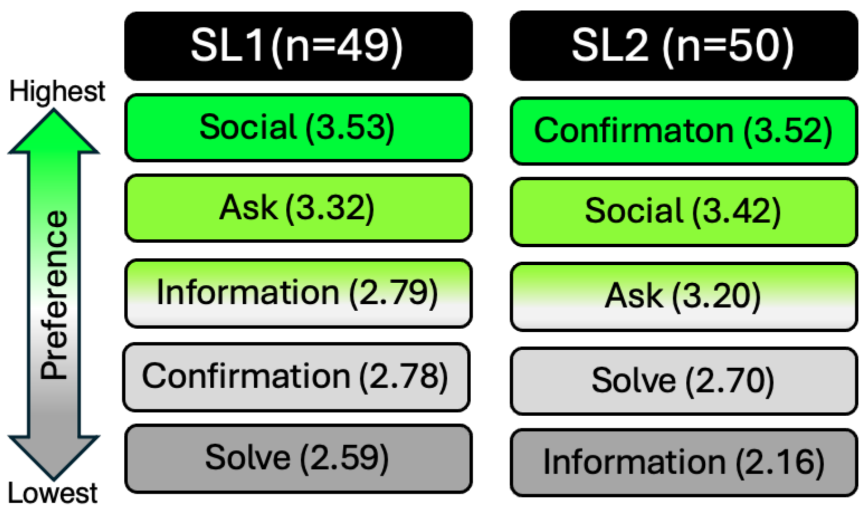


Figure 6.2: **Preferences for Repair Strategies: Arabic native users (SL1), and English non-native users (SL2) based on Friedman Mean Rank Tests. The figure displays the mean ranks for each repair strategy, where higher values indicate greater preference. Strategies are ordered from highest (Green) to lowest preference (Gray); vertical spacing is illustrative only.**

6.3.1.3 Within-Groups Pairwise Comparisons: Wilcoxon Signed-Ranks Test

To better understand specific pairwise differences in strategy preferences within each interaction group, post-hoc Wilcoxon Signed-Rank Tests were conducted for each pairwise comparison among the five strategies. This analysis identifies exactly which pairs of repair strategies differed significantly from one another within each interaction group. A Bonferroni correction [265] was applied to control for multiple comparisons, adjusting the significance threshold to $p < .005$ (0.05/10) [62]. Adjusted p-values are shown in Table 6.3.

The tests revealed that, across all participants, the Social (Z = –5.464, r = .55), Ask (Z = –4.716, r = .47), and Confirmation (Z = –3.064, r = .31) strategies were significantly preferred over the Information strategy. Within the SL1 group, Social was significantly preferred over both Information (Z = –3.066, r = .31) and Solve (Z = –3.043, r = .31). For the SL2 group, Social (Z = –2.090, r = .21), Ask (Z = –3.994, r = .40), and Confirmation (Z = –3.786, r = .38) were significantly preferred over Information. Additionally, SL2 participants preferred Confirmation over Solve (Z = –2.971, r = .30).

The results of the Wilcoxon Signed-Rank pairwise comparisons (Table 6.3) indicate that certain repair strategies consistently emerged as significantly preferred within interaction groups. Among all participants combined, the Social and Ask strategies were significantly preferred over Solve ($p < .001$), and Information ($p < .001$). This highlights a clear preference for socially engaging and concise responses. In the native (SL1) group specifically, participants showed a significant preference for the Social strategy compared to Information and Solve ($p = .002$ for both comparisons). Conversely, participants in the non-native (SL2) group significantly favoured Confirmation over both Information ($p = .001$) and Solve ($p = .003$), suggesting that non-native users prefer explicit confirmations. Additionally, the Information strategy was consistently ranked lower, especially among SL2 participants, who notably favoured direct confirmations or socially engaging responses over explanatory feedback.

6.3.1.4 Between-Group Comparisons: Mann–Whitney Tests

To examine differences in strategy preferences between native (SL1) and non-native (SL2) interaction groups, between-group comparisons were performed using Mann–Whitney U tests (Bonferroni-adjusted significance level, $p < 0.005$) [59]. Before conducting these tests, the dataset was restructured into long-format at the comparison level, where each row represented an individual pairwise preference judgment. Since each participant completed 10 comparisons, with each repair strategy appearing four times per participant, this restructuring resulted in 396 preference judgments across all 99 participants, corresponding to 1,980 individual strategy-level judgments. Preference scores for each strategy were calculated using mean ranks derived from these judgments, allowing us to identify the most preferred strategies and determine whether these preferences varied by interaction

Table 6.3: **Adjusted Significance Levels from Wilcoxon Signed-Rank Tests for Each Pairwise Comparison Using Bonferroni Correction ($p < 0.005$); '**' Indicates Significant Values.**

Paired Comparisons of Strategies	All Participants p-value	SL1 p-value	SL2 p-value
Social vs. Confirmation	0.272	0.043	0.630
Social vs. Information	0.000**	0.002**	0.000**
Social vs. Solve	0.000**	0.002**	0.045
Social vs. Ask	0.581	0.574	0.814
Confirmation vs. Information	0.002**	0.854	0.001**
Confirmation vs. Solve	0.014	0.591	0.003**
Confirmation vs. Ask	0.344	0.059	0.582
Information vs. Solve	0.299	0.558	0.037
Information vs. Ask	0.000**	0.010	0.000**
Ask vs. Solve	0.002**	0.010	0.077

language.

As shown in Figure 6.3, significant differences between the two groups emerged for the Confirmation and Information strategies. Specifically, SL2 (non-native) participants showed a significantly stronger preference for the Confirmation strategy compared to SL1 (native) participants ($U = 16150$, $n_{SL1} = 196$, $n_{SL2} = 200$, $p < .001$). Conversely, the Information strategy was significantly more preferred by SL1 participants than by SL2 participants ($U = 16584$, $n_{SL1} = 196$, $n_{SL2} = 200$, $p = .001$). These results indicate that interaction language is associated with systematic differences in repair strategy preferences between native and non-native conditions.

Although differences in preferences for the remaining three strategies (Social, Solve, and Ask) did not reach statistical significance, notable patterns emerged. Both Social and Ask strategies were highly preferred by participants in both groups, potentially reflecting a cultural preference among Saudi users for strategies that are socially engaging or concise and direct, particularly during goal-oriented tasks. In contrast, the Solve strategy consistently received lower preference rankings across both language conditions, indicating a lower overall preference for repair strategies that present multiple options. No significant differences were observed between native and non-native interaction language conditions for this strategy. While this pattern may reflect interactional challenges associated with option-based repair, such as increased processing demands in voice-only interactions, the specific reasons underlying this preference were not directly assessed in the present study.

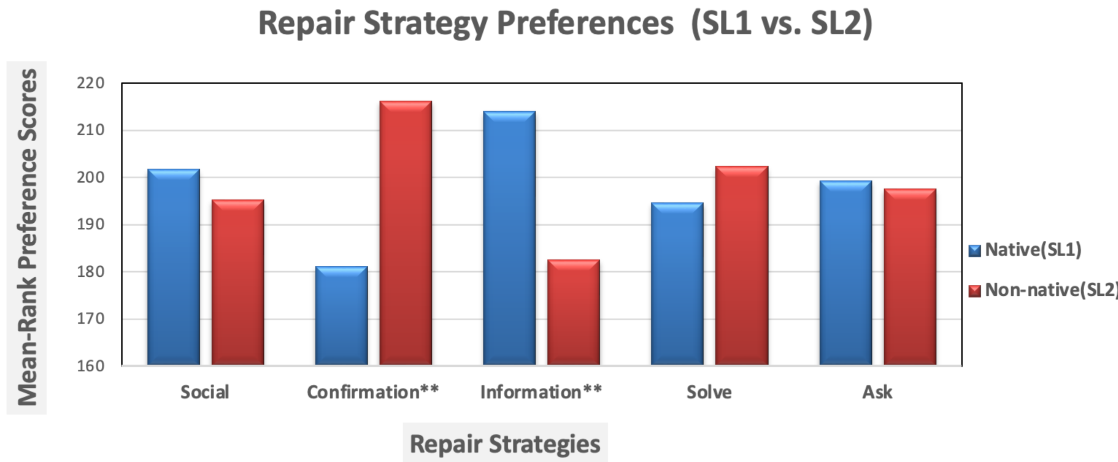


Figure 6.3: **Bar chart comparing the mean preference scores of five repair strategies between native (SL1) and non-native (SL2) Saudi participants. Scores reflect the cumulative ranking of each strategy across all pairwise comparison trials (N = 396), where higher values indicate greater preference. Strategies marked with (**) showed statistically significant differences in preference between language groups (Mann–Whitney U test, Bonferroni-adjusted $p < 0.005$).**

6.3.2 Effect of Individual Factors (RQ3b)

In this subsection, the influence of individual-level factors such as computer self-efficacy and prior experience with technology and DVAs on preferences for the five repair strategies is examined, along with demographic variables such as gender and age. For the non-native (SL2) group, the effect of language proficiency level (basic, intermediate, and proficient) on these preferences was also examined.

A statistical modelling approach was employed for the analysis. Binary logistic regression models were constructed for each repair strategy to predict its likelihood of being the preferred choice. Each model incorporated the interaction between language and individual factors as independent variables. In total, five binary logistic regression models were run, one for each repair strategy, with the focus placed on statistically significant results. Full binary-logistic regression outputs for all repair strategies, including non-significant predictors, are provided in Appendix C.

The assumptions underlying the regression models were tested, including multicollinearity and the presence of influential outliers. Multicollinearity was assessed using variance inflation factor and tolerance values [78]. All variance inflation factor values were below 5 and tolerance values above 0.1, indicating no multicollinearity concerns across the five models. Additionally, standardized residuals and Cook's distance values were examined to assess the presence of outliers or influential cases [58]. Standardized residuals fell within the acceptable range (–3 to +3), with observed values ranging from –1.96 to 2.16. Cook's distance values were all well below 1.0 (maximum = 0.057), suggesting no cases had disproportionate influence on the models [58, 78]. Centred leverage values were also low (maximum = 0.113), indicating that no participants exerted undue leverage.

6.3.2.1 Experience with Technology and DVAs

Participants with higher levels of prior experience using DVAs were found to be significantly more likely to prefer the Confirmation strategy ($\beta = 0.361, SE = 14, p = 0.012$), indicating that increased familiarity with voice assistants raised the odds of preferring confirmation-based repairs. Conversely, these participants were significantly less likely to favour the Information strategy ($\beta = 0.404, SE = 145, p = 0.005$), indicating that experience level is associated with reduced preference for explanation-based repair. No significant effects were observed for general technology experience in predicting repair strategy preferences in the regression models.

($r(97) = -.08, p = 0.10$).

6.3.2.2 Computer Self-Efficacy

Participants with high confidence in using DVAs were less likely to favour the Solve strategy when the DVA provided options to repair breakdowns ($\beta = -0.323, SE = 0.147, p = 0.028$). Additionally, participants with high computer self-efficacy interacting in their non-native language (SL2) were more likely to prefer the Confirmation strategy ($\beta = 0.49, SE = 0.24, p = 0.03$), whereas those with high computer self-efficacy interacting in their native language (SL1) were more likely to prefer the Ask strategy ($\beta = 0.40, SE = 0.20, p = 0.45$). These findings indicate that self-efficacy influences strategy preferences differently depending on the language context.

6.3.2.3 Language Proficiency

Logistic regression analysis indicated that language proficiency did not significantly predict preferences for repair strategies when controlling for other individual-level factors. This suggests that language proficiency alone may not be a strong determinant of repair strategy preference within the current experimental context.

6.3.2.4 Gender and Age

No significant differences in repair strategy preferences were found based on gender. However, age emerged as a significant predictor for several strategies. In the sample, lower participant age (i.e., younger individuals relative to the sample's overall age distribution) was associated with a lower likelihood of preferring the Solve strategy($\beta = -0.03$, $SE = 0.01$, $p = 0.01$), where the DVA provides multiple options to repair misunderstandings. While this pattern suggests that age may be associated with differing expectations of system behaviour, the mechanisms underlying this relationship remain speculative and warrant further qualitative investigation.

Similarly, when interacting in a non-native language, lower participant age was associated with greater preference for the Ask strategy ($\beta = 0.04$, $SE = 0.02$, $p = 0.02$) and lower preference for the Confirmation strategy ($\beta = -0.06$, $SE = 0.02$, $p = 0.003$). These results suggest that younger participants may favour more direct forms of repair over strategies that involve confirmation.

6.3.3 Users' perspectives of Repair Strategies (RQ4)

To address RQ4, semi-structured interviews were conducted to explore users' perspectives on different repair strategies and communication breakdowns, providing insights into how they interact with DVAs. During the interview phase, participants were not directed to answer with reference to any specific commercial DVA. As a result, their responses referenced various commercially available assistants (e.g., Siri, Alexa), based on their own prior experiences. The following three themes capture the complexity of these interactions, emphasizing prior user experience, repair strategy preferences, and DVA adaptation to language. Table 6.4 provides an overview of the three main themes and their sub-themes identified through thematic analysis of participant interviews.

Table 6.4: **Summary of Themes and Sub-themes from Thematic Analysis**

Main Theme	**Sub-themes (*quotes*)**
- User Experience and Adaptation with DVAs	- **Prior Experience** (*Siri still gives irrelevant answers [P4]*) - **Interaction Language Used** (*I avoid using it in English for daily use [P9]*) - **User Adaptability** (*I speak slowly in Arabic so it understands [P5]*)
Repair Strategy Preferences and Perspective	- **Directness and Minimum Level Required** (*Don't want to deal with extra information [P8]*) - **Explanation and Clarity** (*Confirmation shows me what it understood first [P7]*) - **Naturalness and Intelligence** (*Providing options feels unintelligent [P12]*)
- DVA Adaptation to Language Proficiency and Speech Input	- **Dynamic Language Proficiency Adaptation** (*It should adapt to my language level [P3]*) - **Personalized Speech Recognition Adaptation** (*My accent makes it hard to understand me [14]*)

6.3.3.1 Theme 1: User Experiences and Adaptation with DVAs

This theme addresses how users interact with DVAs based on their prior experiences and adaptability.

Prior Experience with DVAs Participants generally reported positive experiences with DVAs and noted substantial improvements over time. However, they still encountered occasional difficulties understanding requests, with some responses perceived as unintelligent or irrelevant:

> *"I've been using Siri since 2012...his performance has definitely gotten better over the years, and the results are faster and better in terms of understanding...but it's still he [Siri] still gives answers that are irrelevant."*[P4]

> *" I use Siri and most of the time when I'm driving, I don't need to hold the phone. My experience with it is very good, I wouldn't say excellent, but very good."*[P2]

Common tasks included phone calls, searching for locations, playing music, and managing calendars. However, interactions were primarily limited to simple tasks: most participants avoided DVAs for complex actions due to a lack of trust in the DVA's accuracy. As P5 noted, they were "*used to giving commands by typing*" and avoided using DVAs for tasks such as booking flights, as "*voice is harder to be understood than typing the commands.*" While some aspects of their experience had been satisfactory, reliance on DVAs was limited:

> *"The overall experience was okay, but I didn't rely on it, because I realized that it didn't provide me with a better service than when I did type by myself."* [P6]

Interaction Language Used Most participants utilized DVAs in both their native and second languages, generally showing a preference for using their native language. Choice of interaction language appeared to depend on the task and the purpose of the interaction.

> *"When I need it on a daily basis, I don't want to use it in English. I want something that speaks to me in my native language. Why do I need to change my native language when the machine can be made to understand it?"*[P9]

> *"I mostly use it in Arabic, but I use it in English if I need to translate or know the meaning of a word in English."* [P2]

Participants also observed a noticeable difference in the quality of responses from DVAs based on the interaction language used. They reported that responses in English tended to be more accurate and detailed, providing comprehensive information crucial for tasks requiring detailed understanding. As P10 remarked, using DVAs in Arabic often led to "*poor, often inaccurate and sometimes just plain silly*" results, likely due to the limited availability and lower quality of Arabic content sources. This contrast was echoed in participants' reflections on information richness:

> *"I feel that there is a big difference in the sources of information from which Siri answers. I mean, answering the same question in Arabic is less efficient than answering it in English... Sometimes when I ask about a discoverer or inventor in Arabic, he usually gives me Wikipedia, but in English, he [Siri] gives me more than one website and source for the same information."* [P4]

Participants expressed uncertainty and frustration regarding the causes of misunderstandings when using DVAs, especially when interacting in their second language. They often questioned whether these misinterpretations stemmed from their own pronunciation, voice volume, or the

DVA's inability to retrieve relevant information. Some were disconcerted by unexpected shifts in interaction language such as when speaking Arabic and mentioning an English term, which led to a sudden switch to English. Others were frustrated by being forced to switch from voice to text input, especially in hands-free situations, such as driving. These issues disrupted the natural flow of interaction and added to their dissatisfaction:

> *"I honestly didn't know if it was me or the program[Siri],... I was raising my voice in an exaggerated manner and then he realized "* [P2]

> *"The audio responses are great, but the text responses annoy me. Sometimes when I'm driving, I can't write or read, and the issue is he[Siri] gives me options and says 'choose'. Well, I'm still driving."*[P4]

> *"Sometimes I speak to him in Arabic, and if I say an English term, he[Google Assistant] suddenly switches to English. Then I have to continue in English..I don't like that."* [P5]

User Adaptability Participants adapted their behaviour and language to enhance interactions with DVAs. When faced with misunderstandings in their native language (Arabic), they often responded by repeating their input and clarifying their speech, either by simplifying their vocabulary or switching to Standard Arabic. Conversely, when interacting in English, participants typically raised their voice and frequently repeated commands, focusing more on improving their pronunciation instead of changing their vocabulary:

> *"When it[Google Assistant] does not understand me, I keep repeating myself ..I don't know but I try to focus hard on my pronunciation of the English words to make sure it understands what I'm trying to say."* [P12]

> *"Sometimes, it feels like he just doesn't get what I'm saying, you know? I talk fast and in my local Arabic, so I slow it down and choose simpler words, hoping he'll catch on."*[P10]

Additionally, to avoid communication breakdowns, participants sometimes adjusted their device settings or changed the interaction modality to better accommodate speech recognition in their second language. For instance, one participant preferred using the voice-to-text feature to ensure their complete message was transcribed and sent as text, thus avoiding interruptions (*barge in*) before they had finished stating their request:

> *"Sometimes, I just use the feature where I talk in English, and it turns what I say into text. It lets me think while I'm talking, without any interruptions."* [P13]

"..when I ask Siri to 'Call Ahmed' in English. It searches for 'Ahmed' spelled in English and can't find him because he[Ahmed] is saved in Arabic. To resolve this, I had to change most of the names in my contact list to English so that Siri could recognize them." [P4]

6.3.3.2 Theme 2: Repair Strategy Preferences and Perspectives

This theme explores participants' preferences for repair strategies used by DVAs during communication breakdowns. It highlights a preference for strategies that are efficient, provide clear explanations, and maintain the natural flow of conversation.

Directness and Minimum Level Required Participants expressed a strong preference for direct and minimal repair strategies when the system encounters communication breakdowns. They particularly valued the efficiency and conciseness of the Ask and Confirmation strategies, which quickly clarify misunderstandings without unnecessary details. As P8 noted, they were "*more interested in solving the issue quickly*" than receiving extra information. This desire for simplicity and clarity was echoed in other responses:

"When it doesn't understand me, I really prefer the system to be straightforward and clear. Then, I'll just repeat my request." [P13]

However, despite the preference for directness and conciseness, participants also emphasized the importance of politeness in interactions. They favoured polite markers such as I'm Sorry or Please when a breakdown occurred, which helped make the conversation feel both more human-like and more culturally aligned with Saudi conversational norms. As P9 noted, using polite words is like "*asking for permission to have it repeated*", which can be especially important when users feel uncomfortable repeating themselves. While extensive and repeated apologies were seen as unnecessary, given the DVA's artificial nature, completely omitting politeness was perceived as making the interaction seem robotic and overly authoritative:

"It's okay to say 'sorry,' but not like a big apology. When dealing with the system, I think it should be quick and easy to communicate, not like when talking to other people." [P1]

"Using polite words matters, like when you're talking to someone and say, 'Sorry, I don't understand this point.' It's like you're asking for permission to have it repeated, especially since not everyone is okay with repeating things, especially more than once."[P9]

Additionally, the importance of apologies varied depending on the task context. In time-sensitive situations, such as emergencies, participants prioritized speed and efficiency over politeness. In more casual settings, a brief yet polite acknowledgment of the breakdown was valued for contributing to a more human-like interaction:

"Sometimes, if the request is straightforward or urgent, I don't really need polite words to get the job done. Still, it's good to keep some level of politeness." [P15]

"I think the phrases should vary depending on the situation and what's being asked. Like, don't give me long phrases or apologies in urgent situations—finding the nearest hospital, for example. But for something less critical, like the nearest restaurant, it's not a big deal." [P12]

Explanation and Clarity Participants found repair strategies that provided explanations particularly useful, as these helped to clarify the DVA's understanding levels. The Confirmation, Information, and Solve repair strategies were seen as engaging and exploratory, as they directly and indirectly revealed the DVA's comprehension. For instance, the Confirmation repair strategy was valued for its ability to demonstrate the DVA's grasp of user input by affirming the city requested. However, frequent confirmations were perceived as frustrating and detracted from human-like conversations.

"I like this strategy[Confirmation] because it explains what it understands first and doesn't do anything wrong before confirming." [P7]

"It would be better if it [Siri] explained the reason, because sometimes I can hear clearly but I'm not focused on the noise around me. With that clarification, I can process things faster." [P6]

Similarly, the Solve strategy, which offers multiple options, was deemed helpful because it showed that the DVA partially understood the user's intent by narrowing down the possibilities to a set of options. As P5 explained, this approach helps users determine whether the system is "*close to understanding or way off*", especially when the suggestions differ from the intended request. Nonetheless, this strategy could also introduce unnecessary information, particularly if the suggestions did not closely match the user's specific needs:

"I don't like it when it gives me options; I'd rather it just asks me to repeat the request. The options can be long and sometimes they're not even related to what I asked for." [P6]

Direct explanations embedded within Information repair strategies were valued, particularly after several misunderstandings, rather than immediately following the first breakdown. Initially, such detailed explanations could be seen as premature or overly complex. This could potentially complicate the interaction rather than simplify it:

"I think this strategy [Information] depends on the situation. If it's something serious, it should be made clear the second time around. For regular stuff, the third time might be okay. But the first time, explaining the reason isn't that important." [P4]

"The first time should be an apology and a request to try again. If it still doesn't understand on the second try, then it [Siri] should explain why it can't understand me" [P9]

Naturalness and Intelligence Participants perceived the Solve repair strategy, which provides multiple options, as less natural. While some appreciated the variety of choices, others felt that offering incorrect options made the DVA "*seems unintelligent and far from how a human would interact*" [P12], falling short of expectations for conversations with current AI technology. Still, some found value in the strategy:

"This strategy [Solve] is excellent because it shows that the program is on track with me. It's clear it understands something, but issues arise since words can sound similar, and people sometimes don't understand similar-sounding words." [P15]

When none of the provided options were correct, some participants appreciated having a 'none of these' option. This allowed them to bypass incorrect suggestions and offered a way to reset or start the interaction over. A few participants found the 'none of these' option unnatural and redundant, preferring to reject suggestions by saying 'no' then stating the correct city:

"I like having the 'none of these' option. It's nice because it at least eliminates the wrong choices." [P1]

"'None of these' just doesn't work for me, it's irritating. It doesn't seem smart at all, so these options don't feel like they're close to a real human interaction."[P9]

Confirmation strategies were often viewed as intelligent because they signalled that DVAs were actively engaging with and supporting the user. Participants particularly valued this approach during tasks such as booking flights, where confirming details before taking action is crucial. One participant noted it was helpful because *"it checks before doing anything... I can just say yes"* [P5]. This strategy was also perceived as a natural repair strategy, as it mimics human-like conversational patterns where misunderstandings are clarified through confirmation. For non-native speakers, confirmation was especially appreciated because it simplified interactions, relieving users of the burden of repeating entire utterances, allowing them to respond with a simple 'yes' or 'no.':

"If he doesn't understand me when I speak in English, I first think it might be my pronunciation or accent. That's why I prefer to respond with just 'yes' or 'no.' I'm less likely to mispronounce if my answers are short and to the point." [P15]

6.3.3.3 Theme 3: DVA Adaptation to Language Proficiency and Speech Input

This theme explores the importance of DVAs adapting interactions based on the nature of the task, specific dialects, cultures, and accents of non-native speakers.

Dynamic Language Proficiency Adaptation Participants emphasized the critical role of DVAs adjusting language based on users' proficiency, especially in non-native languages. Sometimes, participants interact with DVAs using second or third languages to improve their fluency and conversational skills. Their preference for language complexity depends on the task. For goal-oriented tasks such as booking flights, simple and direct language is preferred to ensure clarity and prevent misunderstandings. In contrast, when the objective is learning and practising, participants often choose to challenge themselves by increasing the language difficulty to enhance their linguistic capabilities. As P3 noted, it's ideal when the system adapts: "*If I speak fast, it should respond quickly. If I'm a beginner, it could suggest ways to improve and use simpler phrases.*" This desire for adaptive interaction was also reflected by a participant learning Turkish, who preferred high-level responses with colloquial expressions:

> *"I also fluently speak Turkish...and I practice with him [Siri] in colloquial Turkish. If he responds to me at a beginner level, that's not okay with me. I know Turkish well, and he should recognize that. But he keeps responding like I'm a beginner, which feels wrong to me."* [P9]

Personalized Speech Recognition Adaptation Participants frequently emphasized the need for DVAs to adapt not only to Arabic but to regional dialects and cultural nuances within Saudi Arabia. Saudi dialect has received less attention in NLP development compared to other Arabic dialects, and differs significantly from them [7]. This adaptation is crucial to ensure that DVAs align with participants' habits and cultural contexts, thereby reducing frequent misunderstandings:

> *"It's important that the program considers culture because some words have different or even opposite meanings in Saudi Arabia compared to other Arab countries."* [P11]

> *"Siri keeps misunderstanding me because I use my specific Saudi dialect [Najdi], where I add a /š/at the end of words. It should be able to adjust and learn from these dialect variations to understand me next time "*[P9]

Such personalized adaptation of speech recognition is equally important for non-native users. When interacting in a second language, participants reported challenges in being understood by DVAs due to their accents. In mitigation, some participants attempted to adjust or neutralize their accents during interactions with DVAs:

> *" I feel the system understands me better in Arabic. I think my accent in English makes it hard for it to understand me, so I often adjust how I speak during interactions."* [P14]

> *"sometimes even mistakes 'paragraph' for 'bar graph'...The program should recognize multiple accents because our English accent is different from those in Asia or Africa."*[P7]

Personalized speech recognition adaptation is essential to address users' diverse experiences and challenges, bridging human linguistic diversity with digital comprehension.

6.4 Discussion

This section discusses the results in three stages. First, an overview of participants' overall preferences for repair strategies in DVAs is presented, integrating both quantitative and qualitative insights. Second, the impact of interaction language is examined by comparing SL1 and SL2 groups. Finally, the influence of individual-level factors, including prior experience and computer self-efficacy, is explored.

6.4.1 Repair Strategy Preferences (RQ3a & RQ4)

The analysis of user preferences for repair strategies in DVAs indicated a pronounced preference for strategies that combine efficiency and clarity while maintaining a natural conversational flow. Specifically, participants favoured concise and minimal repair strategies, such as the Ask and Confirmation strategies. These approaches align with Grice's Maxim of Quantity, optimizing cognitive load by providing only necessary information[96].

Contrastingly, the qualitative insights revealed a strong appreciation for DVAs providing explanations of breakdowns, as implemented in the Information repair strategy. However, quantitative results ranked this strategy as less preferred on initial repair attempts. This discrepancy can be explained by the sequence in which strategies are deployed. The Information repair strategy was initially less preferred but became more popular after unsuccessful attempts with other strategies such as the Ask strategy. Based on Grice's conversational maxims, Miehling et al. [169] demonstrates that while the information content in AI dialogues should ideally unfold over multiple turns as requested by the user, the initial deployment of the Information repair strategy might satisfy the Maxim of Quantity but potentially violate the Maxim of Manner, which requires that information be presented in a clear, orderly fashion.

Additionally, the preference for directness does not diminish the importance of politeness and cultural sensitivity in interactions. The findings indicate that the inclusion of polite markers such as "sorry" or "please" significantly enhances the perceived humanness and friendliness of the DVAs. Notably, the Social repair strategy, politely apologize, was the highest-ranked among all Saudi participants, regardless of their interaction language. This preference aligns with cultural norms and communication styles ingrained in Saudi society. According to Hofstede's cultural theory [112], Saudi Arabia is considered a collectivist society where maintaining harmony and relationship-preserving communication are prioritized. High power distance and collectivism in Saudi culture encourage respectful, deferential communication, which explains participants' stronger preference for polite system responses. In such contexts, an apologetic tone is often seen not as inefficiency but as a sign of respect. Systems that apologize for breakdowns are more usable and generally accepted than those that do not apologize [196].

By contrast, users from individualistic cultures may prioritize efficiency and clarity over social harmony. While this study primarily focused on the effects of interaction language, the observed preferences also reflect deeper cultural values that influence user expectations and interaction styles. Future research could explore these cultural dimensions more explicitly, particularly by comparing user groups across different cultural backgrounds. Recent studies emphasize the importance of culturally sensitive and socially integrated CAs, demonstrating how co-designing interactions with end users can enhance user experience [228] and fulfil users' politeness expectations [155].

The qualitative findings also suggest that participants value polite apologies to reduce frustration and robotic interaction but caution against overuse. Repeated apologies without resolving the issue can lead to increased frustration, as multiple repair attempts without success may exacerbate irritation. Furthermore, preference for such Social repair strategies is highly context-dependent. This demonstrates the pervasive and consistent nature of politeness in interactions, relying on a common set of linguistic markers for social coordination. The importance and necessary quantity of each marker is often dictated by the context, presenting challenges for studies in computational politeness [207]. In scenarios requiring urgent responses, direct repair strategies that prioritize efficiency and clarity are more favourable. This aligns with prior research indicating that although politeness generally enhances social interactions, it can impede the need for immediate clarity in high-stakes situations[31]. Users, therefore, may favour straightforward communication over polite language to prevent potential delays in critical contexts.

The Solve strategy was ranked lowest among all repair strategies in the quantitative results. This was consistent across participants (SL1 and SL2), regardless of interaction language. While some valued this approach for enhancing functionality and allowing them to exclude incorrect options, others perceived it as unnatural and less human-like. Two possible explanations emerge. First, in the experiment, all provided options were incorrect, which may have reduced the perceived usefulness of the strategy. However, providing incorrect options reflects real-world use cases. Second, offering multiple options may have imposed additional cognitive load on users, making interaction more demanding, especially in voice-only settings. Interestingly, this finding contrasts with previous research, where providing options was the most preferred repair strategy [17]. However, that study was conducted using text-based scenarios and included the correct option among the choices, which may explain the discrepancy. Compared to voice-only settings, research shows that cognitive workload is reduced when voice interaction is supported by other modalities, such as visual or gesture interaction [65, 194, 275, 211].

6.4.2 Interaction Language Effect (RQ3a)

A distinct pattern emerges when contrasting the responses of SL1 and SL2 participants, particularly with respect to the Information strategy. Native users showed a pronounced preference for this strategy when compared with non-natives. This inclination could be attributed to their greater comfort and reduced cognitive load when processing complex information in their native language. Language familiarity likely facilitated deeper comprehension of the information presented, enabling participants to effectively utilize and respond to breakdown explanations without the burden of language-related cognitive strain.

Conversely, non-native (SL2) participants may find overly detailed information overwhelming, reducing their preference for this repair strategy. The findings align with previous research, which shows an increase in cognitive load for non-native interlocutors compared to natives during dialogue interaction [275, 174, 209], as linguistic dimensions may significantly differ from their native language [263]. This highlights that DVAs should adapt their repair strategies based on the user's language proficiency by simplifying and optimizing information to reduce cognitive demands on non-native users.

Additionally, the preference for the Confirmation strategy differed significantly between SL1 and SL2 participants. SL2 participants have a marked preference for the Confirmation strategy. This can be largely attributed to the strategy's simplicity and directness, which reduces the cognitive burden on users interacting with DVAs in a non-native language. The Confirmation strategy typically requires users to respond with a simple affirmative or negative, such as "yes" or "no". This straightforward interaction minimizes the need for non-native users to construct complex sentences or entirely repeat utterances. This approach could significantly reduce the cognitive workload associated with language production in a second language [275, 273]. Previous research [273] investigated the mental workload of native and non-native users interacting with voice interfaces. Non-native users experienced significantly higher mental workload than natives across smart speakers and smartphone devices.

Furthermore, non-native users often seek greater clarity to avoid misunderstandings due to less fluency. The Confirmation strategy meets this need by providing clear feedback, reducing the need for complex language production, and boosting confidence in using DVAs. In contrast, native users may be more comfortable with repeating input. Thus, confirmation serves as a key facilitator of accessibility in non-native interactions [274].

6.4.3 Individual-level Factors (RQ3b)

With respect to prior experience with DVAs, the results showed that more experienced users tended to favour the Confirmation strategy. This stems from their familiarity with DVAs and the value they place on reassurance that the system understands them. Conversely, these users showed less preference for strategies that provide detailed explanations of errors, such as the Information strategy. This may be due to their understanding of the capabilities and limitations of these systems. These findings are consistent with Ashktorab et al. [17], who similarly found that participants with more prior chatbot experience were more likely to prefer Confirmation strategies. However, while Ashktorab et al [17] reported that participants with greater general technological experience tended to favour explanation-based strategies, the results of the current study showed no significant effect of general technology experience, aside from a weak negative correlation with preference for the Solve strategy. This perspective was corroborated by qualitative feedback from the participants, who noted there is often no need to explain an error such as background noise during the initial repair attempt, as users are typically aware of such causes and can adjust their environment to resolve this.

Higher computer self-efficacy generally correlates with a preference for repair strategies involving less DVA assistance, such as a reduced preference for the multiple-options approach in the Solve strategy. This indicates that individuals with greater confidence in their technological abilities prefer more streamlined and autonomous interactions, suggesting a desire to manage and direct interactions without unnecessary system intervention. This aligns with research where participants with lower self-efficacy found chatbot interactions more helpful and less frustrating, appreciating the supportive features [42]. These results collectively underscore the influence of self-efficacy on user interaction preferences, highlighting a spectrum where users with lower self-efficacy benefit from clear, supportive interactions as they build confidence, while those with higher self-efficacy seek efficiency and minimal guidance.

In relation to interaction language, the findings suggested that computer self-efficacy may influence repair strategy preferences differently depending on whether users interact in their native or non-native language. Non-native users with higher self-efficacy demonstrated a preference for the Confirmation strategy, seeking explicit assurances that the system has understood correctly, which is crucial in interactions involving a second language. Conversely, native users with higher self-efficacy favoured the Ask strategy, indicating a preference for direct interaction that allows them to actively correct misunderstandings on their terms. This suggests that high self-efficacy users value efficiency, while language proficiency shapes their preferred strategies, highlighting the need for adaptable designs.

Regarding the influence of demographics, while no significant gender differences were found, a notable negative correlation emerged between age and preference for the Solve repair strategy. Younger participants expressed a reduced preference for option-providing strategies, possibly reflecting generational differences in interacting with technology. In the interviews, two younger participants described such strategies as less preferable and less human-like, likening them to multiple-choice exams, an experience they considered unrealistic and detached from their expectations of DVAs.

Younger users, who are often more accustomed to conversational interfaces enhanced by advanced NLP technologies, may therefore perceive multiple options as cumbersome or outdated. Prior research shows younger users tend to favour immediate, streamlined conversational exchanges, whereas older users may tolerate greater elaboration and slower pacing (e.g., [261]). A recent systematic literature review by Boostani et al. [32] further showed that older adults and children often require more time and attempts for conversational interfaces to interpret their input correctly, and they achieve lower overall success in obtaining relevant answers compared to younger adults. These age-specific disparities suggest that younger users are generally more efficient in navigating conversational interfaces, and thus more likely to view option-heavy strategies as unnecessary or cumbersome. These findings highlight the importance of considering generational perspectives and evolving technological expectations when designing intuitive DVAs.

In addition to language proficiency, several participants highlighted the impact of local accents on the DVA's ability to understand them. Regional pronunciation differences, even within the same language, often led to misunderstandings, causing frustration and diminishing trust in the system [30, 137]. While this study focused on language use, the role of accent variation, particularly in Arabic dialects, emerged as a potentially stronger mediator of repair success than language proficiency alone, since even highly proficient speakers may face misunderstandings due to accent differences [269]. Future research could explore how accent-aware or dialect-sensitive systems might enhance repair effectiveness and inclusivity.

6.5 Conclusion

This study examined how interaction language influences user preferences for five repair strategies in DVAs. Using pairwise comparison, a scenario-based study was conducted to compare these preferences and to explore the impact of individual factors. The results reveal that native and non-native users differ in their repair strategy preferences. While those using a non-native language favoured confirmation-based strategies due to linguistic limitations, native users preferred more direct and explanatory strategies. Additionally, users with high confidence in their ability to interact with a DVA tend to favour repair strategies that are less system-assisted.

Quantitative analysis identified differences in how native and non-native users experience breakdowns and repair strategies. Beyond linguistic factors, context and task type significantly shape repair preferences, underscoring the importance of adaptive repair mechanisms. Importantly, by comparing native and non-native users, this study highlights the need for personalized and context-aware repair strategies in DVAs, considering cultural, linguistic, and individual differences. Future research should explore multimodal repair strategies and real-world adaptive systems to enhance user experience. Addressing these aspects will contribute to the development of more effective and user-centric DVAs.

6.6 Chapter Summary

While the previous chapter (Study 1) examined cross-cultural differences in repair strategy preferences, this chapter focused on a within-culture investigation of how interaction language and individual-level factors influence user preferences. Adopting a mixed-methods approach, the study combined quantitative results with qualitative insights to provide a deeper and exploratory understanding of the patterns as well as the reasons behind user preferences. These findings highlight the nuanced impact of language use and personal characteristics on interaction with SDSs. The next chapter presents a general discussion that integrates findings from all of the empirical studies, offers design recommendations for implementing effective repair strategies in SDSs, and outlines the limitations of the thesis and directions for future research.

Part III

Integrated Discussion and Conclusions

Chapter 7

Integrated Analysis and General Discussion

This chapter brings together the findings of the two empirical studies (Chapters) 5 and 6) to provide an integrated examination of repair strategy preferences across cultural and interaction language conditions. The chapter includes a secondary integrative analysis of the three participant groups (BL1, SL1, and SL2), conducted using existing datasets to examine cross-group patterns that could not be addressed within the individual studies.

The chapter begins with this integrated comparison (Section 7.1). Then, it highlights the most statically significant strategies that showed between-group differences, namely Information and Confirmation (Section 7.2.1), before turning to the remaining repair strategies that showed less stable or non-significant patterns (Section 7.2.2). This chapter then provides practical design recommendations, taking into account cultural communication styles, interaction language, individual-level factors, and system-level features such as repair sequencing and threshold settings (Section 7.3). Finally, the discussion closes with a reflection on the complexity of interaction with SDSs and the interplay of cultural, cognitive, and contextual influences (Section 7.4).

7.1 Integrated Analysis of Three Participant Groups

To address a cross-study research objective that cannot be examined within the individual empirical studies alone, this section presents a secondary integrative analysis combining all three participant groups: British native speakers (BL1), Saudi native speakers (SL1), and Saudi speakers interacting in a non-native language (SL2). While Study 1 (Chapter 5) examined the effect of cultural background and Study 2 (Chapter 6) examined the effect of interaction language in isolation, neither study independently allows for a direct comparison across both dimensions simultaneously.

The integrated analysis is therefore motivated by the need to examine whether patterns of repair strategy preference observed separately across cultural background and interaction language converge, diverge, or interact when these groups are considered together. Leveraging the consistent experimental design across both studies, this analysis uses existing datasets to explore cross-group differences in repair strategy preferences that speak to the combined influence of cultural and linguistic context in SDS interaction.

Study 1 found a statistically significant difference in preferences for the Information strategy (Mann–Whitney Test, $p < .001$), with Saudi speakers (SL1) showing a stronger preference compared to British participants (BL1). Study 2, which controlled for cultural background by focusing solely on Saudi participants, revealed two statistically significant strategies: Information ($p < .001$) and Confirmation ($p = .002$). The former remained more preferred by SL1, while the latter was more favoured by SL2 when interacting in English.

To synthesise findings across the empirical studies ($N = 151$), Figure 7.1 presents the overall ranking of repair strategies collapsed across participant groups. This representation is intended to provide a descriptive overview of general preference tendencies rather than a basis for inferential analysis. All statistical comparisons and hypothesis testing were conducted at the group level and are reported in the respective study chapters.

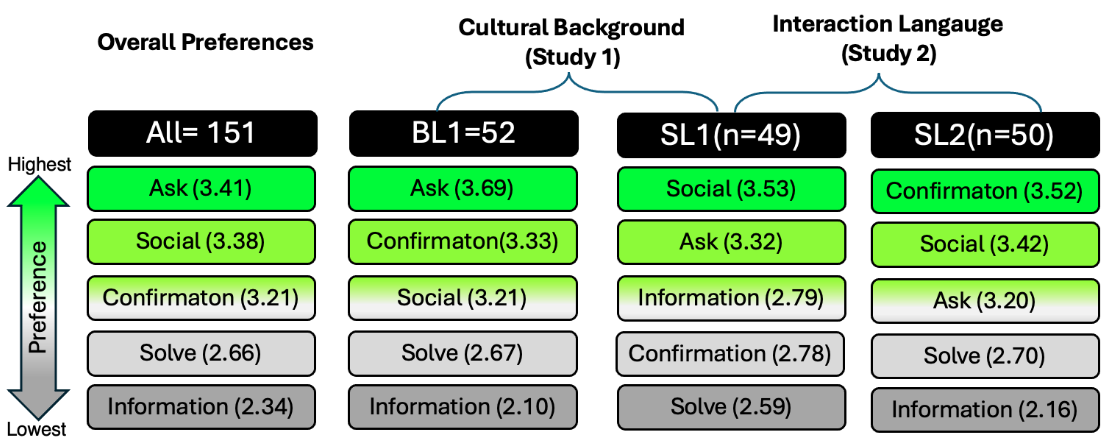


Figure 7.1: **Descriptive overview of overall repair strategy rankings based on Friedman test results across all participants ($N = 151$) and within each participant group. Rankings reflect mean ranking scores, with higher values indicating greater preference. While group-specific results were presented in the respective study chapters, this figure highlights the overall preference trends across cultural and linguistic groups; vertical spacing is illustrative only.**

To further examine cross-group differences, a Kruskal-Wallis test was performed across all three participant groups [59]. Figure 7.2 presents a visual comparison of the mean-ranked preferences for all five strategies across the three groups (BL1, SL1, and SL2). The analysis confirmed that only Information ($H = 14.60$, $p < .001$) and Confirmation ($H = 12.83$, $p = .002$) showed statistically significant differences in preferences. While preferences for Ask, Social, and Solve remained relatively stable across groups, substantial differences emerged for Information and Confirmation, reinforcing their cultural and linguistic sensitivity.

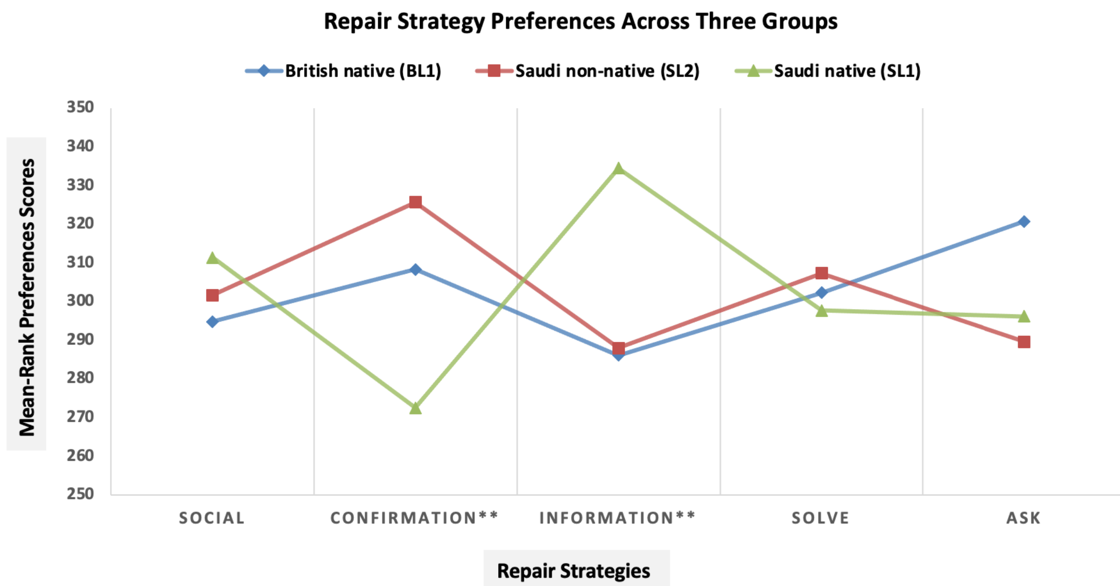


Figure 7.2: **Kruskal–Wallis mean-rank preference scores for the five repair strategies across BL1, SL1, and SL2. Higher scores indicate stronger preference. Significant between-group differences were found for Confirmation ($H = 12.83$, $p = .002$; $SL2 > BL1 > SL1$) and Information ($H = 14.61$, $p < .001$; $SL1 > BL1 \approx SL2$). Social ($H = 1.26$, $p = .533$) and Solve ($H = .415$, $p = .813$) showed no significant differences, while Ask showed a non-significant trend with BL1 highest ($H = 5.08$, $p = .079$).**

The stronger preference for Confirmation among Saudi speakers (SL2) may reflect a need for verification when interacting in a second language. This aligns with findings from second language communication research, which suggest increased cognitive demands in non-native interactions [274, 275]. In contrast, the higher preference for Information among Saudi speakers (SL1) may reflect cultural expectations for more elaborative explanations, consistent with high-context communication norms [103].

Across all three groups, the Kruskal-Wallis test also revealed no statistically significant differences for the Social strategy ($H = 1.26$, $p = .533$), which appeared to be perceived as universally appropriate, showing stable preference patterns. In contrast, the Solve strategy consistently received lower rankings ($H = 0.41$, $p = .813$), suggesting it is less favoured as a repair approach in SDSs, particularly when all provided options are incorrect.

7.2 Theoretical Implications in Cross-Cultural and Linguistic Contexts

To interpret the patterns observed in repair strategy preferences, this section draws on both cultural theories and interactional communication frameworks. Hall's high-low context framework and Hofstede's cultural dimensions (e.g., individualism-collectivism, uncertainty avoidance) offer insights into how cultural norms shape expectations for brevity, elaboration, and clarity in communication. In parallel, cognitive load theory explains how language proficiency and processing demands influence the appeal of concise versus detailed responses, while communication accommodation theory (CAT) provides a lens for understanding how speakers adjust their strategies to align with perceived user needs. These complementary perspectives guide the interpretation of both significant and non-significant between-group differences in the following subsections.

7.2.1 Strategies with Significant Between-Group Differences

The main repair strategies showing statistically significant differences in at least one comparison were Information and Confirmation. Across both empirical studies, Study 1 (cultural background) and Study 2 (interaction language), the Information strategy, which provides a reason for the conversational breakdown, consistently and statistically showed significant differences for Saudi participants (SL1) when compared to other user groups (Mann–Whitney test: $p < .001$ vs. BL1; $p = .001$ vs. SL2). While Confirmation did not differ significantly between British (BL1) and Saudi (SL1) interacting in their native language in Study 1 (Mann–Whitney test, $p = .018$), it was statistically significant in the interaction language comparison in Study 2 (SL1 vs. SL2, $p < .001$). In this thesis, out of the five repair strategies, Information and Confirmation emerged as the two most sensitive to cultural and linguistic effects.

Overall, irrespective of cultural or linguistic background, the Information strategy ranked lowest among all five repair strategies (see Figure 7.1). One explanation, supported by the qualitative findings, is that in goal-oriented tasks (e.g., booking a flight), participants aim to complete the task efficiently, with minimal cognitive load and as few conversational turns as possible, similar to the Ask strategy. Another explanation relates to the sequence of the repair attempt. In the current thesis, the dialogue scenario presented the first conversational breakdown and the first repair attempt. While some participants valued the elaboration and detail offered by the Information strategy, qualitative insights indicate that such responses are less favoured at the first repair attempt but may gain greater value after unsuccessful initial repairs (see P4's comments in Section 6.3.3.2).

This highlights the importance of sequencing repair strategies (i.e., conversation-level management) in SDSs, which should take into account not only sociocultural factors (e.g., cultural background, interaction language), but also task-related costs, the current state of the dialogue, and the cost–benefit trade-offs of repairing breakdowns [122]. Based on Grice's conversational maxims [96], Miehling et al. [169] demonstrate that dialogue system responses should provide a sufficient amount of information (Maxim of Quantity) while being presented in a clear and well-organized manner (Maxim of Manner). Although the Information strategy may satisfy the Maxim of Quantity, its initial deployment in a first-attempt repair could risk violating the Maxim of Manner if the information is perceived as overly detailed or cognitively demanding at that initial stage.

Previous research has investigated the sequence of user repair behaviours when interacting with SDSs (e.g., [152, 192, 179]). In a schedule management setting, Opfermann et al. [192] found that users tend to respond to the first re-prompt with minimal repeats or rephrasing, reserving the addition of problem manifestations or elaborations for the second and third re-prompts. Expecting the same adaptive behaviour from SDSs, the sequencing pattern observed in user-initiated repairs suggests that detailed information may be more effective in later repair stages, when users themselves are more inclined to provide such elaborations. However, dialogue sequencing is a challenge for SDSs, as it needs to adapt dynamically to individual users and "cannot be pre-determined but evolves on a turn-by-turn basis as a result of the interactional work by the participants in the conversation as they aim to achieve their goals and demonstrate mutual understanding" [122, p. 37].

Among the five repair strategies examined, Information and Confirmation emerged as the most sensitive to cultural and linguistic contexts, showing statistically significant differences in at least one of the between-group comparisons. Information consistently distinguished Saudi participants (SL1) from other groups, while Confirmation showed significance primarily in the interaction language comparison with SL2 participants. The following discussion interprets these patterns through two key explanatory themes: cultural communication style and cognitive load in second-language interaction.

7.2.1.1 Cultural Communication Style

Across the three participant groups, the Information strategy was most highly preferred by SL1 (see Figure 7.2, $p < .001$). While not their top-ranked strategy overall, its relatively higher preference aligns with characteristics of high-context communication and high uncertainty avoidance. This is consistent with Hall's [103] conceptualization of high-context cultures, such as Saudi culture, which rely heavily on detailed contextual cues and elaborate explanations to facilitate understanding. Similarly, Hofstede's cultural dimension of uncertainty avoidance [114] reinforces this preference: explicit explanations help reduce ambiguity and provide reassurance. For Saudi participants (SL1), the Information strategy met cultural expectations for richer and more elaborated exchanges, especially when interacting in their native language. By explicitly stating the cause of a breakdown, it reduced uncertainty and offered the clarity valued in high uncertainty avoidance cultures.

In contrast, low-context cultures, such as the UK, tend to favour a more minimalistic approach, prioritising brief and direct requests without extensive background information [103]. British participants (BL1), representing a low-context culture, showed lower preference for the Information strategy, favouring instead concise and direct verbal interactions that emphasize functionality and efficient problem-solving such as Confirmation and Ask strategies. A more detailed discussion of repair strategies in relation to cultural theories is provided in Cultural Interpretation of Strategy Preference Section 5.4.2 of Study 1 (Chapter 5).

While cultural communication style helps explain some of the observed patterns, particularly the higher Saudi (SL1) preference for Information strategy. Notably, the significant increase in Confirmation preference among Saudi (SL2) participants when interacting in a second language suggests that cultural background alone may offer only a partial explanation. This points to cognitive load in second-language interaction as a more primary driver in certain contexts, which becomes especially apparent when examining the SL2 group in Study 2 (Chapter 6).

7.2.1.2 Cognitive Load in L2 Interaction

Among the three participant groups, Saudi interacting in their non-native language (SL2) showed the highest preference for the Confirmation strategy, where the system explicitly confirms its understanding (e.g., "You want to travel to Luton. Is that correct?"). This was followed by British participants (BL1), with Saudi participants (SL1) showing the lowest preference (see Figure 7.2). Statistical analysis revealed that the Confirmation strategy differed significantly only between SL1 and SL2 participants, when investigating interaction language effects in Study 2 (Mann–Whitney U test; $p < .001$). No significant difference was found between SL1 and BL1 participants, when investigating cultural background effects in Study 1 (Mann–Whitney U test, $p = .018$). This suggests that cultural background alone did not account for the observed variation.

Cognitive Load SL2's preferences were influenced by interaction language and associated cognitive demands. Interacting in a second-language increases linguistic processing demands, thereby heightening cognitive load [209, 275, 273]. Insights from interviews revealed that Saudi participants, when using their non-native language (SL2), were concerned that miscommunication might stem from pronunciation and accent issues. As a result, they valued confirmation as a way to receive immediate and clear feedback. Confirmation strategy's simplicity, requiring only a "yes" or "no", minimises the need for complex sentence construction or full repetition, reducing the mental workload in second-language contexts.

Prior research supports this interpretation: non-native users interacting with voice interfaces experience significantly higher mental workload than native users across devices, including smart speakers and smartphones [273]. For example, Pyae and Scifleet [209] found that while native and non-native English speakers were equally confident in listening to and understanding a Google Smart Speaker, non-native speakers felt less confident giving spoken commands. Similarly, Harrington et al. [106] showed that Black older adults interacting with a voice health care assistant were concerned that their accents might cause breakdowns and often resorted to code-switching to be understood. Code-switching refers to the practice of alternating between speech varieties, accents, or linguistic styles depending on context [98]. In a related study, Wenzel and Kaufman [269, p. 8] found that people of colour and non-native English speakers reported harms when interacting with DVAs, as "users must assimilate to a standard white American middle or upper-middle class accent, effectively altering their identity, just to be recognized by their voice assistant."

In HHI research, confirmation checks serve as both a hearing check and an implicit opportunity for correction, enabling the interlocutor to respond with a minimal token rather than repeating the entire utterance [244]. For non-native speakers, the Confirmation repair strategy is considered "a quite typical case of repair due to lack of proficiency in the language" [244, p. 296]. In an HCI context, this may explain why SL2 in the current thesis preferred Confirmation strategies over other strategies, as they impose lower cognitive and linguistic demands. By reducing production effort and clarifying system understanding, confirmation can enhance users' confidence in non-native language interactions.

Cultural Communication Style (and CAT) By contrast, British participants (BL1) ranked the Confirmation strategy second (see Figure 7.1). Low-context cultures, such as the UK, favour clarity and efficiency in dialogue [103], making confirmation a natural fit. SL2 participants' preference for this strategy may also reflect accommodations to the communicative norms of the host language (English). According to Communication Accommodation Theory (CAT), accommodation is a widespread and fundamental process in which speakers adjust their communicative behaviour to manage comprehension and social relationships [83]. Such adjustments can involve aligning lexical choices, speech rate, and interaction style to match the interlocutor's behaviour.

In human–machine contexts, research shows that users also adapt their style when interacting with social robots and voice-based assistants [230, 61]. Craig et al. [61] found that voice-based assistants perceived as over-accommodative, for instance by slowing speech, simplifying vocabulary, or providing translations, were evaluated more positively and seen as higher-quality communicators, particularly when such adjustments were perceived as appropriate to the user's needs. In Study 2, Saudi participants (SL2) interacting in a non-native language (English) may have subconsciously converged toward the direct, minimal confirmation style characteristic of English-language interaction. This shift could function both as a strategy to improve system recognition and as an adaptive response to the communicative norms embedded in the system's English interface.

Information Strategy Contrast Turning to the Information strategy, Saudi participants interacting in their non-native language (SL2) showed a notable decrease in preference compared with SL1 participants. This difference can be interpreted through the lens of cognitive load. Non-native interaction increases linguistic processing demands, making lengthy or elaborate prompts less efficient to comprehend in real time during voice-only interaction. Under such conditions, SL2 participants may prefer more concise repair strategies that reduce processing effort (e.g., the Confirmation strategy), even if these strategies provide less contextual detail. This shift reflects an adaptation to the immediate communicative environment, where linguistic efficiency can outweigh cultural preferences for elaboration.

Interestingly, both Saudi participants interacting in their non-native language (SL2) and British participants (BL1) ranked the Information strategy relatively low (see Figure 7.2). While the reduced preference among SL2 participants may be linked to the increased cognitive load of processing non-native speech, the low preference among BL1 participants appears to stem from different communicative norms [114, 103]. In low-context and low–uncertainty avoidance cultures, such as the UK, conciseness and directness are typically prioritised over explanatory detail. Employing Hofstede's uncertainty avoidance dimension, Haddad et al. [101] compared minimal interfaces (containing only the elements essential for the task) with richer interfaces that included additional support information. They found that Western Caucasian users (low-context culture) of European descent preferred the minimal interface or had no preference, whereas East Asian participants (e.g., Japan and China, high-context cultures) preferred the richer interface and reported feeling less anxious when using it. These results reinforce the interpretation that BL1 participants' lower preference for the Information strategy is rooted in a cultural preference toward efficiency and minimalism, in contrast to elaboration.

Based on a meta-analysis of the effects of cultural values on communication styles in HHI, Merkin et al. [164] found that cultural values have a consistent, but modest, influence on communication patterns across contexts; however, culture explains only part of the variation in communication styles. They caution that "assumptions about one's communication preferences based on one's cultural background must be made with caution and factors other than cultural values must also be taken into account" [164, p. 14]. The increased cognitive demands of second-language interaction represent one such factor that can strongly shape communication preferences. In fact, although cultural values influence expectations and communications, the cognitive load associated with SL2 processing can outweigh these cultural tendencies. This helps explain why Saudi participants, who in their native language context (SL1) showed preferences consistent with high-context cultural expectations, shifted their preferences when interacting in their non-native language as in SL2 group . Consequently, the similar ranking of Information and Confirmation strategies among BL1 and SL2 participants could reflect a convergence in outcomes, but for different reasons. It is the cognitive-load reduction in the case of SL2, and communication style preferences in the case of BL1.

7.2.2 Strategies Without Statistically Significant Between-Group Differences

While the Information and Confirmation strategies demonstrated statistically significant sensitivity to cultural and linguistic context, the remaining three strategies (Ask, Social, and Solve) did not show statistically significant between-group differences across the participant groups. These findings indicate that preferences for these strategies were relatively stable across cultural and interaction-language conditions. Accordingly, the discussion below focuses on describing the observed preference patterns and situating them within the broader interactional and design context of SDSs, without attributing explanatory effects where statistical support is absent. Where relevant, qualitative insights are used to contextualise these patterns rather than to infer causal or group-based differences.

7.2.2.1 Ask

Overall, irrespective of cultural or linguistic background, the Ask strategy ranked highest among all five repair strategies (see Figure 7.1). The Ask strategy is characterised by its directness and conciseness, simply requesting the user to repeat their input without additional elaboration or relational language, making it the shortest and most straightforward repair approach. Insights from the semi-structured interviews suggest that, in goal-oriented tasks, participants often value concise system responses that allow them to resume their task efficiently. In this respect, the Ask strategy aligns well with the functional demands of task-focused interactions and may be particularly appropriate as an initial repair attempt, where a simple repetition request is often sufficient.

Evidence from other goal-oriented tasks supports this preference. In the travel domain, Aberdeen et al. [1] found that when a system failed to understand a user's input, simply asking the user to repeat their prompt was more effective than alternative repair strategies such as Confirmation. Similarly, in the gaming domain, Weitz et al. [266] reported that while rephrasing was the most successful user-initiated strategy, asking for repetition was the most successful system-initiated repair strategy. They also observed that users perceived the AI dialogue partner as intelligent and likeable, although sometimes overestimated its capabilities, which could lead to miscommunication.

Although between-group differences for the Ask strategy were not statistically significant, British participants (BL1) ranked this strategy highest among all groups (see Figure 7.2). However, as these differences did not reach statistical significance, no strong claims can be made regarding cultural or linguistic influences on preferences for the Ask strategy. Across groups, preferences for this strategy appeared relatively stable, suggesting that its appeal may be broadly linked to its simplicity rather than to group-specific factors.

7.2.2.2 Social

The Social strategy, while structurally similar to the Ask strategy, incorporates polite expressions such as apologies (e.g., "I'm sorry") and courteous markers (e.g., "please"), thereby adding a human-like tone and relational warmth while remaining relatively concise. This combination of directness and politeness may help explain why the Social strategy ranked highly overall across participant groups (see Figure 7.1). Prior work supports the general value of politeness in repair contexts. For example, Bulyko et al. [39] showed that apologetic error responses are associated with lower user frustration, while Klein et al. [132] demonstrated that emotionally supportive responses can improve user attitudes and sustain interaction.

Although some variation in ranking was observed across participant groups, between-group differences for the Social strategy did not reach statistical significance. As such, no strong claims can be made regarding systematic cultural or linguistic effects on preferences for this strategy. Overall, preferences for Social repair responses appeared relatively stable across groups, suggesting that their appeal may be linked to their general interpersonal qualities rather than to group-specific factors.

Qualitative findings, however, provide important contextual nuance regarding the use of Social strategies. Interview participants emphasised that apologies and polite markers should be applied selectively rather than indiscriminately. While politeness was generally viewed as beneficial, participants noted that overuse of Social strategies in urgent or high-stakes contexts (e.g., "finding the nearest hospital" [P12]) could hinder efficiency, whereas their inclusion in lower-pressure situations was perceived more positively.

This interpretation aligns with prior work by Akgun et al. [5], who found that the effectiveness of apologetic error messages depends on participants' mood states. Similarly, in website interactions, Tzeng [252] found that negative politeness strategies (e.g., "sorry") were predominantly favoured to handle socially undesirable events, with evaluations differing significantly according to participants' politeness orientations. Recently, Mahmood et al. [158] emphasized the importance of distinguishing between high-stakes and lower-stakes queries in intent recognition failures, recommending a balanced provision of necessary information in DVAs. Taken together, these findings suggest that Social strategies can enhance the interpersonal quality of interactions, but their effectiveness depends on appropriate contextual deployment rather than on cultural or linguistic group membership alone.

7.2.2.3 Solve

Across all participant groups and interaction-language conditions, the Solve strategy consistently received the lowest preference rankings (see Figure 7.1). Importantly, between-group differences for this strategy were not statistically significant, indicating that the low ranking of Solve was stable across cultural and linguistic contexts rather than driven by group-specific effects.

This consistently low preference must be interpreted in light of the experimental design. In this thesis, all options presented within the Solve strategy were intentionally incorrect, reflecting real-world breakdown scenarios and avoiding bias that could arise if a correct option were included (see Section 4.3.3 in Chapter 4). While this design enabled a controlled comparison across repair strategies, it also limits the extent to which preferences for option-based repair can be interpreted independently of this constraint. Participants' low rankings may therefore reflect their response to the design of the strategy rather than to the concept of option-based repair more broadly.

The present findings contrast with prior work in text-based interaction, where option-based repair strategies have been more positively evaluated (e.g., [17]). One important distinction is that studies reporting higher preference for Solve-type strategies typically included the correct option among the alternatives. In addition, interaction modality likely plays a role. In voice-only environments, users must process options sequentially, which may impose higher cognitive demands compared to text-based interfaces where options can be visually scanned and revisited [122]. Unlike text-based settings, where users can quickly scan and select from written options, the present thesis was conducted in a voice-only environment. Participants were required to listen to the full list of options before responding, thereby imposing a higher cognitive load and reducing efficiency. While text-based interfaces allow users to move back and forth at their own pace, voice interaction is inherently linear, requiring participants to process information sequentially [122].

Taken together, the consistently low preference for the Solve strategy across groups should be interpreted cautiously. Rather than indicating a general rejection of option-based repair, the findings highlight how design choices (such as the absence of a correct option and the linear nature of voice interaction) may shape user evaluations. These results underscore the importance of considering task modality and repair design when implementing option-based strategies in SDSs.

7.3 Practical Design Recommendations

This section presents practical design recommendations for repair strategies in SDSs, informed by the findings of this PhD thesis. The recommendations draw on multiple sources of evidence, including (i) quantitative analyses of user preferences across cultural background and interaction language, and (ii) qualitative insights from post-task interviews exploring users' perceptions of repair strategies. In addition, some recommendations reflect broader design-oriented implications that synthesise these findings with prior work on SDS interaction, rather than outcomes that were directly manipulated or empirically tested within the experimental studies (e.g., repair sequencing and system confidence thresholds).

The recommendations are organised around three key dimensions: 1) the suitability of strategies for first-attempt repairs, 2) user background considerations (e.g., cultural, linguistic, and individual-level differences), and 3) conceptual system-confidence threshold considerations. Together, these dimensions provide a structured framework for guiding the design of adaptive and user-sensitive SDS repair mechanisms. Table 7.1 summarises these recommendations, offering an overview of practical design implications derived from this thesis.

Table 7.1: **Recommendations for the Use of Different Repair Strategies in SDSs.**

Strategy Name	Suitability for First-Attempt Repair	User Background Considerations	Confidence Threshold Considerations
Ask	Consistently preferred as a first-attempt repair strategy	No clear differentiation across user background; generally valued for its simplicity and conciseness.	Can be used at any confidence level
Social	Preferred in first-attempt repair, but minimally used	No clear differentiation observed across user background	Can be used at any confidence level, but should be avoided in high-stakes contexts
Confirmation	Frequently preferred in first-attempt repair	Preferred by L2 users for its clarity and simplicity, especially those with high computer self-efficacy	Can be used at any confidence level for L2 users, but required high confidence for L1 users
Information	Less preferred as a first-attempt repair	Preferred by L1 and less experienced users as they can easily comprehend the explanation provided, especially in high-context cultures	Can be used at any confidence level but should be avoided in high-stakes contexts
Solve	Consistently low preference as a first-attempt repair	Preferred by less tech experienced users and those with lower computer self-efficacy, regardless of interaction language	Can be used if at least one of provided options has a high confidence level

7.3.1 Suitability for First-Attempt Repair

The suitability of repair strategies for first-attempt interventions can be inferred from the consistent methodological design of this PhD thesis, in which each task scenario deliberately introduced a single conversational breakdown followed by a first-repair attempt from the system. While the thesis did not explicitly compare first versus subsequent repair turns, this design allows to draw recommendations about which strategies are most effective and preferred in initial repair.

Quantitative findings strongly supported the Ask strategy as the most effective and preferred option for first-attempt repair (see Figure 7.1). Across cultural and linguistic groups, Ask, Social, and Confirmation strategies were consistently ranked higher than more elaborative strategies, such as providing explanations (Information) or lists of options (Solve). This underscores their broad appeal. These strategies were valued for their simplicity and efficiency, as they are required minimal cognitive effort.

Qualitative insights add further depth. Participants emphasized a strong preference for concise repair strategies, particularly in goal-oriented tasks where efficiency and progress were critical. Several interviewees explained that lengthy responses (e.g., explaining the cause of the breakdown) or overly polite formulations were unnecessary during breakdowns. In goal-oriented interactions, users' main priority was to resolve the breakdown and complete the task quickly. This suggests that while politeness and explanations are appreciated in some contexts, they are not prioritized in the initial attempt at repair, especially when task urgency is high. Taken together, these findings suggest that Ask, Social, and Confirmation strategies may be well suited for use as default options in first-attempt repair in SDSs, within the scope of the task scenarios and interaction contexts examined in this thesis.

7.3.2 User Background Consideration

Patterns observed in the quantitative results revealed systematic differences in repair strategy preferences across cultural and linguistic groups, as well as individual-level factors such as computer self-efficacy and prior experience. These patterns provide valuable design insights for tailoring repair strategies in SDSs to diverse user backgrounds.

In terms of interaction language, the Confirmation strategy was significantly favoured by users interacting in their non-native language (L2). Its use of explicit confirmations may provide reassurance and reduce ambiguity, particularly by lowering the cognitive demands associated with repeating full inputs in a non-native language. Cultural communication style also shaped repair strategy preferences. Participants from collectivist cultural contexts (e.g., Saudi Arabia) showed a significantly stronger preference for the Information strategy compared to participants from more individualist cultures (e.g., the UK). This pattern is consistent with theoretical accounts of cultural communication styles, which suggest that high-context and collectivist cultures place greater value on contextual information during communication [114, 103].

Individual-level factors further influenced repair preferences. Information strategy were more acceptable to native speakers (L1), particularly those from high-context cultures who could easily follow and interpret explanatory guidance. Conversely, Solve strategy proved most beneficial for non-native speakers (L2), less experienced technology users, or those with lower computer self-efficacy, as the structured and directive style reduced uncertainty and supported task progression regardless of language background.

7.3.3 System's Confidence Threshold Considerations

In this thesis, recommendations regarding confidence threshold considerations are grounded in the study's methodological design. Specifically, for strategies that provide options or suggestions (i.e., Confirmation and Solve), the correct user intent was deliberately excluded from the presented alternatives. This approach minimized priming bias and better simulated real-world scenarios in which the system operates under low-confidence conditions (see Section 4.3.3 for more details). By adopting this design, the thesis was able to explore how users respond to repair strategies when the system cannot reliably identify users' intended input under low-confidence conditions. Thus, providing design recommendations on the suitability of different strategies in low confidence thresholds offers practical guidance for real-world applications.

The Ask strategy is highly robust across different levels of system confidence, making it a reliable default option, particularly for first-attempt repairs. By placing the burden of repair on the user in a simple and concise way, it can be applied regardless of whether the system is operating with high or low confidence. Similarly, the Social strategy, while valuable for adding relational and human-like elements, has limitations in high-stakes contexts. Based on insights from the qualitative analysis (e.g., P12), its use appears more appropriate in casual or low-stakes interactions, where building rapport can enhance engagement without unnecessarily lengthening the dialogue. In urgent, high-stakes situations where efficiency is critical, the Social strategy should be avoided, even though it can technically be applied at any confidence level.

Confirmation strategies show a more detailed relationship to system confidence thresholds. For users interacting in their second language (L2), they can be effective at both low and high confidence levels, since they reduce ambiguity and cognitive load by requiring only a simple response (e.g., "yes" or "no"). In contrast, for first-language (L1) users, Confirmation strategies appear more appropriate at high confidence thresholds, as they allow the dialogue to proceed quickly with minimal input (e.g., P8). At lower thresholds, however, they may risk creating frustration, since a negative response (e.g., "no") could trigger additional repair turns before the dialogue can progress. In such cases, native speakers may prefer to repeat their inputs (as in the Ask strategy), which is more efficient when no language barrier exists. Although not directly examined in this thesis, future research is needed to examine whether, in scenarios where system confidence is high, Confirmation strategies may offer a cognitively efficient repair option by allowing users to proceed with minimal input (e.g., "yes").

Information strategies, which involve providing explanations rather than suggesting a user intent, can also be used at any confidence level. However, similar to Social strategies, they are not advisable in high-stakes contexts where efficiency is critical. Because they add extra information, especially during initial repairs. Explanations may slow progress and distract from quick task success in such scenarios. Finally, Solve strategies, which provide multiple suggestions or options, are more constrained by confidence considerations. They are best used only when at least one of the provided options has a high confidence score. This ensures that users are not overwhelmed with a list of unreliable alternatives, while still benefiting from structured guidance when confidence is sufficiently strong for at least one option.

7.4 Summary of Findings

This integrative discussion highlights the inherent complexity of HCI, particularly in SDSs, where breakdowns are inevitable. Unlike HHI, system repair strategies need to balance efficiency, clarity, and user expectations, which are shaped not only by cultural communication style but also by the cognitive demands of interaction. These findings underscore that communication with dialogue systems is a dynamic conversation between the user and the system.

The results demonstrate that cultural background does shape preferences for repair strategies in HCI. Participants from high-context, high-uncertainty avoidance cultures (e.g., Saudi users) showed higher preferences for strategies that provide explicit elaboration and contextualization, especially interacting in their native language (e.g., Information). This reflects their communication style in HHI, where detailed explanations and contextual richness are valued. However, while culture plays a significant role in certain contrasts, it is not consistently dominant across all strategies or interactional contexts. Thus, cultural influence should be seen as one important, but not exclusive layer in shaping SDS repair preferences.

Equally important, the thesis's findings reveal that interaction language can outweigh cultural influences. When Saudi participants (SL2) engaged in their second language (English), their preferences shifted, reflecting increased cognitive load and the communication accommodation. Similarly, British participants (BL1) and Saudi participants using English as a non-native language (SL2) aligned in certain repair preferences despite cultural differences. These patterns suggest that situational constraints, such as processing demands and interactional context, can temporarily override cultural communication tendencies.

In goal-oriented SDS tasks, repair strategy preferences are shaped by factors beyond cultural background and interaction language. These include individual-level influences such as prior experience and computer self-efficacy, as well as system-level design features like the sequencing of repair attempts and confidence thresholds. Taken together, these elements often shape user expectations and preferences more strongly than culture alone. For example, users with greater experience in interacting with SDSs tend to prefer Confirmation strategies, as they are more aware of the system's limitations. Similarly, participants with higher self-confidence in their ability to use SDSs are more likely to simply repeat their input when interacting in their native language, but prefer confirmation when using a non-native language. These patterns underscore that repair strategy design should account for the combined influence of culture, language, cognition, and system behaviour, rather than relying on any single factor.

Overall, the theoretical implications emphasize that SDS repair strategy design requires a multi-layered perspective. While cultural communication style provides an important background frame, preferences are ultimately shaped by interaction language, cognitive processing demands, and system characteristics. This layered understanding offers both theoretical clarity and practical direction for developing adaptive and user-sensitive repair strategies in SDSs.

7.5 Chapter Summary

While the previous chapter (Study 2) presented a focused investigation of interaction language and individual-level influences, this chapter brought together findings from all empirical studies to offer an integrated understanding of user preferences for repair strategies in SDSs. It discussed the results in light of the broader theoretical and methodological frameworks introduced earlier, highlighting both cross-cultural and within-culture patterns. In addition, this chapter provided design recommendations for developing cultural and linguistic adaptive SDSs. The final chapter concludes the thesis by presenting the main conclusions, summarizing key contributions, identifying limitations, and outlining directions for future research.

Chapter 8

Conclusions

This final chapter brings the thesis to a close by synthesizing its main quantitative and qualitative findings and reflecting on their broader significance. It begins with a concise conclusion that integrates the key insights from the scoping review and empirical studies (Section 8.1). Building on this, the chapter outlines the thesis's contributions to cross-cultural and linguistic communications, individual-level factors, experimental design, and practical application (Section 8.2), before acknowledging its limitations and scope (Section 8.3). Then, the chapter informs a discussion of promising directions for future research (Section 8.4). Finally, it closes with brief remarks that highlight the broader significance of the work and its implications for advancing understanding and design in HCI (Section 8.5).

8.1 Conclusions

SDSs are becoming increasingly prevalent, enabling users to access information and services across platforms such as computers, smart speakers, and personal assistants. Commercially available systems (e.g., Apple Siri, Amazon Alexa, and Google Assistant) support a wide range of tasks, from booking flights and checking weather forecasts to playing music. Despite these advances, conversational breakdowns remain inevitable, just as they occur in HHI, but are often more problematic when one interlocutor is a machine. Such breakdowns can negatively affect user trust, satisfaction, and continued use of SDSs. In response, research has increasingly focused on developing effective repair strategies (e.g. [278, 163, 180]), which are essential for improving system performance and enhancing user experience.

As SDSs are deployed globally and across multiple languages, they serve a highly diverse user base. Yet communication breakdowns remain common, and relatively little is known about how users from different cultural and linguistic backgrounds prefer such breakdowns to be handled.

This gap highlights the need for a deeper understanding of how repair strategies can be adapted to account for user diversity.

This PhD thesis addresses this gap by quantitatively and qualitatively examining how user characteristics influence preferences and perspectives on communication repair strategies in SDSs. The overarching research question guiding this PhD thesis was: **"How do user characteristics influence user preferences and perspectives on communication repair strategies in spoken dialogue systems during conversational breakdowns?"** To answer this, the research combined a systematic scoping review (Chapter 3) with two empirical studies (Chapters 5 and 6).

The scoping review (Chapter 3) systematically examined the literature on repair strategies in voice-only interaction. To provide consistency across disciplines, it developed two comprehensive frameworks: one categorizing system-repair strategies (Section 3.3.2) and the other categorizing user-repair strategies (Section 3.3.3). The review also identified gaps in existing research, such as the lack of studies on user characteristics, particularly in non-Western contexts (e.g., Saudi Arabia). The system-repair framework derived from this review served as the foundation for the subsequent empirical studies.

The first empirical study (Chapter 5) explored the role of cultural background in shaping user preferences for repair strategies in goal-oriented tasks. By controlling for language proficiency, the study conducted a cross-cultural experiment comparing participants from the UK and Saudi Arabia, each interacting with the system in their native language. The results indicated that cultural background significantly influenced preferences, particularly regarding the degree of elaboration and contextual detail embedded in repair strategies such as the Information strategy. The first study demonstrated that cultural background shaped preferences but also revealed variations that could not be fully explained by culture alone.These findings suggested that factors beyond culture, such as interaction language, cognitive demands, and individual-level factors, may also shape preferences, pointing to the need for further exploration within a single cultural context.

The second empirical study (Chapter 6) addressed this need by adopting a mixed-methods approach to investigate the influence of interaction language and individual-level factors, such as prior experience with SDSs and computer self-efficacy, within a single cultural setting. To minimize cultural bias, the study compared Saudi participants interacting either in their native language (Arabic) or their non-native language (English). Using a methodology consistent with the first study, the quantitative results highlighted cross-linguistic variation, while the qualitative component deepened understanding by capturing participants' perspectives and lived experiences of breakdowns and repair with SDSs.

The findings revealed that although cultural background shapes preferences for repair strategies, the cognitive demands inherent in second-language interaction often outweighed cultural tendencies. Saudi participants (SL2) interacting in English shifted toward less cognitively demanding strategies (e.g. Confirmation) as an adaptation to the immediate communicative environment. Moreover, individual-level factors further shaped preferences. Participants with greater prior experience of SDSs tended to favour Confirmation strategy, reflecting their awareness of system limitations. Those with higher confidence in interacting with SDSs preferred simply repeating their input in native-language interactions (e.g., Ask strategy), but shifted toward the Confirmation strategy when using a second language. Qualitative insights further emphasized the contextual complexity of HCI, showing that the purpose of the interaction (e.g. second-language practising) and situational urgency (e.g. direction to hospital) also influenced user preferences for repair strategies. This pattern aligns with Park et al. [195], who, based on data from 929 individuals across 17 countries, found that individual variations explained 85% of the variance in communication style preferences, compared to only 15% attributable to cultural background.

Taken together, this thesis demonstrates that user preferences for repair strategies in SDSs are shaped not only by cultural background but also by interaction language, prior experience, self-efficacy, and situational context. These findings highlight the importance of designing adaptive repair mechanisms that account for user diversity, rather than relying on one-size-fits-all solutions. They also point to the inherently dynamic nature of HCI, where user behaviour cannot be fully understood without considering both individual and contextual dimensions. By foregrounding these interdependencies, this PhD thesis emphasizes the need for adaptive and context-sensitive design for repair strategies in SDSs. Building on this foundation, the following section outlines the key contributions of this PhD thesis.

8.2 Summary of Contributions

This thesis makes several key contributions to advancing the understanding of communication repair in SDSs. First, it develops two comprehensive frameworks for categorizing system- and user-initiated repair strategies: an adapted framework for system-initiated repair strategies based on prior text-based interaction literature (e.g. [26]), and a novel user-initiated repair framework developed from the scoping review. Second, through two empirical studies, it demonstrates how cultural background, interaction language, and individual-level factors, such as self-efficacy and prior experience, influence user preferences for repair strategies, thereby highlighting the importance of user diversity in system design. Third, it provides rich qualitative insights into users' lived experiences and expectations, complementing the quantitative findings and underscoring the contextual complexity of HCI. Methodologically, the thesis extends the use of pairwise comparison in SDS research, showing its effectiveness for capturing complex user preferences across cultural and linguistic contexts. Finally, it translates these findings into practical recommendations for designing adaptive, user-sensitive repair mechanisms in goal-oriented SDSs.

8.3 Limitations of the Thesis

While this thesis provides significant insights, several limitations should be acknowledged.

A first limitation relates to the scoping review. The analysis relied on predefined frameworks for categorising repair strategies in CAs, which may not fully capture the multifaceted and evolving nature of repair strategies in SDSs. Although these frameworks provided a structured understanding, the process was constrained by the author's interpretations and by inconsistencies in terminology across studies. Efforts were made to categorise strategies based on definitions, examples, and the objectives described in the studies reviewed. The inclusion of combined repair strategies further complicated categorisation, as it blurred the boundaries between strategy types. Moreover, the review's scope was restricted to four selected databases, predefined search terms, and inclusion criteria, which may have excluded relevant studies, particularly from non-Western contexts. Future research could employ more theory-building approaches or integrate recent taxonomies (e.g., [10]) to produce more comprehensive frameworks.

A second limitation concerns the methodological design of the empirical studies. Both experiments used a controlled setup in which the SDS always failed to recognize the correct option and presented incorrect alternatives. While this design allowed for consistency and comparability, it does not fully reflect the unpredictability of real-world interactions, where both correct and incorrect system responses occur. Furthermore, no commercial SDSs (e.g., Alexa, Siri) were used; instead, responses were pre-scripted and delivered using Amazon Polly's text-to-speech engine. Although this ensured standardisation across participants, it may limit the generalisability of the findings to real-time interactions with commercial systems.

A further limitation of the experimental design concerns the inclusion of deliberately incorrect content within the Confirmation and Solve repair strategies. This design choice was intended to simulate realistic breakdown scenarios and to avoid biasing preferences by presenting a correct suggestion or correct option. However, it may have influenced the observed preference patterns in two ways. First, incorrect suggestions or options may have reduced the perceived usefulness and trustworthiness of the Confirmation and Solve strategies. Second, because preferences were elicited via pairwise comparisons, this reduced appeal may have indirectly increased the likelihood of selection for the alternative strategies with which they were paired. Accordingly, preference scores in this thesis should be interpreted as relative judgments under the specific experimental conditions and strategy implementations used. Future work could evaluate variants in which Confirmation includes a correct suggestion, or Solve includes a correct option to better disentangle the effects of strategy type from the accuracy of suggested content.

Third, the empirical studies focused exclusively on a goal-oriented task (flight booking). Qualitative findings suggested that contextual factors, such as task urgency or whether users were practising a second language, also influenced repair preferences. By excluding open-domain and social tasks, the research may have overlooked how different interaction purposes shape expectations of repair.Future work should therefore extend this investigation to a broader range of tasks and domains.

More broadly, it remains unclear whether the preference patterns observed in this thesis would generalise to social or open-domain conversational contexts. In goal-oriented tasks, such as flight booking, users typically prioritise efficiency, clarity, and rapid task completion, which may explain the strong preference for concise repair strategies (e.g., Ask and Confirmation). In contrast, in social or relational interactions, users may place greater value on politeness, rapport-building, or conversational naturalness, potentially increasing the relative appeal of Social or Information-based repair strategies. In social dialogues, the primary aim is to sustain engagement, and repair is often less focused on resolving misunderstandings and more on continuing the conversation [17]. As a result, the findings of this thesis should be interpreted as context-dependent, reflecting preferences within a goal-oriented interaction setting rather than universal repair strategy preferences across all conversational domains.

With regard to Study 1, the cultural effect was examined only through a comparison of two cultures, Saudi Arabia (high-context) and the UK (low-context), within a single goal-oriented task. While this revealed meaningful cultural differences, the narrow scope limits the generalisability of the findings. Cultural influence on communication is complex and context-dependent; as highlighted in prior HCI research [167], communication patterns are shaped not only by culture but also by dialogue domain, task type, and individual user states. Broader cross-cultural comparisons and more varied contexts are therefore needed to better capture the detailed role of culture.

Finally, Study 2 focused on individual-level factors such as prior experience, computer self-efficacy, and language proficiency, but did not address other potentially influential traits. Research on HHI has shown that individual differences can outweigh cultural effects in shaping communication preferences [195]. Future work should therefore incorporate a wider range of cognitive and psychological characteristics, such as tolerance for ambiguity, cognitive load, and personality traits, to deepen understanding of individual variation in repair strategy preferences.

8.4 Future Research Directions

Building on the findings and limitations of this thesis, several promising directions emerge for future research.

First, the proposed frameworks could be further refined and extended by integrating recent taxonomies and applying them in different contexts. While the scoping review outlined broad categories of repair, user behaviour varies widely and requires more detailed mapping. For example recently, Alloatti et al. [10] investigated user adaptations during conversational breakdowns with CAs and introduced a novel tag system to classify user repair attempts in text-based interaction.

Adapting such taxonomies to voice-only interaction could provide a more comprehensive account of how users attempt repair in SDS contexts.

Second, the scoping review also revealed that most research on system-repair strategies has focused on Confirmation, Ask, and Social strategies, with relatively little attention to Information and Disclosure. Although this thesis addressed Information strategies empirically, Disclosure strategies remain largely unexplored. Future studies could investigate how users perceive and respond to these less-studied strategies and the contexts in which they may be effective.

Third, future studies should employ interactive systems, including commercial SDSs, to capture the dynamic and unpredictable nature of real-world interaction. While this thesis benefited from the consistency of a controlled experimental design, real-world conversations are far more variable. Systems may sometimes offer correct responses, repairs may succeed or fail, and users may escalate to alternative strategies or abandon the interaction altogether. Research on repair sequences and outcomes in SDSs could therefore yield valuable insights. For instance, Ashktorab et al. [17] found that simple strategies like Ask ranked highly in successful repair scenarios, whereas users preferred escalation to a human agent after repeated failures. Investigating similar patterns in SDS contexts would help clarify how repair outcomes shape user behaviour and trust.

Additionally, this thesis primarily interpreted findings through established cultural and communication frameworks, such as Hall's high- and low-context theory [103] and Hofstede's cultural dimensions [114], and also incorporated insights from CAT and cognitive load perspectives. While these frameworks provided a valuable foundation, they represent only one type of interpretive lens. Future research could benefit from drawing on alternative perspectives, such as politeness theory or interactional sociolinguistics, which may reveal additional dimensions of how users negotiate meaning during breakdowns. Adopting multiple theoretical lenses would enable a more comprehensive understanding of how cultural, linguistic, and individual factors interact in shaping repair preferences.

Finally, expanding the investigation of individual-level factors remains vital. Psychological and cognitive characteristics,s such as tolerance for ambiguity, personality traits, generational differences, professional background, and educational experience, may shape communication preferences in ways that diverge from dominant cultural norms. As Dumitrescu [71, p. 90] observes, the global interconnectedness facilitated by modern communication technologies "tends to unify communicative practices and styles and to bring people... together as one more or less uniform, beyond and despite cultural boundaries." The findings of this thesis echo this observation, showing that individual and contextual factors play a critical role in HCI. Future research should therefore examine these characteristics across varied contexts and explore adaptive SDSs that personalise repair strategies to user profiles and evaluate them longitudinally.

8.5 Closing Remarks

This thesis has shown that preferences for repair strategies in SDSs are shaped by the interplay of cultural, linguistic, and individual factors, reflecting the inherent complexity of HCI. Beyond its empirical findings, this work highlights the value of an interdisciplinary perspective by drawing on cultural theories, communication studies, and HCI. It provides a more holistic view of repair than any single lens could offer. As Edward Hall [102, p. 218] observed, “culture is communication and communication is culture”, a reminder that these dimensions are inseparably linked. Recognising this complexity underscores both the richness of studying human–machine communication and the importance of designing adaptive and context-sensitive systems for diverse global users.

Bibliography

[1] Aberdeen, J., and Ferro, L. Dialogue patterns and misunderstandings. In *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Error Handling in Spoken Dialogue Systems* (Château d'Oex, Vaud, Switzerland, Aug. 2003), ISCA Archive, pp. 17–21.

[2] Abfalter, D., Mueller-Seeger, J., and Raich, M. Translation decisions in qualitative research: a systematic framework. *International Journal of Social Research Methodology 24*, 4 (2021), 469–486.

[3] Agresti, A. *Categorical data analysis*, vol. 792. John Wiley & Sons, 2012.

[4] Ahmed, N., Liu, X., Al-Hazwani, I., and Huang, E. M. Cultural dimensions and mental health technology: A systematic review of hofstede's dimensions in shaping mental health experiences. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–13.

[5] Akgun, M., Cagiltay, K., and Zeyrek, D. The effect of apologetic error messages and mood states on computer users' self-appraisal of performance. *Journal of Pragmatics 42*, 9 (2010), 2430–2448.

[6] Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, february 21-26, 2011, revised selected papers*. Springer, 2012, pp. 114–130.

[7] Al-Twairesh, N., Al-Matham, R., Madi, N., Almugren, N., Al-Aljmi, A.-H., Alshalan, S., Alshalan, R., Alrumayyan, N., Al-Manea, S., Bawazeer, S., et al. Suar: Towards building a corpus for the saudi dialect. *Procedia computer science 142* (2018), 72–82.

[8] Alghamdi, E., Halvey, M., and Nicol, E. System and user strategies to repair conversational breakdowns of spoken dialogue systems: a scoping review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (2024), pp. 1–13.

[9] Alghamdi, E., Halvey, M., and Nicol, E. The effect of interaction language on preferences for communication repair strategies in digital voice assistants (dvas): a comparative study. In *Proceedings of the Mensch und Computer 2025*. 2025, pp. 146–162.

[10] Alloatti, F., Grasso, F., Ferrod, R., Siragusa, G., Di Caro, L., and Cena, F. A tag-based methodology for the detection of user repair strategies in task-oriented conversational agents. *Computer Speech & Language* (2024), 101603.

[11] Andrews, M. Against good advice: Reflections on conducting research in a country where you don't speak the language. *The Oral History Review 22*, 2 (1995), 75–86.

[12] Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., and Sakr, M. Effects of culture on the credibility of robot speech: A comparison between english and arabic. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (2015), pp. 157–164.

[13] Anna-Maria Meck, C. D., and Vogt, T. Failing with grace: Exploring the role of repair costs in conversational breakdowns with in-car voice assistants. *International Journal of HumanÄìComputer Interaction 0*, 0 (2023), 1–19.

[14] Arksey, H., and O'Malley, L. Scoping studies: towards a methodological framework. *International journal of social research methodology 8*, 1 (2005), 19–32.

[15] Arnold, A., Kolody, S., Comeau, A., and Miguel Cruz, A. What does the literature say about the use of personal voice assistants in older adults? a scoping review. *Disability and Rehabilitation: Assistive Technology 19*, 1 (2024), 100–111.

[16] Ashktorab, Z., Desmond, M., Johnson, J. M., Pan, Q., Dugan, C., Brachman, M., and Spina, C. Sme-in-the-loop: Interaction preferences when supervising bots in human-ai communities. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (2023), pp. 2281–2303.

[17] Ashktorab, Z., Jain, M., Liao, Q. V., and Weisz, J. D. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–12.

[18] Asisof, A. Towards a comprehensive repair framework for human-chatbot interaction: the case of rephrasing. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (2022), pp. 1–3.

[19] Avgustis, I., Shirokov, A., and Iivari, N. "please connect me to a specialist": scrutinising 'recipient design'in interaction with an artificial conversational agent. In *IFIP Conference on Human-Computer Interaction* (2021), Springer, pp. 155–176.

[20] Aw, E. C.-X., Tan, G. W.-H., Cham, T.-H., Raman, R., and Ooi, K.-B. Alexa, what's on my shopping list? transforming customer experience with digital voice assistants. *Technological Forecasting and Social Change 180* (2022), 121711.

[21] Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review 84* (1977).

[22] Bandura, A. *Self-efficacy in changing societies*. Cambridge university press, 1997.

[23] Bandura, A., et al. Social foundations of thought and action. *Englewood Cliffs, NJ 1986*, 23-28 (1986), 2.

[24] Baumeister, R. F. *Encyclopedia of social psychology*, vol. 1. Sage, 2007.

[25] Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., and Hiniker, A. Communication breakdowns between families and alexa. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–13.

[26] Benner, D., Elshan, E., Schöbel, S., and Janson, A. What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents. In *International Conference on Information Systems (ICIS)* (2021), pp. 1–17.

[27] Bhattacharjee, A., Suh, J., Ershadi, M., Iqbal, S. T., Wilson, A. D., and Hernandez, J. Understanding communication preferences of information workers in engagement with text-based conversational agents. *arXiv preprint arXiv:2410.20468* (2024).

[28] Bickmore, T., and Cassell, J. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2001), pp. 396–403.

[29] Bohus, D. *Error awareness and recovery in conversational spoken language interfaces*. PhD thesis, Carnegie Mellon University, 2007.

[30] Bohus, D., and Rudnicky, A. I. Sorry, i didn't catch that! an investigation of non-understanding errors and recovery strategies. *Recent trends in discourse and dialogue* (2008), 123–154.

[31] Bonnefon, J.-F., Feeney, A., and De Neys, W. The risk of polite misunderstandings. *Current Directions in Psychological Science 20*, 5 (2011), 321–324.

[32] Boostani, R., Munteanu, C., and Kuzminykh, A. Towards age-inclusive conversational interfaces: Understanding requirements across age groups. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces* (2025), pp. 1–13.

[33] Borlund, P. A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. *Journal of Documentation 72*, 3 (2016), 394–413.

[34] Borlund, P., and Schneider, J. W. Reconsideration of the simulated work task situation: A context instrument for evaluation of information retrieval interaction. In *Proceedings of the third symposium on Information interaction in context* (2010), pp. 155–164.

[35] Bourguet, M.-L. Towards a taxonomy of error-handling strategies in recognition-based multi-modal human–computer interfaces. *Signal Processing 86*, 12 (2006), 3625–3643.

[36] Braggaar, A., Verhagen, J., Martijn, G., and Liebrecht, C. Conversational repair strategies to cope with errors and breakdowns in customer service chatbot conversations. In *International Workshop on Chatbot Research and Design* (2023), Springer, pp. 23–41.

[37] Braun, V., and Clarke, V. Using thematic analysis in psychology. *Qualitative research in psychology 3*, 2 (2006), 77–101.

[38] Brown, P. *Politeness: Some universals in language usage*, vol. 4. Cambridge university press, 1987.

[39] Bulyko, I., Kirchhoff, K., Ostendorf, M., and Goldberg, J. Error-correction detection and response generation in a spoken dialogue system. *Speech Communication 45*, 3 (2005), 271–288.

[40] Cackett, M., Fulton, L., and Pearl, D. Sound as an interface, methods to evaluate voice user interface (vui) experiences in various contexts. In *Proceedings of the Future Technologies Conference* (2022), Springer, pp. 120–129.

[41] Cairns, P. 34. experimental methods in human-computer interaction.

[42] Cao, Y., Carmona, V. I. S., Liu, X., Hu, C., Iskender, N., Beyer, A., Möller, S., and Polzehl, T. On the impact of self-efficacy on assessment of user experience in customer service chatbot conversations. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore* (2022), Springer, pp. 253–262.

[43] Chen, C., Lifset, E. T., Han, Y., Roy, A., Hogarth, M., Moore, A. A., Farcas, E., and Weibel, N. Screen or no screen? lessons learnt from a real-world deployment study of using voice assistants with and without touchscreen for older adults. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (2023), pp. 1–21.

[44] Chen, H., Liu, X., Yin, D., and Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter 19*, 2 (2017), 25–35.

[45] Chen, P.-J., Hsu, I.-H., Huang, Y.-Y., and Lee, H.-Y. Mitigating the impact of speech recognition errors on chatbot using sequence-to-sequence model. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2017), IEEE, pp. 497–503.

[46] Cheng, Y., Yen, K., Chen, Y., Chen, S., and Hiniker, A. Why doesn't it work? voice-driven interfaces and young children's communication repair strategies. In *Proceedings of the 17th ACM conference on interaction design and children* (2018), pp. 337–348.

[47] Choisel, S., and Wickelmaier, F. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America 121*, 1 (2007), 388–400.

[48] Choularton, S., and Dale, R. User responses to speech recognition errors: consistency of behaviour across domains. In *Proceedings of SST* (2004), vol. 4.

[49] Clark, H. Grounding in communication. *Perspectives on socially shared cognition/American Psychological Association* (1991).

[50] Clark, H. H. *Using language*. Cambridge university press, 1996.

[51] Clark, H. H., and Schaefer, E. F. Collaborating on contributions to conversations. *Language and cognitive processes 2*, 1 (1987), 19–41.

[52] Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., et al. The state of speech in hci: Trends, themes and challenges. *Interacting with computers 31*, 4 (2019), 349–371.

[53] Clausen, M., Kyhn, M. P., Papachristos, E., and Merritt, T. Exploring humor as a repair strategy during communication breakdowns with voice assistants. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (2023), pp. 1–9.

[54] Cohen, J. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[55] Cohen, M. H., Giangola, J. P., and Balogh, J. *Voice user interface design*. Addison-Wesley Professional, 2004.

[56] Colman, M., and Healey, P. The distribution of repair in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2011), vol. 33.

[57] Compeau, D. R., and Higgins, C. A. Computer self-efficacy: Development of a measure and initial test. *MIS quarterly* (1995), 189–211.

[58] Cook, R., et al. Residuals and influence in regression.

[59] Corder, G. W., and Foreman, D. I. Nonparametric statistics for non-statisticians. *(No Title)* (2009).

[60] Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., and Bandeira, N. " what can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services* (2017), pp. 1–12.

[61] Craig, M. J., Lin, X., Edwards, C., and Edwards, A. Can you repeat that again? investigating the mediating effects of perceived accommodation appropriateness for accommodative voice-based assistants. *Computers in Human Behavior: Artificial Humans 2*, 2 (2024), 100102.

[62] Cramer, D., and Howitt, D. L. The sage dictionary of statistics: a practical resource for students in the social sciences.

[63] Crescenzi, A., and Li, L. Assessing realism in simulated work tasks. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (2022), pp. 266–271.

[64] Cuadra, A., Li, S., Lee, H., Cho, J., and Ju, W. My bad! repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (2021), 1–24.

[65] Cui, Z., Gong, H., Wang, Y., Shen, C., Zou, W., and Luo, S. Enhancing interactions for in-car voice user interface with gestural input on the steering wheel. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2021), pp. 59–68.

[66] Daudt, H. M., van Mossel, C., and Scott, S. J. Enhancing the scoping study methodology: a large, inter-professional team's experience with arksey and o'malley's framework. *BMC medical research methodology 13* (2013), 1–9.

[67] David, H. A. *The method of paired comparisons*, vol. 12. London, 1963.

[68] Derinalp, P., and Ozyurt, M. Adaptation of the student attitudes toward artificial intelligence scale to the turkish context: Validity and reliability study. *International Journal of Human–Computer Interaction 41*, 8 (2025), 4653–4667.

[69] Di Maro, M. Computational grounding: An overview of common ground applications in conversational agents. *IJCoL. Italian Journal of Computational Linguistics 7*, 7-1, 2 (2021), 133–156.

[70] Dippold, D. "can i have the scan on tuesday?" user repair in interaction with a task-oriented chatbot and the question of communication skills for ai. *Journal of pragmatics 204* (2023), 21–32.

[71] Dumitrescu, V. M. Culture as communication: Communication style across and within cultures. *Synergy*, 1 (2013), 84–91.

[72] Engelhardt, S., Hansson, E., and Leite, I. Better faulty than sorry: Investigating social recovery strategies to minimize the impact of failure in human-robot interaction. In *WCIHAI@ IVA* (2017), pp. 19–27.

[73] Esterwood, C., and Robert Jr, L. P. Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human behavior 142* (2023), 107658.

[74] Fan, M., and Zhu, L. Think-aloud verbalizations for identifying user experience problems: Effects of language proficiency with chinese non-native english speakers. In *Proceedings of the Ninth International Symposium of Chinese CHI* (2021), pp. 22–32.

[75] Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods 39*, 2 (2007), 175–191.

[76] Feine, J., Gnewuch, U., Morana, S., and Maedche, A. A taxonomy of social cues for conversational agents. *International Journal of human-computer studies 132* (2019), 138–161.

[77] Feng, S. How to convey resilience: Towards a taxonomy for conversational agent breakdown recovery strategies.

[78] Field, A. *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

[79] Følstad, A., and Taylor, C. Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In *International Workshop on Chatbot Research and Design* (2019), Springer, pp. 201–214.

[80] Foosherian, M., Kernan Freire, S., Niforatos, E., Hribernik, K. A., and Thoben, K.-D. Break, repair, learn, break less: Investigating user preferences for assignment of divergent phrasing learning burden in human-agent interaction to minimize conversational breakdowns. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia* (2022), pp. 151–158.

[81] Frampton, M., and Lemon, O. Using dialogue acts to learn better repair strategies for spoken dialogue systems. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (2008), IEEE, pp. 5045–5048.

[82] Galbraith, M. An analysis of dialogue repair in voice assistants. *arXiv preprint arXiv:2311.03952* (2023).

[83] Gallois, C., Ogay, T., and Giles, H. Communication accommodation theory. *Theorizing about intercultural communication* (2005), 121–148.

[84] Ge, X., Li, D., Guan, D., Xu, S., Sun, Y., and Zhou, M. Do smart speakers respond to their errors properly? a study on human-computer dialogue strategy. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21* (2019), Springer, pp. 440–455.

[85] Gessinger, I., Cohn, M., Zellou, G., and Möbius, B. Cross-cultural comparison of gradient emotion perception: Human vs. alexa tts voices. In *INTERSPEECH* (2022), pp. 4970–4974.

[86] Gibson, R. Intercultural business communication. *TESL-EJ 9*, 1 (2005).

[87] Gieselmann, P. Comparing error-handling strategies in human-human and human-robot dialogues. In *Proc. 8th Conf. Nat. Language Process.(KONVENS). Konstanz, Germany* (2006), pp. 24–31.

[88] Giri, V. N. Culture and communication style. *The Review of Communication 6*, 1-2 (2006), 124–130.

[89] Gist, M. E., Schwoerer, C., and Rosen, B. Effects of alternative training methods on self-efficacy and performance in computer software training. *Journal of applied psychology 74*, 6 (1989), 884.

[90] Goldberg, J., Ostendorf, M., and Kirchhoff, K. The impact of response wording in error correction subdialogs. In *ISCA tutorial and research workshop on error handling in spoken dialogue systems* (2003).

[91] Green, A., Eklundh, K. S., Wrede, B., and Li, S. Integrating miscommunication analysis in natural language interface design for a service robot. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2006), IEEE, pp. 4678–4683.

[92] Greenberg, S., and Buxton, B. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2008), pp. 111–120.

[93] Greene, J. C., Caracelli, V. J., and Graham, W. F. Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis 11*, 3 (1989), 255–274.

[94] Gregoire, J. Itc guidelines for translating and adapting tests. *International Journal of testing 18*, 2 (2018), 101–134.

[95] Greiner, D., and Lemoine, J.-F. Bridging the gap: user expectations for conversational ai services with consideration of user expertise. *Journal of Services Marketing 39*, 2 (2025), 76–94.

[96] Grice, H. P. Logic and conversation. In *Speech acts*. Brill, 1975, pp. 41–58.

[97] Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K., and Heyman, S. The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human communication research 22*, 4 (1996), 510–543.

[98] Gumperz, J. Discourse strategies. *Cambridge UP* (1982).

[99] Gunarto, H. Applications of ai-empowered electric vehicles for voice recognition in asian and austronesian languages. In *Artificial Intelligence-Empowered Modern Electric Vehicles in Smart Grid Systems*. Elsevier, 2024, pp. 81–112.

[100] Gupta, S., and Bostrom, R. P. A revision of computer self-efficacy conceptualizations in information systems. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems 50*, 2 (2019), 71–93.

[101] Haddad, S., McGrenere, J., and Jacova, C. Interface design for older adults with varying cultural attitudes toward uncertainty. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), pp. 1913–1922.

[102] Hall, E. T. *The Silent Language*. Doubleday, Garden City, NY, 1959.

[103] Hall, E. T. *Beyond culture*. Anchor, 1976.

[104] Hall, E. T. Context and meaning. *Intercultural communication: A reader 9* (2000), 34–43.

[105] Han, E., Yin, D., and Zhang, H. Interruptions during a service encounter: Dealing with imperfect chatbots.

[106] Harrington, C. N., Garg, R., Woodward, A., and Williams, D. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–15.

[107] Heigham, J., and Croker, R. *Qualitative research in applied linguistics: A practical introduction*. Springer, 2009.

[108] Hertzog, M. A. Considerations in determining sample size for pilot studies. *Research in nursing & health 31*, 2 (2008), 180–191.

[109] Hirschberg, J., Litman, D., and Swerts, M. Identifying user corrections automatically in spoken dialogue systems. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics* (2001).

[110] Hirschberg, J., Swerts, M., and Litman, D. Labeling corrections and aware sites in spoken dialogue systems. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue* (2001).

[111] Hirst, G., McRoy, S., Heeman, P., Edmonds, P., and Horton, D. Repairing conversational misunderstandings and non-understandings. *Speech communication 15*, 3-4 (1994), 213–229.

[112] Hofstede, G. *Culture's consequences: International differences in work-related values*, vol. 5. sage, 1984.

[113] Hofstede, G. Cultural differences in teaching and learning. *International Journal of intercultural relations 10*, 3 (1986), 301–320.

[114] Hofstede, G. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

[115] Hofstede, G. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture 2*, 1 (2011), 8.

[116] Hosseinpanah, A., and Krämer, N. C. Lost in interpretation? the role of culture on rating the emotional nonverbal behaviors of a virtual agent. In *International Conference on Human-Computer Interaction* (2021), Springer, pp. 350–368.

[117] Huang, B., and Sénécal, S. How should voice assistants be heard? the mitigating effect of verbal and vocal warmth in voice assistant service failure. *The Service Industries Journal* (2023), 1–21.

[118] Jain, M., Kota, R., Kumar, P., and Patel, S. N. Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 chi conference on human factors in computing systems* (2018), pp. 1–6.

[119] Janssen, A., Passlick, J., Rodríguez Cardona, D., and Breitner, M. H. Virtual assistance in any context: A taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering 62* (2020), 211–225.

[120] Jee, N., Kumar, S., Patel, R. R., Mandal, R., Singh, R. K., and Vardhan, H. Advancements in voice assistants: A study of speech recognition and emotional intelligence. In *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)* (2024), IEEE, pp. 283–286.

[121] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys 55*, 12 (2023), 1–38.

[122] Jokinen, K., and McTear, M. *Spoken dialogue systems.* Morgan & Claypool Publishers, 2009.

[123] Karolus, J., Wozniak, P. W., Chuang, L. L., and Schmidt, A. Robust gaze features for enabling language proficiency awareness. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (2017), pp. 2998–3010.

[124] Kazemzadeh, A., Lee, S., and Narayanan, S. Acoustic correlates of user response to error in human-computer dialogues. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)* (2003), IEEE, pp. 215–220.

[125] Khan, T., Williams, M. A., and Pitts, M. J. Cross-cultural differences in automotive hmi design: a comparative study between uk and indian users' design preferences. *Journal of Usability Studies 11*, 2 (2016), 45–65.

[126] Kiesel, J., Bahrami, A., Stein, B., Anand, A., and Hagen, M. Toward voice query clarification. In *The 41st international acm sigir conference on research & development in information retrieval* (2018), pp. 1257–1260.

[127] Kiesel, J., Bahrami, A., Stein, B., Anand, A., and Hagen, M. Clarifying false memories in voice-based search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (2019), pp. 331–335.

[128] Kim, J., Jeong, M., and Lee, S. C. " why did this voice agent not understand me?" error recovery strategy for in-vehicle voice user interface. In *Proceedings of the 11th International*

Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings (2019), pp. 146–150.

[129] Kim, S., Salter, D., DeLuccia, L., and Tamrakar, A. Study on text-based and voice-based dialogue interfaces for human-computer interactions in a blocks world. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (2020), pp. 227–229.

[130] Kim, Y. Y. Inquiry in intercultural and development communication. *Journal of communication* (2005).

[131] Kisser, L., and Siegert, I. Erroneous reactions of voice assistants" in the wild"–first analyses. *Elektronische Sprachsignalverarbeitung* (2022), 113–120.

[132] Klein, J., Moon, Y., and Picard, R. W. This computer responds to user frustration. In *CHI'99 extended abstracts on Human factors in computing systems* (1999), pp. 242–243.

[133] Kox, E. S., Kerstholt, J. H., Hueting, T. F., and de Vries, P. W. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous agents and multi-agent systems 35*, 2 (2021), 30.

[134] Kraljevski, I., and Hirschfeld, D. Hyperarticulation of corrections in multilingual dialogue systems. In *INTERSPEECH* (2017), pp. 2531–2535.

[135] Krämer, N. C., von der Pütten, A., and Eimler, S. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. *Human-computer interaction: The agency perspective* (2012), 215–240.

[136] Krauss, R. M., and Chiu, C.-Y. Language and social behavior.

[137] Kukk, K., and Alumäe, T. Improving language identification of accented speech. *arXiv preprint arXiv:2203.16972* (2022).

[138] Kulviwat, S., C. Bruner II, G., and P. Neelankavil, J. Self-efficacy as an antecedent of cognition and affect in technology acceptance. *Journal of Consumer Marketing 31*, 3 (2014), 190–199.

[139] Kvale, S. Interviews: Learning the craft of qualitative research interviewing, 2009.

[140] Kyriakoullis, L., and Zaphiris, P. Culture and hci: a review of recent cultural studies in hci and social networks. *Universal Access in the Information Society 15* (2016), 629–642.

[141] Law, E. L.-C., van As, N., and Følstad, A. Effects of prior experience, gender, and age on trust in a banking chatbot with (out) breakdown and repair. In *IFIP Conference on Human-Computer Interaction* (2023), Springer, pp. 277–296.

[142] Lee, C., Jung, S., Lee, D., and Lee, G. G. Example-based error recovery strategy for spoken dialog system. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* (2007), IEEE, pp. 538–543.

[143] Lee, D., Sah, Y. J., and Lee, S. Improving usability perception of error-prone ai speakers: Elaborated feedback mitigates negative consequences of errors. *International Journal of Human–Computer Interaction 35*, 17 (2019), 1645–1652.

[144] Lee, M., and Lee, S. "i don't know exactly but i know a little": Exploring better responses of conversational agents with insufficient information. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–5.

[145] Li, C.-H., Yeh, S.-F., Chang, T.-J., Tsai, M.-H., Chen, K., and Chang, Y.-J. A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–12.

[146] Li, T. J.-J., Radensky, M., Jia, J., Singarajah, K., Mitchell, T. M., and Myers, B. A. Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology* (2019), pp. 577–589.

[147] Liao, Q. V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., Wang, D., Muller, M., Shami, N. S., and Geyer, W. All work and no play? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–13.

[148] Liljenberg, M., Tian, K., and Yao, M. Cross-cultural user design: divergences in chinese and western human computer interface interaction. In *HCI International 2019–Late Breaking Posters: 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21* (2019), Springer, pp. 39–45.

[149] Lin, W., Chen, H.-C., and Yueh, H.-P. Using different error handling strategies to facilitate older users' interaction with chatbots in learning information and communication technologies. *Frontiers in Psychology 12* (2021), 785815.

[150] Lindquist, E. F. Design and analysis of experiments in psychology and education.

[151] Ling, E. C., Tussyadiah, I., Tuomi, A., Stienmetz, J., and Ioannou, A. Factors influencing users' adoption and use of conversational agents: A systematic review. *Psychology & marketing 38*, 7 (2021), 1031–1051.

[152] Litman, D., Hirschberg, J., and Swerts, M. Characterizing and predicting corrections in spoken dialogue systems. *Computational linguistics 32*, 3 (2006), 417–438.

[153] Liu, Z., Li, H., Chen, A., Zhang, R., and Lee, Y.-C. Understanding public perceptions of ai conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–17.

[154] Luger, E., and Sellen, A. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), pp. 5286–5297.

[155] Lumer, E., and Buschmeier, H. Should robots be polite? expectations about politeness in human–robot interaction. *Frontiers in Robotics and AI 10* (2023), 1242127.

[156] Mahmood, A., Fung, J. W., Won, I., and Huang, C.-M. Owning mistakes sincerely: Strategies for mitigating ai errors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–11.

[157] Mahmood, A., and Huang, C.-M. Gender biases in error mitigation by voice assistants. *Proceedings of the ACM on Human-Computer Interaction 8*, CSCW1 (2024), 1–27.

[158] Mahmood, A., Wang, J., Yao, B., Wang, D., and Huang, C.-M. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies 195* (2025), 103406.

[159] Maimani, A., and Roudaut, A. Frozen suit: Toward a changeable stiffness suit and its application for haptic games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, Denver, CO, USA* (2017), pp. 6–11.

[160] Mavrina, L., Szczuka, J., Strathmann, C., Bohnenkamp, L. M., Krämer, N., and Kopp, S. "alexa, you're really stupid": A longitudinal field study on communication breakdowns between family members and a voice assistant. *Frontiers in Computer Science 4* (2022), 791704.

[161] McTear, M. Handling miscommunication: Why bother? *Recent trends in discourse and dialogue* (2008), 101–122.

[162] McTear, M., O'Neill, I., Hanna, P., and Liu, X. Handling errors and determining confirmation strategies—an object-based approach. *Speech communication 45*, 3 (2005), 249–269.

[163] Meck, A.-M., Draxler, C., and Vogt, T. Failing with grace: Exploring the role of repair costs in conversational breakdowns with in-car voice assistants. *International Journal of Human–Computer Interaction* (2023), 1–19.

[164] Merkin, R., Taras, V., and Steel, P. State of the art themes in cross-cultural communication research: A systematic and meta-analytic review. *International Journal of Intercultural Relations 38* (2014), 1–23.

[165] Meyer, E. *The culture map: Breaking through the invisible boundaries of global business.* 2016.

[166] Miehle, J. Communication style modelling and adaptation in spoken dialogue systems.

[167] Miehle, J., Minker, W., and Ultes, S. What causes the differences in communication styles? a multicultural study on directness and elaborateness. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[168] Miehle, J., Minker, W., and Ultes, S. When to say what and how: Adapting the elaborateness and indirectness of spoken dialogue systems. *Dialogue & Discourse 13*, 1 (2022), 1–40.

[169] Miehling, E., Nagireddy, M., Sattigeri, P., Daly, E. M., Piorkowski, D., and Richards, J. T. Language models in dialogue: Conversational maxims for human-ai interactions. *arXiv preprint arXiv:2403.15115* (2024).

[170] Möller, S., Engelbrecht, K.-P., and Oulasvirta, A. Analysis of communication failures for spoken dialogue systems. In *INTERSPEECH* (2007), pp. 134–137.

[171] Morse, J. M. *Essentials of qualitatively-driven mixed-method designs*. Routledge, 2016.

[172] Most, T. The use of repair strategies by children with and without hearing impairment. *Language, Speech, and Hearing Services in Schools 33*, 2 (2002), 112–123.

[173] Motta, I., and Quaresma, M. Users' error recovery strategies in the interaction with voice assistants (vas). In *Congress of the International Ergonomics Association* (2021), Springer, pp. 658–666.

[174] Moussalli, S., and Cardoso, W. Intelligent personal assistants: Can they understand and be understood by accented l2 learners? *Computer Assisted Language Learning 33*, 8 (2020), 865–890.

[175] Moussawi, S., Koufaris, M., and Benbunan-Fich, R. The role of user perceptions of intelligence, anthropomorphism, and self-extension on continuance of use of personal intelligent agents. *European Journal of Information Systems 32*, 3 (2023), 601–622.

[176] Müller, R., Paul, D., and Li, Y. Reformulation of symptom descriptions in dialogue systems for fault diagnosis: How to ask for clarification? *International Journal of Human-Computer Studies 145* (2021), 102516.

[177] Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., and Aromataris, E. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology 18* (2018), 1–7.

[178] Murad, C., Candello, H., and Munteanu, C. What's the talk on vui guidelines? a meta-analysis of guidelines for voice user interface design. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (2023), pp. 1–16.

[179] Myers, C., Furqan, A., Nebolsky, J., Caro, K., and Zhu, J. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (2018), pp. 1–7.

[180] Myers, C. M., Laris Pardo, L. F., Acosta-Ruiz, A., Canossa, A., and Zhu, J. "try, try, try again:" sequence analysis of user interaction data with a voice user interface. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (2021), pp. 1–8.

[181] Nass, C., and Moon, Y. Machines and mindlessness: Social responses to computers. *Journal of social issues 56*, 1 (2000), 81–103.

[182] Nass, C., Steuer, J., and Tauber, E. R. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1994), pp. 72–78.

[183] Nass, C. I., and Brave, S. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, 2005.

[184] Neuenswander, K. L., Bryant, G. A., and Stroessner, S. J. Gendered and machine-like features in voices affect social judgments. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)* (2024), IEEE, pp. 1408–1415.

[185] Nguyen, Q. N., and Sidorova, A. Understanding user interactions with a chatbot: A self-determination theory approach.

[186] Nguyen, T. H., Trivedi, R. H., Fukukawa, K., and Adomako, S. Investigating drivers of customer experience with virtual conversational agents. *Information Systems Frontiers* (2025), 1–27.

[187] Nomura, T., and Kinoshita, Y. Gender stereotypes in cultures: experimental investigation of a possibility of reproduction by robots in japan. In *2015 International Conference on Culture and Computing (Culture Computing)* (2015), IEEE, pp. 195–196.

[188] Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods 16*, 1 (2017), 1609406917733847.

[189] of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, C. *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press, 2001.

[190] Oh, E. Y., and Song, D. Developmental research on an interactive application for language speaking practice using speech recognition technology. *Educational Technology Research and Development 69*, 2 (2021), 861–884.

[191] Oishi, S., Hahn, J., Schimmack, U., Radhakrishan, P., Dzokoto, V., and Ahadi, S. The measurement of values across cultures: A pairwise comparison approach. *Journal of research in Personality 39*, 2 (2005), 299–305.

[192] Opfermann, C., and Pitsch, K. Reprompts as error handling strategy in human-agent-dialog? user responses to a system's display of non-understanding. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2017), IEEE, pp. 310–316.

[193] O'Neill-Brown, P. Setting the stage for the culturally adaptive agent. In *Proceedings of the 1997 AAAI fall symposium on socially intelligent agents* (1997), AAAI Press Menlo Park, CA, pp. 93–97.

[194] Pabst, A., Tabb, T., Stewart, K. M., and Danielescu, A. A qualitative study on the expectations and concerns around voice and gesture interactions in vehicles. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (2023), pp. 2155–2171.

[195] Park, H. S., Levine, T. R., Weber, R., Lee, H. E., Terra, L. I., Botero, I. C., Bessarabova, E., Guan, X., Shearman, S. M., and Wilson, M. S. Individual and cultural variations in direct communication style. *International Journal of Intercultural Relations 36*, 2 (2012), 179–187.

[196] Park, S. J., MacDonald, C. M., and Khoo, M. Do you care if a computer says sorry? user experience design through affective messages. In *Proceedings of the designing interactive systems conference* (2012), pp. 731–740.

[197] Payne, J., Szymkowiak, A., Robertson, P., and Johnson, G. Gendering the machine: Preferred virtual assistant gender and realism in self-service. In *International Workshop on Intelligent Virtual Agents* (2013), Springer, pp. 106–115.

[198] Pearl, C. *Designing voice user interfaces: Principles of conversational experiences.* " O'Reilly Media, Inc.", 2016.

[199] Perez Garcia, M., and Saffon Lopez, S. Building trust between users and telecommunications data driven virtual assistants. In *IFIP international conference on artificial intelligence applications and innovations* (2018), Springer, pp. 628–637.

[200] Pérez-Marín, D. A review of the practical applications of pedagogic conversational agents to be used in school and university classrooms. *Digital 1*, 1 (2021), 18–33.

[201] Perez-Ortiz, M., and Mantiuk, R. K. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686* (2017).

[202] Peterson, J., Pearce, P. F., Ferguson, L. A., and Langford, C. A. Understanding scoping reviews: Definition, purpose, and process. *Journal of the American Association of Nurse Practitioners 29*, 1 (2017), 12–16.

[203] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology 88*, 5 (2003), 879.

[204] Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems* (2018), pp. 1–12.

[205] Porcheron, M., Fischer, J. E., and Sharples, S. " do animals have accents?" talking with agents in multi-party conversation. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (2017), pp. 207–219.

[206] Prahl, A., and Goh, W. W. P. "rogue machines" and crisis communication: When ai fails, how do companies publicly respond? *Public Relations Review 47*, 4 (2021), 102077.

[207] Priya, P., Firdaus, M., and Ekbal, A. Computational politeness in natural language processing: A survey. *ACM Computing Surveys 56*, 9 (2024), 1–42.

[208] Putri, B. B. K., Noer, A. H., Pebriani, L. V., and Purba, F. D. Indonesian translation and adaptation of social appearance anxiety scale (saas) for early adolescent girls in indonesia. *Psychology Research and Behavior Management* (2025), 475–485.

[209] Pyae, A., and Scifleet, P. Investigating differences between native english and non-native english speakers in interacting with a voice user interface: A case of google home. In *Proceedings of the 30th Australian conference on computer-human interaction* (2018), pp. 548–553.

[210] Pyae, A., and Scifleet, P. Investigating the role of user's english language proficiency in using a voice user interface: A case of google home smart speaker. In *extended abstracts of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–6.

[211] Qian, Z., Fu, J., and Zhou, Y. Exploring cultural and intergenerational dynamics in voice assistant design for chinese older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 9*, 1 (2025), 1–18.

[212] Rau, P. P., Li, Y., and Li, D. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior 25*, 2 (2009), 587–595.

[213] Raux, A., Bohus, D., Langner, B., Black, A. W., and Eskenazi, M. Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *Ninth International Conference on Spoken Language Processing* (2006).

[214] Reeves, B., and Nass, C. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK 10*, 10 (1996), 19–36.

[215] Rheu, M., Shin, J. Y., Peng, W., and Huh-Yoo, J. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction 37*, 1 (2021), 81–96.

[216] Riefle, L., Brand, A., Mietz, J., Rombach, L., Szekat, C., and Benz, C. What fits tim might not fit tom: Exploring the impact of user characteristics on users' experience with conversational interaction modalities.

[217] Rubin, H. J., and Rubin, I. S. *Qualitative interviewing: The art of hearing data.* sage, 2011.

[218] Santos Jr, H. P., Black, A. M., and Sandelowski, M. Timing of translation in cross-language qualitative research. *Qualitative health research 25*, 1 (2015), 134–144.

[219] Schaffer, B. S., and Riordan, C. M. A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational research methods 6*, 2 (2003), 169–215.

[220] Schegloff, E. A. Practices and actions: Boundary cases of other-initiated repair. *Discourse processes 23*, 3 (1997), 499–545.

[221] Schegloff, E. A. When'others' initiate repair. *Applied linguistics 21*, 2 (2000), 205–243.

[222] Schegloff, E. A., Jefferson, G., and Sacks, H. The preference for self-correction in the organization of repair in conversation. *Language 53*, 2 (1977), 361–382.

[223] Schmitt, A., and Minker, W. *Towards adaptive spoken dialog systems*. Springer Science & Business Media, 2012.

[224] Schneider, J. Mental model shifts in human-llm interactions. *Journal of Intelligent Information Systems* (2025), 1–16.

[225] Schuetzler, R. M., Grimes, G. M., and Scott Giboney, J. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems 37*, 3 (2020), 875–900.

[226] Seaborn, K., Gessinger, I., Yoshida, S., Cowan, B. R., and Doyle, P. R. Cross-cultural validation of partner models for voice user interfaces. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (2024), pp. 1–10.

[227] Seaborn, K., Iseya, S., Hidaka, S., Kobuki, S., and Chandra, S. Play across boundaries: Exploring cross-cultural maldaimonic game experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–15.

[228] Seaborn, K., Sawa, Y., and Watanabe, M. Coimagining the future of voice assistants with cultural sensitivity. *Human Behavior and Emerging Technologies 2024*, 1 (2024), 3238737.

[229] Seminck, O. Conversational ai: Dialogue systems, conversational agents, and chatbots by michael mctear. *Computational Linguistics 49*, 1 (2023), 257–259.

[230] Shen, H., and Wang, M. Effects of social skills on lexical alignment in human-human interaction and human-computer interaction. *Computers in Human Behavior 143* (2023), 107718.

[231] Shin, J., Narayanan, S. S., Gerber, L., Kazemzadeh, A., Byrd, D., et al. Analysis of user behavior under error conditions in spoken dialogs. In *INTERSPEECH* (2002), Denver, CO.

[232] Skantze, G. Exploring human error handling strategies: Implications for spoken dialogue systems. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems* (2003).

[233] Skantze, G. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication 45*, 3 (2005), 325–341.

[234] Skantze, G. *Error handling in spoken dialogue systems-managing uncertainty, grounding and miscommunication*. Gabriel Skantze, 2007.

[235] Soley, M., and Pandya, K. V. Culture as an issue in knowledge sharing: A means of competitive advantage. *Electronic Journal of Knowledge Management 1*, 2 (2003), pp1–8.

[236] Song, M., Zhang, H., Xing, X., and Duan, Y. Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. *Journal of Retailing and Consumer Services 73* (2023), 103323.

[237] Song, S., Baba, J., Nakanishi, J., Yoshikawa, Y., and Ishiguro, H. Mind the voice!: Effect of robot voice pitch, robot voice gender, and user gender on user perception of teleoperated robots. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–8.

[238] Souvignier, B., Kellner, A., Rueber, B., Schramm, H., and Seide, F. The thoughtful elephant: Strategies for spoken dialog systems. *IEEE Transactions on Speech and Audio Processing 8*, 1 (2000), 51–62.

[239] Squires, A. Methodological challenges in cross-language qualitative research: A research review. *International journal of nursing studies 46*, 2 (2009), 277–287.

[240] Steenkamp, J.-B. E. The role of national culture in international marketing research. *International marketing review 18*, 1 (2001), 30–44.

[241] Steichen, B., Kosasih, W., and Becerra, C. Inferring search user language proficiency from eye gaze data. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (2022), pp. 211–220.

[242] Stent, A. J., Huffman, M. K., and Brennan, S. E. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication 50*, 3 (2008), 163–178.

[243] Sugimoto, K., Takizawa, Y., Iida, J., Nakayama, H., and Furlong, M. J. Japanese cultural adaptation and validation of the social emotional health survey-higher education. *Educational and Developmental Psychologist 41*, 2 (2024), 178–190.

[244] Svennevig, J. Echo answers in native/non-native interaction. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) 13*, 2 (2003), 285–309.

[245] Swerts, M., Litman, D. J., and Hirschberg, J. Corrections in spoken dialogue systems. In *Interspeech* (2000), pp. 615–618.

[246] Taras, V., Kirkman, B. L., and Steel, P. Examining the impact of culture's consequences: a three-decade, multilevel, meta-analytic review of hofstede's cultural value dimensions. *Journal of applied psychology 95*, 3 (2010), 405.

[247] Tian, L., and Oviatt, S. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI) 10*, 2 (2021), 1–32.

[248] Traum, D., and Dillenbourg, P. Miscommunication in multi-modal collaboration. In *AAAI Workshop on Detecting, Repairing, And Preventing Human–Machine Miscommunication* (1996), pp. 37–46.

[249] Traum, D. R. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium* (1999), pp. 124–131.

[250] Trovato, G., Zecca, M., Sessa, S., Jamone, L., Ham, J., Hashimoto, K., and Takanishi, A. Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by egyptians and japanese. *Paladyn, Journal of Behavioral Robotics 4*, 2 (2013), 83–93.

[251] Tzeng, J.-Y. Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human-Computer Studies 61*, 3 (2004), 319–345.

[252] Tzeng, J.-Y. Matching users' diverse social scripts with resonating humanized features to create a polite interface. *International journal of human-computer studies 64*, 12 (2006), 1230–1242.

[253] Uchida, T., Minato, T., Koyama, T., and Ishiguro, H. Who is responsible for a dialogue breakdown? an error recovery strategy that promotes cooperative intentions from humans by mutual attribution of responsibility in human-robot dialogues. *Frontiers in Robotics and AI 6* (2019), 29.

[254] Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., and Seaborn, K. Trust in human-ai interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (2022), pp. 1–7.

[255] Vailshery, L. S. Number of digital voice assistants in use worldwide 2019-2024. *Statista Research Departement.[Online]. Available: https://www. statista. com/statistics/973815/worldwide-digital-voice-assistant-in-use* (2021).

[256] Van Nes, F., Abma, T., Jonsson, H., and Deeg, D. Language differences in qualitative research: is meaning lost in translation? *European journal of ageing 7* (2010), 313–316.

[257] Wali, A., Mahamad, S., and Sulaiman, S. Task automation intelligent agents: A review. *Future Internet 15*, 6 (2023), 196.

[258] Wang, H., Wang, L., Du, Y., Chen, L., Zhou, J., Wang, Y., and Wong, K.-F. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789* (2023).

[259] Wang, J. From self-efficacy to human-computer interaction design. In *Journal of Physics: Conference Series* (2019), vol. 1168, IOP Publishing, p. 032060.

[260] Wang, X., Hwang, Y., and Guchait, P. When robot (vs. human) employees say "sorry" following service failure. *International Journal of Hospitality & Tourism Administration 24*, 4 (2023), 540–562.

[261] Wang, Y.-L., and Lo, C.-W. The effects of response time on older and young adults' interaction experience with chatbot. *BMC psychology 13*, 1 (2025), 150.

[262] Wasti, S. M., Pu, K. Q., and Neshati, A. Large language user interfaces: Voice interactive user interfaces powered by llms. *arXiv preprint arXiv:2402.07938* (2024).

[263] Watson, C., Liu, W., and MacDonald, B. The effect of age and native speaker status on intelligibility. In *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)* (2013).

[264] Weigand, E. Misunderstanding: The standard case. *Journal of pragmatics 31*, 6 (1999), 763–785.

[265] Weisstein, E. W. Bonferroni correction. *https://mathworld. wolfram. com/* (2004).

[266] Weitz, K., Vanderlyn, L., Vu, N. T., and André, E. " it's our fault!": insights into users' understanding and interaction with an explanatory collaborative dialog system.

[267] Welch, C., and Piekkari, R. Crossing language boundaries: Qualitative interviewing in international business. *Management International Review 46* (2006), 417–437.

[268] Wenzel, K., Devireddy, N., Davison, C., and Kaufman, G. Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–14.

[269] Wenzel, K., and Kaufman, G. Designing for harm reduction: Communication repair for multicultural users' voice interactions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–17.

[270] Wilhelm, M., Schwaetzer, E., Otten, T., Zobel, T. E. I., and Schumacher, K. Troubleshooting conversations: Exploring chatbot repair strategies. In *Proceedings of Mensch und Computer 2024*. 2024, pp. 386–391.

[271] Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., and Martins, A. Brave new world: service robots in the frontline. *Journal of service management 29*, 5 (2018), 907–931.

[272] Wolters, K. M., Engelbrecht, K.-P., Gödde, F., Möller, S., Naumann, A., and Schleicher, R. Making it easier for older people to talk to smart homes: The effect of early help prompts. *Universal Access in the Information Society 9* (2010), 311–325.

[273] Wu, Y., Edwards, J., Cooney, O., Bleakley, A., Doyle, P. R., Clark, L., Rough, D., and Cowan, B. R. Mental workload and language production in non-native speaker ipa interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (2020), pp. 1–8.

[274] Wu, Y., Porcheron, M., Doyle, P., Edwards, J., Rough, D., Cooney, O., Bleakley, A., Clark, L., and Cowan, B. Comparing command construction in native and non-native speaker ipa interaction through conversation analysis. In *Proceedings of the 4th conference on conversational user interfaces* (2022), pp. 1–12.

[275] Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P. R., Clark, L., and Cowan, B. R. See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers. In *22nd international conference on human-computer interaction with mobile devices and services* (2020), pp. 1–9.

[276] Würtz, E. Intercultural communication on web sites: A cross-cultural analysis of web sites from high-context cultures and low-context cultures. *Journal of computer-mediated communication 11*, 1 (2005), 274–299.

[277] Xie, T., Pentina, I., and Hancock, T. Friend, mentor, lover: does chatbot engagement lead to psychological dependence? *Journal of service Management 34*, 4 (2023), 806–828.

[278] Xu, Z., Wang, G., Zhai, S., and Liu, P. When automation fails: Examining the effect of a verbal recovery strategy on user experience in automated driving. *International Journal of Human–Computer Interaction* (2023), 1–11.

[279] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., and Hu, X. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* (2023).

[280] Yi, Z., Ouyang, J., Xu, Z., Liu, Y., Liao, T., Luo, H., and Shen, Y. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys 58*, 6 (2025), 1–38.

[281] You, H.-C., Luo, D.-X., and Ho, L.-Y. A kansei engineering approach to virtual personality of embodied voice assistants. In *International Conference on Kansei Engineering & Emotion Research* (2024), Springer, pp. 206–219.

[282] You, H.-C., and Weng, H.-Y. Effects of nonverbal communication on chatbot's perceived personality and user satisfaction. In *9th International Conference on Kansei Engineering and Emotion Research. KEER2022. Proceedings* (2022), Kansei Engineering and Emotion Research (KEER), pp. 265–274.

[283] Yu, C.-E., and Ngan, H. F. B. The power of head tilts: gender and cultural differences of perceived human vs human-like robot smile in service. *Tourism Review 74*, 3 (2019), 428–442.

[284] Yuan, S., Brüggemeier, B., Hillmann, S., and Michael, T. User preference and categories for error responses in conversational user interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (2020), pp. 1–8.

[285] Yus, F. Misunderstandings and explicit/implicit communication. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) 9*, 4 (1999), 487–517.

[286] Zargham, N., Pfau, J., Schnackenberg, T., and Malaka, R. "i didn't catch that, but i'll try my best": Anticipatory error handling in a voice controlled game. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–13.

[287] Zeng, J., Fan, D., Zhou, X., and Tang, J. Chatbot with resilience: The impact of repair strategies on customer satisfaction in conversational breakdowns. In *Wuhan International Conference on E-business* (2024), Springer, pp. 306–317.

[288] Zgorzelski, A., Schmitt, A., Heinroth, T., and Minker, W. Repair strategies on trial: Which error recovery do users like best? In *Eleventh Annual Conference of the International Speech Communication Association* (2010).

[289] Zue, V., and Seneff, S. Spoken dialogue systems. *Springer Handbook of Speech Processing* (2008), 705–722.

Appendix A

Participant Information Sheets and Consent Forms for the Pilot Study (Ethical ID: 2769)

A.1 Participant Information Sheet

Name of department: Computer and Information Sciences
Title of the study: Verifying Distinctiveness and Realism in Repair Strategies for Digital Voice Assistants (DVAs): A Pilot Study

Introduction:

I am Essam Alghamdi, a PhD student in the Department of Computer and Information Sciences at the University of Strathclyde. My research focuses on how people interact with Digital Voice Assistants (DVAs), such as Siri and Alexa. Specifically, I am investigating user preferences and perspectives on Digital Voice Assistants (DVAs)'s responses and the role of user characteristics in this.

What is the purpose of this research?

The purpose of this research is to (1) confirm the different types of responses that Digital Voice Assistants (DVAs) can provide and (2) evaluate realism of the task scenario used in this research.

Do you have to take part?

Participation in this study is entirely voluntary, and you have the right to withdraw from the research at any time without any negative consequences.

What will you do in the project?

In this pilot study, you will engage in a simulated task of booking a flight using a Digital Voice Assistant (DVA). You will begin by reading a text-based scenario on your screen, which presents a dialogue between you and the DVA. Next, you will listen to a voice response from the DVA. After each interaction, you will answer a multiple-choice question to identify the category that best describes the DVA's response. Additionally, you will complete a short questionnaire to evaluate the realism of the task and how closely it mimics real-world interactions with DVAs.

The study will be conducted entirely online and will take approximately 15 minutes. As a token of appreciation for your time, you will have the option to provide your first name and email address to receive a £5 Starbucks gift card. Your name and email will only be used to send the reward.

Why have you been invited to take part?

You have been invited to participate in this research study because you meet the following criteria:

- You are aged 18 or older and
- You have proficiency in English, whether as a native or a second language.

What information is being collected in the project?
This pilot study will collect data regarding your nationality, current residence, age, gender, preferred type of system response and your first and second languages (if applicable). Additionally, this study will gather your feedback on the realism of the task scenarios by completing a short questionnaire. All collected data will be anonymized to protect your privacy, with no identifiable information being gathered. Each participant will be assigned a unique ID number to ensure anonymity.
Who will have access to the information?
The data will be anonymized, and only the research team will have access to it. Anonymized data (i.e., data that do not identify you personally) cannot be withdrawn once they have been included in the study. Anonymized data may be used in the final thesis for this research and any resulting publications.
Where will the information be stored and how long will it be kept for?
Data will be securely stored on the University of Strathclyde's servers (OneDrive). The anonymized data will be retained for 10 years, after which it will be deleted. Anonymized quotations and findings may be used in publications and the final thesis.

Thank you for reading this information – please ask any questions if you are unsure about what is written here.
All personal data will be processed in accordance with data protection legislation. Please read the University of Strathclyde Privacy Notice for Research Participants for more information about your rights under the legislation.
If you do not want to participate, we thank you for your interest and attention.

What happens next?
If you would like to participate in the research, or require any further information, please contact Essam Alghamdi (essam.alghamdi@strath.ac.uk). You will be asked to confirm and agree to a participant consent form before engaging with the study. We hope that the results of this study will be published. If you wish to receive a copy of any resulting publication, please email us.

Researcher contact details:

Essam Alghamdi
Computer and Information Sciences – University of Strathclyde
26 Richmond Street, Glasgow G1 1XQ
essam.alghamdi@strath.ac.uk
Chief Investigator details:
Professor Martin Halvey
Computer and Information Sciences – University of Strathclyde
26 Richmond Street, Glasgow G1 1XQ
martin.halvey@strath.ac.uk
Second Supervisor details:
Dr Emma Nicol
Computer and Information Sciences – University of Strathclyde
26 Richmond Street, Glasgow G1 1XQ
emma.nicol@strath.ac.uk

This research was granted ethical approval by the University of Strathclyde Computer and Information Sciences Ethics Committee. The reference number of this ethical is 2769.
If you have any questions/concerns or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Computer and Information Sciences Ethics Committee
University of Strathclyde
Livingstone Tower
26 Richmond Street
Glasgow
G1 1XQ
Email: ethics@cis.strath.ac.uk

Would you like to continue with this survey?

◯ Yes, I agree to participate.

◯ No, I do not wish to participate.

→

A.2 Questionnaire Consent Form

Please Read the following terms and click next '-->' button if you agree with all of the terms and wish to proceed:

□ I confirm that I have read and understood the Participant Information Sheet for the above project and the researcher has answered any queries to my satisfaction.
□ I confirm that I have read and understood the Privacy Notice for Participants in Research Projects and understand how my personal information will be used and what will happen to it (i.e. how it will be stored and for how long).
□ I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences.
□ I understand that I can request the withdrawal from the study of some personal information and that whenever possible researchers will comply with my request. This includes the following personal data: my personal information from questionnaire that identify me.
□ I understand that anonymised data (i.e. data that do not identify me personally) cannot be withdrawn once they have been included in the study.
□ I understand that any information recorded in the research will remain confidential and no information that identifies me will be made publicly available.
□ I consent to being a participant in the project.

← →

A.3 Questionnaires

A.3.1 Demographic Questionnaire

Which of the following is your native (first) language?

- ○ English
- ○ Arabic
- ○ Other

Which of the following is your second language (if any)?

- ○ English
- ○ Arabic
- ○ I don't have a second language
- ○ Other

What is your nationality (the country you identify with)?

Please, enter your age (year)?

What is your gender?

- ◯ woman
- ◯ man
- ◯ non-binary
- ◯ prefer not to say
- ◯ prefer to self-describe

→

A.3.2 Verifying the Distinctiveness of Repair Strategies

Overview

Part One: in this survey, you will be presented with a text-based scenario where you are booking a flight using a digital voice assistant, such as Siri or Alexa. During the interaction, the voice assistant (referred to as "**Echo**") will experience a misunderstanding and fail to recognize the city you want to fly to.

After the scenario, you will hear one voice response from Echo, representing a specific strategy for handling the misunderstanding.

Your task is to listen to Echo's response and select the category that best describes the strategy used.

← →

Scenario 1 of 5:

You need to book a flight for tomorrow because you have important plans. **Echo** is your digital voice assistant, similar to Siri or Alexa, and it usually helps you with travel arrangements. However, during the conversation, **Echo** is unable to understand the city you want to fly to, leading to a breakdown in the conversation.

Conversation:

Echo: *"Hello, I am Echo, your digital voice assistant.
How can I help today?"*

You: *"I need to book a flight for tomorrow."*

Echo: *"Got it! And which city are you flying to?"*

You: *"I want to fly to London."*

At this point, **Echo doesn't understand you and initiates this response:**

==

Please listen carefully, You can replay the audio as many times as needed.

0:00 / 0:05

Which of the following categories best describes Echo's response? (Choose only one

- ◯ **Confirmation**: the system (Echo) tries to acknowledge non-understanding by guessing what the user meant and seeks confirmation from the user.
- ◯ **Information:** the system (Echo) explains why it couldn't understand.
- ◯ **Social**: the system (Echo) politely apologizes for the non-understanding.
- ◯ **Solve:** the system (Echo) offers a list of possible options based on partial understanding.
- ◯ **Ask :** the system (Echo) directly asks the user to repeat the input without additional context or elaboration.

← →

A.3.3 Evaluating the Realism of the Task Scenario

Part Two: Now that you have completed the interaction with the digital assistant, we would like to understand your thoughts on how realistic the scenario felt.

Please indicate how much you **agree** or **disagree** with the following statements about the experience, using a scale from 1 (strongly disagree) to 7 (strongly agree).

I can imagine myself in the scenario (booking a flight through a digital voice assistant).

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

I can imagine myself interacting with a digital voice assistant like the one in this study (Echo).

- ◯ Strongly disagree
- ◯ Disagree
- ◯ Somewhat disagree
- ◯ Neither agree nor disagree
- ◯ Somewhat agree
- ◯ Agree
- ◯ Strongly agree

Were there any moments in the scenario where the system's responses (Echo) felt unrealistic?

- ◯ No
- ◯ Yes, please explain

← →

Appendix B

Participant Information Sheets and Consent Forms for the Empirical Studies (Ethical ID: 2781)

B.1 Participant Information Sheet

Name of department: Computer and Information Science
Title of the study: Cross-Cultural Preference for Communication Response Strategies in Digital Voice Assistants (DVAs): The Role of User Characteristics

Introduction

I am Essam Alghamdi, a PhD student in the Department of Computer and Information Sciences at the University of Strathclyde. My research focuses on how people interact with Digital Voice Assistants (DVAs), such as Siri and Alexa. Specifically, I am investigating user preferences and perspectives on system responses and the role of user characteristics such as cultural background, computer self-efficacy, and language proficiency.

What is the purpose of this research?

The purpose of this research is to explore how individual characteristics e.g. cultural background, computer self-efficacy, and language proficiency influence preferences and perspectives regarding response strategies of Digital Voice Assistants (DVAs). Existing literature highlights the importance of user-centred design in DVAs by taking individual characteristics into account, as these factors significantly affect overall interaction experiences.

Do you have to take part?
Participation in this study is entirely voluntary, and you have the right to withdraw from the research at any time without any negative consequences.
What will you do in the project?
In this online survey, you will be initially asked a few multiple-choice questions to determine your language proficiency and computer self-efficacy. Then, you will engage in a simulated task of booking a flight using a Digital Voice Assistant (DVA). You will begin by reading a text-based scenario on your screen, which presents a dialogue between you and the DVA. Next, you will listen to two voice responses from the DVA. After each interaction, you will be asked to select which of the two voice responses you prefer.
The study will be conducted entirely online and will take approximately 13 minutes. As a token of appreciation for your time, you will have the option to provide your first name and email address to receive a £5 Starbucks gift card. Your name and email will only be used to send the reward.
For the interview (Optional): At the end of the questionnaire, you will have the option to indicate whether you would like to participate in a follow-up interview. If you choose to participate, you will be asked to provide your contact information—either your email or WhatsApp number, based on your preference. This information will be used exclusively to schedule a convenient

time for the interview, which will take place online via Microsoft Teams.
The interview will collect information about your previous experiences (if any) with Digital Voice Assistants (DVAs) and explore your preferences and perspectives on DVAs' responses. Interviews are expected to last no longer than one hour. If you choose to take part in the interview, once completed, as a token of appreciation for your time, you will have the option to receive a £10 Starbucks gift card. Your name and email will only be used to send the reward.

Why have you been invited to take part?

You have been invited to participate in this research study because you meet the following criteria:

- You are aged 18 or over.
- You have proficiency in English and/or Arabic, with one of these languages being your native or second language.

What information is being collected in the project?

This research will collect data regarding your language proficiency, computer self-efficacy, and preferences for responses from Digital Voice Assistants (DVAs). Additionally, this research will gather data regarding your nationality, current residence, age, gender, and your first and second languages (if applicable). All collected data will be anonymized to protect your privacy, with no identifiable information being gathered. Each participant will be assigned a unique ID number to ensure anonymity.

For the interview (Optional): If you choose to participate in the interview, you will discuss your previous experiences (if any) with Digital Voice Assistants (DVAs) and share your preferences and perspectives on DVAs' responses. The interview will be audio-recorded, and the recordings will be transcribed and anonymized to protect your privacy. No identifiable information will be collected, and each participant will be assigned a unique ID number to ensure anonymity. Interviews are expected to last no longer than one hour. As a token of appreciation for your time, you will have the option to receive a £10 Starbucks gift card. Your name and email will only be used to send the reward.

Who will have access to the information?

The data will be anonymized, and only the research team will have access to it. Anonymized data and quotes (i.e., data that do not identify you personally) cannot be withdrawn once they have been included in the study. Anonymized data may be used in the final thesis for this research and any resulting publications.

Where will the information be stored and how long will it be kept for?

Data from the questionnaire will be securely stored on the University of Strathclyde's servers (OneDrive). The anonymized data will be retained for 10 years, after which it will be deleted. Anonymized findings and quotations may be used in publications and the final thesis.

For the interview (Optional): Anonymised transcription files will be kept for 10 years, after which they will be deleted. Anonymized quotations from the interviews may be used in publications and the final thesis.

Thank you for reading this information – please ask any questions if you are unsure about what is written here.
All personal data will be processed in accordance with data protection legislation. Please read the University of Strathclyde Privacy Notice for Research Participants for more information about your rights under the legislation.

What happens next?

If you would like to participate in the research, or require any further information, please contact Essam Alghamdi (essam.alghamdi@strath.ac.uk). You will be asked to confirm and agree to a participant consent form before starting the study. We hope that the results of this study will be published. If you wish to receive a copy of any resulting publication, please email us. If you do not want to participate, we thank you for your interest and attention.

Researchers contact details:
Essam Alghamdi
Computer and Information Sciences – University of Strathclyde
26 Richmond Street, Glasgow G1 1XQ
essam.alghamdi@strath.ac.uk

Chief Investigator details:
Professor Martin Halvey
Computer and Information Sciences – University of Strathclyde
26 Richmond Street, Glasgow G1 1XQ
martin.halvey@strath.ac.uk

Second Supervisor details:
Dr Emma Nicol
Computer and Information Sciences – University of Strathclyde
26 Richmond Street, Glasgow G1 1XQ
emma.nicol@strath.ac.uk
This research was granted ethical approval by the University of Strathclyde Computer and Information Sciences Ethics Committee. The reference number of this ethical is 2781.
If you have any questions/concerns, during or after the research, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Computer and Information Sciences Ethics Committee
University of Strathclyde
Livingstone Tower
26 Richmond Street
Glasgow
G1 1XQ
Email: ethics@cis.strath.ac.uk

Would you like to continue with this survey?

- (●) Yes, I agree to participate.
- ( ) No, I do not wish to participate.

>> Next

B.2 Questionnaires Consent Form

Please Read the following terms and click '**>> Next**' button if you agree with all of the terms and wish to proceed:

- I confirm that I have read and understood the Participant Information Sheet for the above project and the researcher has answered any queries to my satisfaction.
- I confirm that I have read and understood the Privacy Notice for Participants in Research Projects and understand how my personal information will be used and what will happen to it (i.e. how it will be stored and for how long).
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences.
- I understand that I can request the withdrawal from the study of some personal information and that whenever possible researchers will comply with my request. This includes the following personal data: my personal information from the questionnaire identifies me.
- I understand that anonymised data (i.e. data that do not identify me personally) cannot be withdrawn once they have been included in the study.
- I understand that any information recorded in the research will remain confidential and no information that identifies me will be made publicly available.
- I consent to being a participant in the project.

<< Back

>> Next

Before you begin:
To participate in this experiment, you will need to listen to short audio clips.
Please make sure that:

- You have access to speakers or headphones to hear the audio clearly.
- You are in a quiet environment where you can focus without distractions.

Do you have access to speakers or headphones, and are you in a quiet place suitable for this experiment?

○ **Yes**, I have access to speakers or headphones and am in a quiet environment.

○ **No**, I do not have access to speakers or headphones and/or am not in a quiet environment.

>> Next

B.3 Interview Consent Form

Interview Consent Form

Name of department: Computer and Information Science
Title of the study: Cross-Cultural Preference for Communication Repair strategies in Spoken Dialogue Systems: The Role of User Characteristics

☐ I confirm that I have read and understood the Participant Information Sheet for the above project and the researcher has answered any queries to my satisfaction.

☐ I confirm that I have read and understood the Privacy Notice for Participants in Research Projects and understand how my personal information will be used and what will happen to it (i.e. how it will be stored and for how long).

☐ I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences.

☐ I understand that I can request the withdrawal from the study of some personal information and that whenever possible researchers will comply with my request. This includes the following personal data:

- my personal information from the interview that identifies me.
- audio recordings of interviews that identify me.
- my personal information from transcripts.

☐ I understand that anonymised data (i.e. data that do not identify me personally) cannot be withdrawn once they have been included in the study.

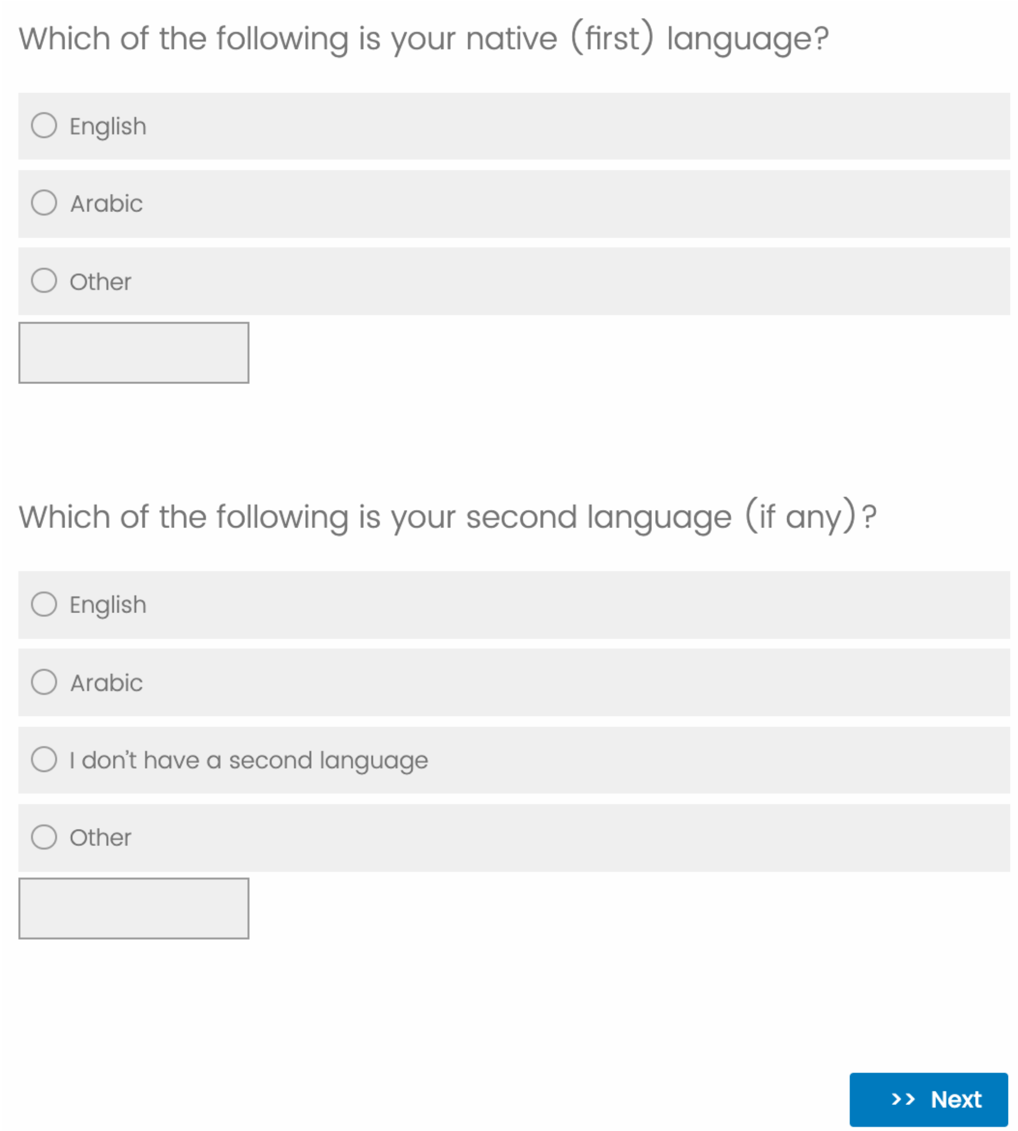
☐ I understand that any information recorded in the research will remain confidential and no information that identifies me will be made publicly available.

☐ I consent to being audio recorded during my interview as part of the project.

☐ I consent to being a participant in the project.

B.4 Questionnaires and Interview

B.4.1 Demographic Questionnaire



Which of the following is your native (first) language?

- English
- Arabic
- Other

Which of the following is your second language (if any)?

- English
- Arabic
- I don't have a second language
- Other

>> Next

What is your gender?

- ◯ woman
- ◯ man
- ◯ non-binary
- ◯ prefer not to say
- ◯ prefer to self-describe

Please, enter your age (year)?

What is your nationality (the country you identify with)?

- ◯ United Kingdom of Great Britain and Northern Ireland
- ◯ Saudi Arabia
- ◯ Other

B.4.2 Prior Experience with Technology and DVAs

The next four questions assess your prior experience with technologies and Digital Voice Assistants (DVAs) like Apple's Siri, Amazon Alexa, or Google Assistant.
Please rate your level of agreement with the following statements regarding your experience.

I consider myself an advanced technology user:

- ○ Strongly agree
- ○ Somewhat agree
- ○ Neither agree nor disagree
- ○ Somewhat disagree
- ○ Strongly disagree

I am eager to try new technologies:

- ◯ Strongly agree
- ◯ Somewhat agree
- ◯ Neither agree nor disagree
- ◯ Somewhat disagree
- ◯ Strongly disagree

I am familiar with Digital Voice Assistances (DVAs):

- ◯ Strongly agree
- ◯ Somewhat agree
- ◯ Neither agree nor disagree
- ◯ Somewhat disagree
- ◯ Strongly disagree

I use Digital Voice Assistances (DVAs) frequently:

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree

>> Next

B.4.3 Computer Self-Efficacy

The following five questions ask you to rate your **CURRENT** ability to use Digital Voice Assistants (DVAs) such as Apple's Siri or Amazon Alexa for an unfamiliar/new task.
Please rate your level of agreement with the following statements:

I have mastered the use of Digital Voice Assistants (DVAs):

- ○ Strongly agree
- ○ Agree
- ○ Somewhat agree
- ○ Neither agree nor disagree
- ○ Somewhat disagree
- ○ Disagree
- ○ Strongly disagree

I cannot yet use Digital Voice Assistants (DVAs) as effectively as I would like:

- ◯ Strongly agree
- ◯ Agree
- ◯ Somewhat agree
- ◯ Neither agree nor disagree
- ◯ Somewhat disagree
- ◯ Disagree
- ◯ Strongly disagree

I am able to perform tasks well using Digital Voice Assistants (DVAs):

- ◯ Strongly agree
- ◯ Agree
- ◯ Somewhat agree
- ◯ Neither agree nor disagree
- ◯ Somewhat disagree
- ◯ Disagree
- ◯ Strongly disagree

It is not yet possible for me to use Digital Voice Assistants (DVAs) at the level I prefer:

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

I think my ability to use Digital Voice Assistants (DVAs) can be improved substantially:

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

<< Back

>> Next

B.4.4 Pairwise Comparison Experiment on Preferences

Scenario 1 of 10:

You need to book a flight for tomorrow because you have important plans. **Echo** is your digital voice assistant, similar to Siri or Alexa, and it usually helps you with travel arrangements. However, during the conversation, **Echo** is unable to understand the city you want to fly to, leading to a breakdown in the conversation.

Conversation:
Echo: *"Hello, I am Echo, your digital voice assistant.*
How can I help today?"
You: *"I need to book a flight for tomorrow."*
Echo: *"Got it! And which city are you flying to?"*
You: *"I want to fly to London."*
At this point, **Echo doesn't understand your last input and initiates one of the following responses.**

==

Instruction:
Please listen carefully, **Which of these two system responses from Echo do you prefer for handling the situation?**

○ ▶ 0:00 / 0:03

○ ▶ 0:00 / 0:03

→ Next

B.4.5 Self-Assessment of English Language Proficiency

Please Use the following statements to self-assess your **listening level** of English as your second language.
Choose the **highest** level you feel accurately represents your ability:

- ◯ **Level A1:** I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.
- ◯ **Level A2:** I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.
- ◯ **Level B1:** I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.
- ◯ **Level B2:** I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.
- ◯ **Level C1:** I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signaled explicitly. I can understand television programmes and films without too much effort.
- ◯ **Level C2:** I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.

→ Next

B.4.6 Interview Questions

<table>
<tr><th colspan="3">Semi-Structured Interview Guideline:
The Effect of Cultural Background and Interaction Language on Preferences for Repair Strategies in Voice Assistants</th></tr>
<tr><th>Topic</th><th>Questions</th><th>Prompts</th></tr>
<tr><td>Introduction</td><td>1. Greeting:
“Hello (__) Thank you for agreeing to talk to me and participate in this study.
How are you today?”

2. Purpose of the interview:
• “The purpose of this interview is to understand your preferences for different responses from voice assistants when communication breakdowns occur.
• voice assistants are software applications designed to assist you with different tasks, such as booking appointments or flights, playing music, or setting alarms.
• Examples of voice assistants are Siri on Apple phones, Google Assistant on Android devices and Alexa from Amazon.
• By "conversational breakdown," I mean, situations where voice assistants cannot understand you.</td><td></td></tr>
<tr><td>Consent</td><td>• In this interview: Your responses will remain confidential and will only be used for the research purpose.
• The interview will be audio recorded and transcribed, and you can stop the interview or withdraw at any time.
• Our discussion today should last between 30 and 40 minutes, but you can talk for as little or as long as you would like.
• In the text chat, I will send you the consent form to have your agreement and then we can begin the interview.</td><td></td></tr>
</table>

	• Do you have any questions or comments you would like to discuss further before beginning the interview?	
Warm-Up Questions	• Have you used a voice assistant, such as Siri or Alexa, before? ○ If yes, could you describe your overall experience with using a voice assistant?	**Prompt:** What are the reasons that you have not used a voice assistant so far?

Main Questions	• **<u>General:</u>**
	  ○ Have there been times when the voice assistant could not understand you? How did you feel about those instants?
	  ○ What was the VA's response in trying to clarify your request?
		Prompt:
		Was this in your first language or a second language?

	• **<u>Social/politeness:</u>** ○ Imagine you are using a voice assistant to book a flight to London, but the assistant doesn't understand something you've said. How important is it to you that the voice assistant uses polite phrases like 'I'm sorry' or 'please' when asking you to repeat something? How does this affect your overall experience?	**Prompt:** In what situations do you find polite language from a voice assistant to be especially important or unnecessary?

	• **<u>Explanting the cause of error</u>**
	  ○ How do you feel about a voice assistant explaining why it could not understand you, such as saying, 'I couldn't understand you because of background noise'?
	  ○ When the system doesn't understand you after several tries, would you prefer it to explain why rather than just asking you to repeat again? Why?
	  ○
		Prompt: Can you share any specific challenges or experiences you've had with this strategy when interacting in a (non)-native language?

	• **<u>Providing Options</u>**
	  ○ In situations where the voice assistant cannot understand you, it tries to address the misunderstanding by providing multiple options. For example, the assistant might say, 'I could not understand you. Did you mean Luton, Sunderland, or London?' How do you feel about this type of response?
	  ○ If none of the options provided by the voice assistant to address a misunderstanding are correct, would that change your preference for this type of response? Why or why not?
		Prompt:
		If the voice assistant responds with incorrect options, like 'Did you mean Luton, Sunderland, or Liverpool?' would you prefer the assistant to include an option such as 'none of these'? Why or why not?

	• **<u>Confirmation:</u>**
	  ○ How do you feel about a voice assistant repeating back what you said to **confirm details**? For example, when booking a flight, the voice assistant says, 'You want to travel to London, is that correct?
	  ○ When the voice assistant cannot understand you, would you prefer a voice assistant to repeat back what you said to confirm, or use a different strategy, such as explaining the error? Please explain your preference.
	• **<u>language proficiency</u>**
	  ○ Imagine you have conversation with a voice assistant in a language that is not your first language.
	    ▪ Would you prefer the assistant to repeat back what you said for confirmation, or would you prefer a different strategy that includes more detailed options and explanations of errors? Why?
	  ○ Do you find it easier to respond to simple yes-or-no questions compared to other types of interactions when using a voice assistant in a non-native language? Why or why not?
		Prompt:
		Do you find it easier when a voice assistant asks you simple yes-or-no questions, compared to other types of interactions, especially in a non-native language? Why or why not?

Closing Questions	• Do you have any suggestions for how DVSs could handle communication breakdowns better?	**Prompts:**

	• Is there anything else you'd like to share about your experiences with voice assistants or these repair strategies?	Do you think voice assistant should be designed differently for specific cultures or countries? Why ? Do you think voice assistant should change how they communicate, especially during misunderstanding, depending on whether the user is speaking in their native or non-native langue why ?
Thanks	• Thank you for your time and for sharing your valuable experiences. If you have any questions or thoughts later, feel free to contact me.	

Appendix C

Supplementary Statistical Analyses

C.1 Group Comparability Analyses (t-tests)

Independent-Samples t-Test Results for Individual-Level Characteristics (SL1 vs. SL2)*			
Individual Factors	*t*	*df*	*p*
Age	-0.852	97	0.396
Computer Self-Efficacy	0.609	97	0.544
Experience with Technology	0.942	97	0.348
Experience with DVA	1.237	97	0.219

* Independent-samples t-tests were conducted to compare Saudi native (SL1) and Saudi non-native (SL2) participants on individual-level characteristics. All tests were two-tailed. No statistically significant differences were observed (all $p > .05$).

C.2 Binary logistic regression predicting preferences for the five repair strategies

C.2.1 Social strategy:

Logistic regression predicting preference for the Social strategy

Predictor	β	SE	Wald	p	OR (Exp(B))	95% CI
Tech experience	−0.103	0.169	0.372	0.542	0.902	[0.648, 1.256]
DVA experience	−0.020	0.142	0.02	0.886	0.98	[0.742, 1.295]
Self-efficacy	−0.051	0.143	0.126	0.723	0.951	[0.718, 1.258]
Gender	0.046	0.227	0.04	0.841	1.047	[0.671, 1.633]
Age	0.013	0.012	1.072	0.301	1.013	[0.989, 1.038]
Constant	0.704	0.763	0.849	0.357	2.021	—

C.2.2 Confirmation strategy:

Logistic regression predicting preference for the Confirmation strategy

Predictor	β	SE	Wald	p	OR (Exp(B))	95% CI
Tech experience	0.082	0.163	0.255	0.614	1.086	[0.788, 1.496]
DVA experience	**0.361**	**0.143**	**6.339**	**0.012**	**1.435**	**[1.083, 1.901]**
Self-efficacy	0.031	0.142	0.048	0.827	1.031	[0.780, 1.363]
Gender	-0.117	0.226	0.265	0.607	0.89	[0.571, 1.387]
Age	-0.009	0.012	0.58	0.446	0.991	[0.967, 1.015]
Constant	-1.263	0.753	2.812	0.094	0.283	—

C.2.3 Information strategy:

Logistic regression predicting preference for the Information strategy

Predictor	β	SE	Wald	p	OR (Exp(B))	95% CI
Tech experience	0.247	0.173	2.04	0.153	1.28	[.91, 1.80]
DVA experience	**–0.412**	**0.146**	**7.96**	**0.005**	**0.66**	**[.50, .88]**
Self-efficacy	0.11	0.147	0.56	0.454	1.12	[.84, 1.49]
Gender	0.238	0.233	1.04	0.308	1.27	[.80, 2.00]
Age	0.005	0.012	0.14	0.713	1.01	[.98, 1.03]
Constant	-0.866	0.776	1.245	0.265	0.421	

C.2.4 Solve strategy:

Logistic regression predicting preference for the Solve strategy

Predictor	β	SE	Wald	p	OR (Exp(B))	95% CI
Tech experience	−0.070	0.166	0.175	0.676	0.933	[0.674, 1.292]
DVA experience	0.072	0.142	0.253	0.615	1.074	[0.813, 1.420]
Self-efficacy	**−0.323**	**0.147**	**4.82**	**0.028**	**0.724**	**[0.543, 0.966]**
Gender	−0.265	0.231	1.317	0.251	0.767	[0.488, 1.206]
Age	**−0.030**	**0.013**	**5.607**	**0.018**	**0.97**	**[0.946, 0.995]**
Constant	2.261	0.772	8.589	0.003	9.594	—

C.2.5 Ask strategy:

Logistic regression predicting preference for the Ask strategy

Predictor	β	SE	Wald	p	OR (Exp(B))	95% CI
Tech experience	-0.153	0.167	0.846	0.358	0.858	[0.619, 1.189]
DVA experience	-0.014	0.141	0.009	0.922	0.986	[0.749, 1.300]
Self-efficacy	0.232	0.144	2.601	0.107	1.261	[0.951, 1.672]
Gender	0.11	0.227	0.234	0.628	1.116	[0.715, 1.741]
Age	0.021	0.012	2.999	0.083	1.022	[0.997, 1.047]
Constant	-0.807	0.756	1.14	0.286	0.446	—