

ATHABASCA UNIVERSITY

A MULTI-AGENT FRAMEWORK TO SUPPORT USER-AWARE
CONVERSATIONAL AGENTS IN AN E-LEARNING ENVIRONMENT

BY

MICHAEL PROCTER

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE in INFORMATION SYSTEMS

ATHABASCA UNIVERSITY

August, 2017

© Michael Procter, 2017

Approval of Thesis

The undersigned certify that they have read the thesis entitled

**“A Multi-Agent Framework to Support User-Aware Conversational Agents
in an E-Learning Environment”**

Submitted by

Michael Procter

In partial fulfillment of the requirements for the degree of

Master of Science in Information Systems

The thesis examination committee certifies that the thesis
and the oral examination is approved

Co-Supervisor's:

Dr. Fuhua (Oscar) Lin
Faculty of Science and Technology
Athabasca University

Dr. Bob Heller
Faculty of Humanities and Social Sciences
Athabasca University

Committee Member:

Dr. Xiaokun Zhang
Faculty of Science and Technology
Athabasca University

External Examiner:

Dr. Shawn Fraser
Faculty of Health Disciplines
Athabasca University

August, 1, 2017

Acknowledgements

First and foremost, I'd like to thank my two supervisors:

Dr. Oscar Lin, for providing the incentive to push boundaries and explore different avenues, while somehow still helping me to remain focused. His experience, wisdom, and knowledge were essential to the completion of this thesis.

Dr. Bob Heller, whose generosity with his time and mentoring over many years of employment and collaboration inspired my decision to pursue my master's degree, and supplied the background and research skills to accomplish it. His encouragement has been invaluable.

I am grateful to my supervisors not only for their individual guidance and support but also for their combined efforts as a team, which was vital the success of this multidisciplinary thesis.

I would like to express my gratitude to the other members of my committee: Dr. Xiaokun Zhang for his valuable insight and feedback on the thesis, particularly on the system architecture; and Dr. Shawn Fraser for his meticulous and thoughtful observations and comments throughout. The final thesis has benefited significantly from his advice.

Thanks also go to Andrew Chiarella for allowing me access to a the participant pool, and Lorna Brown for providing the technical assistance to implement the survey during a very busy time of transition.

I'd very much like to express my appreciation to the many people who just make things work: Alice Tieulie in the Research Centre, Gail Leicht in Office of Research Services, and Krystal Zahara in the Faculty of Graduate Studies. I particularly want to acknowledge Linda Gray in the Faculty of Science and Technology for her responsiveness, and patience, in advising on all things to do with registration and courses.

Finally, I must thank my wife, Cheryl, for supporting me throughout this project, providing her own insights and advice on all aspects of the process, and putting up with far too many late nights.

Abstract

E-learning systems based on a conversational agent (CA) provide the basis of an intuitive, engaging interface for the student. The goal of this thesis is to propose an approach to improve the way that students interact with conversation-based e-learning applications. It attempts to do this through three contributions. The first is the design of an adaptable agent-based framework for improving interactions with conversation-based learning applications. The second contribution is to put forward a new approach to detecting user engagement based on real-time detection of conversational behaviour using the ongoing transcript of the interaction. The last contribution is to validate the approach by using volunteer students to test a proof-of-concept implementation of the framework. The observational and self-report data collected from the student testing provides new insights into how student interact with, and what their priorities are in evaluating, a pedagogical CA. This has implications for future development and research.

Table of Contents

Acknowledgements iii

Abstract v

List of Figures x

List of Tables xi

Chapter I - Introduction 1

 1.1 Research Background 1

 1.2 Research Issues 4

 1.2.1 What the CA detects about the student..... 5

 1.2.2 What the student perceives about the CA..... 6

 1.2.3 Integration with CA 8

 1.3 Research Objectives 9

 1.3.1 Design the agent-based framework 9

 1.3.2 Proof-of-concept implementation..... 10

 1.3.3 Develop agents to detect engagement..... 12

 1.3.4 Evaluate the system using student participants 12

 1.4 Contributions/Significance of Research 13

 1.4.1 Agent-based solution framework..... 13

 1.4.2 Conversation-based engagement detection..... 13

 1.4.3 Empirical results 14

 1.5 Organization of Thesis 14

Chapter II – Literature Review 15

 2.1 Conversational Agents 15

 2.1.1 Application to education and e-learning..... 15

 2.1.2 Task-oriented vs. narrative CAs 16

 2.2 Towards realistic CAs 16

 2.3 Detecting User State 19

 2.3.1 Affect detection 19

 2.3.2 Analyzing the conversational record 22

 2.4 Engagement..... 24

 2.4.1 Defining Engagement 24

2.4.2	Applications and impact of engagement.....	25
2.4.3	Detecting Engagement.....	26
2.4.4	Engagement-aware Responses.....	27
2.5	Context.....	27
2.6	Pedagogy Related to Conversational Engagement	29
Chapter III - Design Of Agent-Based Framework		30
3.1	Framework Objectives and Scope.....	30
3.2	Overall System Analysis and Design.....	31
3.2.1	System specification	31
3.2.2	System roles and agent assignments.....	37
3.3	System Architecture.....	37
3.3.1	Inter-agent message protocols	39
3.4	Detailed Design.....	43
3.4.1	CA-REP agent	44
3.4.2	ST-REP agent	46
3.4.3	ST-MODEL agent	47
3.4.4	Data source agents	48
Chapter IV – Analyzing Dialogue		50
4.1	Background.....	50
4.1.1	Conversational quality and appropriateness	50
4.1.2	Conversational behaviour	51
4.2	Development	54
4.3	Evaluation of Algorithms.....	56
4.3.1	Conversational quality and appropriateness classifier.....	56
4.3.2	User behaviour detection	60
Chapter V – Implementation.....		62
5.1	Overview of System Architecture.....	62
5.1.1	Student Representation (ST-REP)	62
5.1.2	CA Representation (CA-REP).....	63
5.1.3	Student Modeling (ST-MODEL).....	64
5.1.4	CA Modeling (CA-MODEL)	64
5.1.5	Student and CA Data Source Agents (DSA)	64
5.2	Proof of Concept Implementation.....	64
5.2.1	Scope of implementation	65

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

5.2.2	Goals of implementation.....	65
5.2.3	Core DSAs - Conversation text classifiers	66
5.2.4	External communication – connecting users and devices to the system	67
5.3	System Data and Execution	68
5.3.1	System data.....	68
5.3.2	System execution	68
5.4	Implementation Tasks	69
5.4.1	CA data set.....	70
5.4.2	CA communication and decision support.....	70
5.4.3	Student model plans.....	71
5.4.4	Integration strategies.....	72
5.5	Data Source Agents - DSAs.....	72
5.5.1	Constructing a DSA	72
5.5.2	Conversation-based DSAs	73
5.6	System Evaluation	78
5.6.1	Achievement of proof-of-concept goals	78
5.6.2	Performance.....	79
Chapter VI – User Testing and Discussion		81
6.1	Description of User Testing Methodology	81
6.1.1	Questionnaire.....	83
6.1.2	Chat logs	84
6.2	Analysis.....	84
6.2.1	Purpose and expected outcomes	84
6.2.2	Chatlog analysis.....	85
6.2.3	Survey data analysis	90
6.3	Findings.....	96
6.3.1	The efficacy of interventions	96
6.3.2	Factors associated with perceived usefulness.....	103
Chapter VII – Conclusions and Future Work		110
7.1	Discussion and Conclusions	110
7.2	Future Directions for Research and Development.....	113
7.2.1	Further analysis.....	113
7.2.2	Further research – Learning outcomes	114
7.2.3	Further development – Data source agents.....	114

References..... 116
Appendix A – Ethics Review Letter of Approval 133
Appendix B – Questionnaire..... 135

List of Figures

Figure 1: Goal overview 34

Figure 2: Analysis overview 35

Figure 3: Key to PDT diagrams 36

Figure 4: System roles and agent assignments 36

Figure 5: Agent role grouping overview 37

Figure 6: System overview 38

Figure 7: Initialization protocol 39

Figure 8: Initialization of data source agents 40

Figure 9: Protocol for establishing the student data set 41

Figure 10: Dissemination of updates to student model 42

Figure 11: Conversation protocol 43

Figure 12: Agent overview - CA-REP 45

Figure 13: Agent overview - ST-REP 47

Figure 14: Agent overview - ST-MODEL 48

Figure 15: Agent overview - ST-DATA 49

Figure 16: Excerpt from conversational log 52

Figure 17: System architecture 63

Figure 18: Remote user interface communication 67

Figure 19: Freudbot start page 81

Figure 20: Freudbot interface with End Conversation button 82

Figure 21: Conversation log processing 86

Figure 22: User experience frequency data 91

Figure 23: Social presence frequency data 93

Figure 24: Performance measure frequency data 95

List of Tables

Table 1. Response appropriateness confusion matrix..... 57

Table 2. Response appropriateness classifier performance 57

Table 3. Conversation quality confusion matrix..... 59

Table 4. Conversation quality classifier performance 59

Table 5: Behaviour algorithm testing 60

Table 6: Agent activity phases..... 69

Table 7: LIWC variables for chat log analysis 89

Table 8: ChatAgain/Recommend/Overall frequency tables 92

Table 9: Behaviour types and interventions 96

Table 10: Intervention frequencies 97

Table 11: Freud content and no-match counts before and after interventions 98

Table 12: LIWC social presence measures before/after intervention 2..... 99

Table 13: LIWC social presence measures before/after intervention 3..... 101

Table 14: Social presence ratings association to Chat Again 105

Table 15: Usage experience ratings association with Chat Again..... 106

Table 16: Conversation control ratings association with Chat Again 108

Table 17: Content and no-match measures association with Chat Again 109

Chapter I - Introduction

1.1 Research Background

Conversational agents (CAs) are designed to provide users with the ability to interact with computer software using natural language. In effect, the user is able to chat with an application to obtain information or carry out tasks, receive coaching, practice a language, learn a new skill, or simply converse for the purpose of social interaction or companionship. CAs may take the form of virtual guides (Yuan & Chee, 2005), characters in games and interactive stories (Endrass, Klimmt, Mehlmann, André, & Roth, 2014), social and learning companions (Castellano, Pereira, Leite, Paiva, & McOwan, 2009; Wong, Cavedon, Thangarajah, & Padgham, 2012). They are often embedded into web sites, video games, mobile devices such as “smart” phones, and even children’s toys. Recent advances in automatic speech recognition have resulted in a growing popularity of virtual assistants on mobile devices (SIRI and Google Now), and operating systems such as Cortana on Windows 10™ (Luger & Sellen, 2016).

In this thesis I distinguish between a CA and other natural language interfaces by the nature of the dialogue that takes place. Both the user and the software are expected to participate in an exchange that follows – or at least aspires to follow – the common rules and conventions associated with conversation, such as turn-taking, repair, and cooperation (Warren, 2006). In other words, conversing with a CA should be comparable to the experience of speaking with another human. By this measure, entering a question into a

search engine and receiving a list of links to resources would not be considered having a conversation, while asking a virtual assistant, such as SIRI, the same question might be, if the response is 'human-like'. While both interactions provide the convenience of using free-form natural language input, the motivation for using a CA includes a level of engagement associated with interacting with an intelligent being, albeit an artificial one.

CAs embedded within e-learning applications¹ have the potential to provide an intuitive, user-friendly interface that engages the student. Educational applications of CA technology include animated pedagogical agents (APA) (Heller & Procter, 2009; Johnson, Rickel, Lester, & others, 2000), intelligent tutoring systems (ITS) (D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008), and collaborative learning (Kumar & Rosé, 2011). CAs can play an important role in game-based learning systems (GBL) (Bellotti, Berta, De Gloria, & Lavagnino, 2011; Löckelt, 2011; McClure, Chang, & Lin, 2013).

The potential for education-related applications is particularly significant to distance education students as they can often be made remotely accessible via the Internet (Danforth, Procter, Chen, Johnson, & Heller, 2009; Heller & Procter, 2011), or deployed on home computers and mobile device (Perez-Marin & Pascual-Nieto, 2011) providing students with on demand access to one-on-one and collaborative e-learning resources, available 24x7. Students can also interact with embodied CA's participating in 3D virtual worlds (Grant, Sandeep, & Fuhua, 2013; Heller, Procter, & Rose, 2016).

An interesting role for CAs is to act the part of interviewee (Becker, Kopp, & Wachsmuth, 2007). For example, for medical students to practice their diagnostic skills

¹ The term CA will be used to refer to CA-based e-learning applications

against a virtual patient (Danforth et al., 2009). Heller & Procter (2011) have developed historical figure CAs which allow students to converse with a virtual Sigmund Freud or Jean Piaget. They refer to these as role-playing actor agents. It is this form of CA that will be used as a basis for the research in this thesis.

Conversational agents may employ natural language processing techniques to attempt to understand and process the user's input (J. Lester, Branting, & Mott, 2004). This may also be coupled with artificial intelligence (AI) techniques to reason about the meaning of the user's input, or at least categorize it, and determine an appropriate response (D'mello & Graesser, 2013; Olney et al., 2003). Other CAs, particularly those that might be labelled as "chatbots", may rely on simpler pattern matching methods, while still attempting to simulate intelligence (Kirakowski, O'Donnell, & Yiu, 2007; Wallace, 2009). CAs may be task-oriented, such as an ITS, where success is measured by the degree to which specific pedagogical goals are achieved. Alternatively, the CA may have a narrative or social approach, as in the case of an NPC or an historical figure (Löckelt, 2011). In the former case, conversation provides a convenient and engaging interface to achieving the task. In the latter case, the conversation is the task.

Regardless of the approach, CA's rely, to varying degrees, on the cooperation of the user to suspend disbelief so as to imagine that they are conversing with an intelligent entity or virtual character of some kind (Cassell, 2001; Veletsianos & Miller, 2008). This anthropomorphizing of the CA, by the user, is a key factor in maintaining engagement.

1.2 Research Issues

The overall purpose of the thesis research is to investigate ways to improve this interaction between students and e-learning CA's by concentrating on maintaining engagement.

The experience of interacting with a CA can be diminished when the agent falls short in simulating certain human behaviors (Becker et al., 2007; Callejas, López-Cózar, Ábalos, & Griol, 2011), resulting in a potential loss of engagement and reduced effectiveness as a learning tool. It is for this reason that much of the current research in this field is directed towards making CA's more realistic, life-like, and believable (Bogdanovych, Trescak, & Simoff, 2016; Cassell, 2001; Löckelt, 2011). There are a number of issues associated with achieving this goal of improving the CA. Several factors that can influence the perception of having a human-CA interaction that resembles that of a human-human one. Different approaches have been proposed to achieve these results: embodiment to give human-like characteristics and provide non-verbal cues through facial expressions and animated gestures (Cassell, Vilhjálmsón, & Bickmore, 2004; Johnson et al., 2000); speech input/output; modeling and expressing emotion and personality; detecting user emotion, and the ability to react appropriately to various affect states (Callejas et al., 2011; Kapoor & Picard, 2005); or an ability to recognize personality traits, and how they may affect the interaction (Mairesse, Walker, Mehl, & Moore, 2007).

In describing affective computing, Picard (1997) recognized two distinct capabilities of affect-aware software: the ability to express emotion, and ability to detect emotion. When this concept is applied more generally, we can say that approaches to improving how CAs are perceived come from two perspectives: what the CA can express

about itself – i.e. what the student perceives about the CA (e.g. embodiment, display of emotions), and what the CA can detect about the student (e.g. detecting user emotions and using the information in an intelligent manner).

1.2.1 What the CA detects about the student

Efforts to improve the ability of the CA to detect or predict user characteristics attempt to address some of the challenges in giving a CA comparable qualities to those of a human tutor, notably the ability to build trust and rapport, and providing a personalized service appropriate to the individual student's needs and preferences (Desmarais & Baker, 2011). This may involve building and maintaining some sort of model of the student within the CA. This requires the collection of data about the user, including information that changes over time, as in the case of user affect or behaviour. What can be perceived about the student is often provided by devices, such as cameras, eye-trackers, EEG sensors, and heart-rate monitors.

These devices can be obtrusive, or simply may not typically be available to student outside a laboratory setting. Availability of detection devices may also depend on whether the student is at home, attempting to learn while commuting, or in a public space like a library. Another consideration is that potential measurement devices may, over time, become available in affordable consumer versions. Examples include EEG devices such as the Emotiv headset (www.emotiv.com). As well, new devices are constantly being brought to market. Smart phones now contain information about the user's actions and movement that were unavailable a few years ago. 'Wearable' fitness tracking devices can now track heart rate. In theory this information could be made available to the CA. This suggests two related issues:

Issue 1a: A system that seeks to improve the interaction with the CA by collecting information about the student should ideally be able to adapt to whatever student model data is accessible due to availability of cameras, physiological sensors, and other measurement devices. This data would augment that which can be derived from the conversational record, which should always be available in the case of a CA.

Issue 1b: The system should be capable of adding new devices as they become available or practical to use by students (e.g. fitness trackers and other wearables) and processing the associated data.

The needs and goals of the CA will also dictate what type of information about the student is useful. For example, some CAs may be able to adjust their actions based on data about a range of student emotions, while others may only be able to use a subset of that data, or none of it at all. As well, the types of emotions and behaviours that may be expected to occur also varies depending on the nature of the interaction with the CA. A CA that tests the student, or provides a challenging task, may be more likely to see the student experience frustration or confusion. A CA that provides information through conversation, such as an historical figure, may not elicit such strong emotion.

Issue 2: The system should be adaptable to the capabilities and needs of the CA, based on what student information it can recognize, and how it can respond to it.

1.2.2 What the student perceives about the CA

The degree to which a CA is effective in a learning context relies in part on the perceived credibility of the CA. It seems reasonable to accept that perceived credibility

will affect the level of student engagement as well as the degree to which the CA will be viewed as a trusted authority on the information domain in question.

A number of things contribute to the credibility, and believability, of the CA (Bogdanovych et al., 2016). In the past 15 years, much of the research has focused on making the CA more realistic, or human-like, through embodiment, or the representation of the CA by an image or animated figure. The arguments for this are compelling. Human-human communication typically involves non-verbal cues in the form of facial expressions, eye gaze, head nods, shoulder shrugs, and other gestures. Despite this, attempts to confirm the link between learning performance and embodiment have produced mixed results (Dehn & Van Mulken, 2000; Veletsianos & Miller, 2008).

In this thesis I propose that when the CA plays the role of a tutor or expert on a subject, the student's perception of the CA's intelligence is particularly important to its credibility. Veletsianos & Russell (2013) believe that if a pedagogical agent is simulating a human expert, as a source of information or guidance, then the student will have similar expectations as for a human expert. They cite Baylor & Kim (2005) and their belief that student expect a high level of accuracy from agents acting as content experts. Naturally, for a *conversational* agent in particular, the quality of the dialogue is an important factor in the perception of intelligence, as well as social presence (Heller, 2016). This requires following rules of conversation, such as turn-taking, error detection and repair.

Research by Reeves & Nass (1996) showed that users will interact with media in a similar way to how they would with a human, suggesting a natural inclination to anthropomorphise the agent. The illusion of conversing with an intelligent entity can be somewhat fragile and can be disturbed by repetition, obvious mistakes due to pattern

matching, or other things that reveal a pre-programmed response. Therefore maintaining this illusion is an important goal in designing the CA. Although the user is willing to make some concessions, if pushed too far beyond a reasonable suspension of disbelief the perception can switch to a negative one, which can be difficult to reverse (Luger & Sellen, 2016). Therefore, protecting this illusion is an important design goal.

The quality of the dialogue is enhanced by the ability to detect if the student is engaged and cooperating in the conversation. It also requires handling off topic comments or questions, bad grammar, typos, and other situations where the CA cannot understand the user gracefully. Löckelt (2011) talks about the frustration a user can experience when expectations are set by the CA's generated text, but not met by its ability to understand user input, regardless of where the fault lies. When the CA does not understand the student, it is important to maintain the illusion of conversing with an intelligent being, and minimize the disruption to the learning experience.

Issue 3a: A system that seeks to improve the interaction by focusing on how the student perceives the CA needs to monitor and maintain the quality of the conversation with a focus on maintaining credibility and engagement.

Issue 3b: Ideally be capable of supporting a variety of dimensions, adaptable to what is appropriate to the nature of the CA. These include quality of dialogue as well as characteristics of embodiment, such as animation, modeling CA affect and personality.

1.2.3 Integration with CA

In the simplest case, where one is developing a new CA, the concepts discussed so far can be taken into consideration when designing the software. However, there are

several challenges to be overcome if enhancing an existing CA. Modifying a CA extensively may not be practical depending on its design, or the availability and complexity of the source code if the original developer is no longer available. In some cases, a solution that requires little or no changes to the existing CA reduces the risk of introducing unexpected side effects in the software.

Issue 4: It may not be possible to modify the CA or modification may be limited. A solution should allow for integration with a variety of CAs and should not rely on the ability to alter the CA.

1.3 Research Objectives

The following objectives were motivated by the research issues described in section 1.2. The objectives are listed here and described in greater detail following:

1. Design an agent-based framework that interacts between the user and the CA.
2. Implement a proof-of-concept system based on the framework
3. Design and implement agents to detect user engagement based on analysis of the conversation
4. Collect and analyze interaction and self-report data from students who use the proof-of-concept system.

1.3.1 Design the agent-based framework

In order to address the dynamic nature of modeling both the student and the CA, as stated in issues 1a and 1b, a system based on autonomous intelligent agents is proposed. Agents, by their nature, are designed to adapt to changes in the environment. Each source

of student data and CA behavior are to be represented by an agent, providing information to central agents responsible for maintaining a model of the student or the CA.

An additional feature of the agent-based framework is that it supports both approaches to improving the experience of interacting with a CA, aiming to both enhance what the CA can express to the student, as well as what the CA can understand about the student. Additional agents provide an interface with both the student and the CA, supporting the modification of the CA's behaviour based on student and CA data, and providing the visible elements for the student's perception of the CA.

These interface, or "representation", agents also afford a means for the agent framework to integrate with the existing CA with minimal modification to its software. This is intended to resolve issue 4, or the varying degree to which different CAs can be modified. By implementing the solution "externally" from the CA itself, the decision-making logic and modelling capabilities can be developed in the agents, with few changes to the existing software. Change to the dialogue with the user can be made in the CA or, if necessary, in the representation agent itself. It is possible to achieve the integration without making any changes to the CA, though this requires more processing on the part of the representative agent.

The details of the design requirements for the agent-based framework are provided in Chapter 3. Chapter 5 describes the different integration options.

1.3.2 Proof-of-concept implementation

A proof-of-concept implementation of the framework was developed and tested. This serves several purposes:

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- Demonstrate the process of integrating an existing CA to the agent-based framework
- Provide a platform to test and evaluate the performance of an implementation
- Provide a test platform to collect live data from students interacting with the system
- Demonstrate the ability to add or add intelligence and change functionality of the CA using the representation agent

The motivation for this objective is to test the ability of the proposed agent-based approach to address Issue 4 the ability to integrate with a CA that was developed separately from the framework. An existing conversation agent known as Freudbot (Heller, Proctor, Mah, Jewell, & Cheung, 2005) was used as the basis of the implementation of the system. Details of the implantation of the proof of concept system, and how it was integrated with Freudbot, are provided in Chapter 5.

To demonstrate how the agent approach can be used to add intelligence or modify the behaviour of the CA with minimal changes to the CA, an important strategy was modified using the representation agent. The process of selecting the most appropriate response when the CA does cannot match the user input to a known pattern was changed from a random selection to an escalating model based on the number of “misses” that had occurred, and other conditions. This is described in Chapter 5.

Two agents were developed to provide student data, described in 1.3.3. A third agent was created to demonstrate the process of collecting data from a device worn by the student and the communication protocols to support this.

1.3.3 Develop agents to detect engagement

As part of the proof-of-concept implementation, two agents were developed to provide data about the student. These have a special role in that they are designed to always be available, providing a default set of measures for the system. The goal is that they do not depend on any special measuring device, relying instead on a real-time analysis of the conversational record. Conversation is the common component of all CAs and the log of the conversation is always available since the user-CA dialogue is transmitted through the agents. The agents which analyze this dialogue focus on the quality of, and patterns in, the student's contribution to the conversation. These agents support interventions by the CA, consistent with the concerns described in Issue 3a.

These agents are designed to provide a measure of the student's engagement in the activity of conversing with the CA. Engagement has been identified as an important component to learning (Szafir & Mutlu, 2012) and as such would be a measure that would be of interest to a large spectrum of CA types. An underlying element of these agents, is the development of a new approach to estimating engagement based on conversational behaviour. This approach is described in Chapter 4.

This demonstrated the general process of developing the agents which represent student data sources, based on template agents designed from the framework. These agents also provided a basis for testing the use of the system with volunteers.

1.3.4 Evaluate the system using student participants

To validate the framework, the proof-of-concept implementation was used by a group of student volunteers. Students chatted with the Freudbot CA via the agent-based system for at least 10 minutes. Objective measures were obtained through an analysis of

the conversational and agent action logs. Participant data was collected from a questionnaire filled out by the student volunteers immediately after the chat session.

The information gathered from this experiment can be used both to evaluate the system and to provide data for future development and tuning of the agents and process. Subjective data provides feedback on how well student perception of the experience matches the expectations of the researcher. They also provide some insight into the types of students who viewed the exchange positively, or not, and their associated behaviour when interacting with the CA. The methodology and results are described in Chapter 6.

1.4 Contributions/Significance of Research

1.4.1 Agent-based solution framework

One goal of this research is to develop a system based on autonomous intelligent agents that is capable of adapting to the individual needs and capabilities of a CA-based learning application, the learning goals of the student, and the user affect, personality, and context information that is available. An important outcome will be the development of a framework for a multi-agent system based on intelligent, autonomous agents to model both the CA and the user. Additional outcomes include an implementation which demonstrates the utility of the framework, providing template agents that implement the underlying communication and execution protocols. These can serve as an example, or can be adapted to work with other CAs and different educational contexts.

1.4.2 Conversation-based engagement detection

New text-analysis techniques were proposed and developed for estimating user engagement based on the quality and nature of the conversation by the student by analyzing

the conversational log. Agents were developed, based on and compatible with the framework, to implement these approaches.

1.4.3 Empirical results

A second goal of the research is to investigate the ability of this system to enhance the interaction between the user and the CA. A study has been designed to test the effectiveness of the system on the students' interactions with a conversational agent. It is expected that the data collected from student interactions will contribute some insight into what strategies help to improve students' perception of a CA as a learning tool.

1.5 Organization of Thesis

The remainder of the thesis is organized as follows. Chapter 2 provides a review of the literature on conversational agents, detecting affect and other user characteristics, the importance of engagement, and multi-agent systems. Chapter 3 describes the design of the agent framework, the overall system architecture, detailed design, agent capabilities and communication protocols. Chapter 4 introduces the concepts behind the text analysis implemented for the thesis research, its development and evaluation. Chapter 5 goes on to describe a “proof of concept” implementation and an evaluation of the framework and the implementation. Chapter 6 details the methodology for testing of the implementation by volunteers, the data collected, ending with results of a statistical analysis of survey and conversational log data. Finally, Chapter 7 summarizes the conclusions of the thesis and describes directions for future research.

Chapter II – Literature Review

2.1 Conversational Agents

Conversational agents are designed to allow a natural language interaction with users with the intention of providing an engaging experience that mimics that of speaking with another human. This technology has been employed in a wide range of applications. Nunamaker describes an automated interviewer for border security which can detect potential deception (Nunamaker Jr., Derrick, Elkins, Burgoon, & Patton, 2011). Smith et al. (2007) uses intelligent CAs as "bit part" players in a role-play e-drama where other actors are human driven avatars. The objective behind Bellotti et al. (2011) is to build an NPC for a serious gaming (SG) environment that can provide information to participants, while being easy to maintain.

2.1.1 Application to education and e-learning

The application of CAs in an educational context developed alongside the evolution of early intelligent tutor systems to use virtual characters as their interface (Veletsianos & Russell, 2014). The research headed by Graesser using AutoTutor (Rus, D'Mello, Hu, & Graesser, 2013) is a prime example of an ITS with a dialogue-based interface, with two decades of on-going development reported in the literature. CAs can be used in Animated Pedagogical Agents, a term Johnson, Rickel and Lester (2000) created to describe the combination of ITS with animated user interfaces. Perez-Marin & Pascual-Nieto (2011) believes that we are headed toward a future in which ubiquitous and pervasive

pedagogical conversational agents will support students in reviewing their studies on multiple platforms.’

Conversational agents plays an important role in the use of virtual characters in e-learning environments.

An interesting application of conversational agents is that of interviewee. (Becker et al., 2007). For example, allowing medical students to practise their diagnostic skills against a virtual patient (Danforth et al., 2009). Heller & Procter (2011) have developed historical figure CAs which allow students to converse with a virtual Sigmund Freud or Jean Piaget using an interview-based dialogue.

2.1.2 Task-oriented vs. narrative CAs

CAs may be task-oriented, such as an ITS (Graesser, Conley, & Olney, 2012) or social-oriented (Veletsianos & Russell, 2013), as in the case of an NPC or a historical figure (Heller & Procter, 2011). (Löckelt, 2011) uses the term ‘narrative’ to refer to social-oriented CAs, pointing out one of the key differences is that task-oriented systems do not typically rely on an immersive effect to the degree that narrative ones do. Wong et al. (2012) describes some of the challenges in designing an interactive toy that handles both task-oriented interactions as well as more "chatty" conversations.

2.2 **Towards realistic CAs**

Much of the literature focuses on how to design conversational agents to be realistic, or believable. Typically this means that the CA behaves in a human-like way. Lester and Stone define believability as “the extent to which users interacting with an agent

come to believe that they are observing a sentient being with its own beliefs, desires, and personality” (Lester & Stone, 1997, p17).

The motivation for doing this includes increasing engagement, and improving task performance. The two are often related. For example, a user who is engaged in the interaction may feel more motivated to carry out the task. However it is also possible for engaging aspects of the CA to distract from the task (Yee, Bailenson, & Rickertsen, 2007). Lester & Stone (1997) developed a competition-based approach to balancing an APA’s “believability enhancing” actions with pedagogical sequences. This allowed them to ensure that behaviours associated with making the agent life-like did not distract from problem solving tasks.

What does it mean to be more “human-like” exactly? It can refer to perceived intelligence, an ability to understand the user’s emotions and/or express emotions, or having a personality. These goals are often achieved with the aid of some sort of visual representation of the CA. This may be a static image, an animated figure or head, or even a 3D avatar situated in a virtual world. However, some research is directed to improving the conversation itself (Battaglino & Bickmore, 2015; Graesser, Li, & Forsyth, 2014).

Norman (1994) notes the tendency for users of virtual characters to anthropomorphize, attributing human-like traits to software that displays or simulates some form of intelligence, and the resulting disappointment when the application is unable to accomplish tasks at a human level of performance. This is consistent with our own experience with historical figure agents such as Freudbot (Heller & Procter, 2011), when users attempt to ask sophisticated questions, encouraged by what appears to be an artificial intelligence. Löckelt (2011) found a similar effect can occur with expectations set by the

level of realism of the rendered character where the user expects better conversational performance from a virtual character that is a realistic rendering of a human.

Löckelt (2011) identifies the importance of modelling personality traits in agents, basing the agent's behaviour on a dynamically updated model of its affective state. He describes the OCEAN model (openness, conscientiousness, extroversion, agreeableness, neuroticism) for parameterizing affect change resulting from interaction. Callejas et al. (2011) summarizes literature by "trait theorists", describing various proposed dimensions of personality ranging from 2 to 16 dimensions. The Five Factor Model (also Big Five, or OCEAN model) is the most widely used and "has become a standard in psychology".

Nunamaker et al. (2011) studied the effectiveness of dynamic vs static expressions in embodied conversational agents (ECA) and found that nods or head shakes were preferred by users over the same image that remained static. Even a disembodied voice was preferred over photorealistic heads that are not "lifelike". This is consistent with a study by Heller & Procter (2009) comparing user preferences for a static image, an animated head with static expressions, and no image when speaking with a conversational agent that represents Sigmund Freud. Part of the explanation for these observations may be attributed to the "uncanny valley" effect when an embodied conversational agent is very realistic but does not fully simulate human behaviour, as described in Callejas et al. (2011) and Nunamaker et al. (2011).

The importance of emotional awareness in human-computer interaction has prompted a great deal of research activity in recent years. Callejas et al. (2011) believes that emotions provide personality, leading to user's adopting a better attitude to the agent. They describe the similarity-attraction principle, which proposes that people prefer to

interact with a CA that has a similar personality to their own. Picard (1997) describes affective computing applications in terms of two abilities: the ability to express emotion, and the ability to detect emotion. Nunamaker et al. (2011) uses the expression of emotion by the agent to increase believability. Becker et al. (2007) also concentrates on the internal emotional state of the agent as opposed to the emotional state of the user. In one experiment conducted by Becker, the agent is playing a card game against a user. GSR and EMG sensors are used to measure the user's response to the agent's expressed emotions (this is not fed back to the agent). Negative empathic behavior by the agent was found to illicit "negatively valenced emotions" in the user.

Affect-aware agents are capable of detecting the user's emotions. Nunamaker et al. (2011) describes an agent that is able to detect increased stress levels using a single sensor (vocal signal) when the user attempts to provide deceptive response to interview questions. D'Mello et al. (2008) discusses how AutoTutor can be modified to use affect detection in decision making. "This adaptation would increase the bandwidth of communication and allow AutoTutor to respond at a more sophisticated metacognitive level". Afzal & Robinson (2011) propose that the strong role that affect plays in human teacher-learner activities suggests the importance of "affective diagnosis" in computer-based learning.

2.3 Detecting User State

2.3.1 Affect detection

Research on emotion stretches back over a century and today crosses multiple disciplines, including philosophy, cognitive and social psychology, anthropology and neuroscience (Calvo & D'Mello, 2010). Emotion can be classified using discrete

emotional categories, or dimensionally (Zeng, Pantic, Roisman, & Huang, 2009). The most popular example of emotional categories is the six basic emotions described by Ekman (Ekman & Rosenberg, 1997), which includes happiness, sadness, fear, anger, disgust, and surprise and is the basis of FACS, the Facial Action Coding System developed by Ekman. Classifying emotion dimensionally refers to expressing emotion as values along a series of scales (dimensions), two of the most important being evaluation (positive/negative) and activation (likelihood of taking action active/passive as result of emotion). Zeng describes a third appraisal-based approach as "one of the most influential" in modern psychology, similar to and an extension of dimensional, but difficult to program. Appraisal theory is based on understanding the significance an individual places on a situation, object, or event. Emotions are reaction to appraisal of situation or event and how it affects the person - several dimensions of appraisal are proposed: beneficial/harmful, probable/improbable, agency (caused by oneself/someone else), reward/punishment, control/no control (Calvo & D'Mello, 2010).

While the majority of affect detection research focuses on Ekman's basic emotions, some emotions can be considered to be more relevant to learning-related activities. Baker, D'Mello, Rodrigo, & Graesser (2010) compares Ekman's 6 basic emotions (fear, anger, happiness, sadness, surprise, disgust) to those more relevant to learning: boredom, confusion, delight, flow (involvement), frustration, surprise. Kapoor & Picard (2005) focused on detecting interest and disinterest in children. Pekrun, Goetz, Frenzel, Barchfeld, & Perry (2011) identify enjoyment, hope, pride, relief, anger, anxiety, shame, hopelessness, and boredom as "critically important for students' motivation, learning, performance, identity development, and health."

Detecting user affect relies on physiological, behavioral and psychological approaches (Zimmermann, Guttormsen, Danuser, & Gomez, 2003). The question arises as to what modalities, and how many, are best suited for affect detection. Kleinsmith, Bianchi-Berthouze, & Steed (2011) seeks the "minimal information necessary for automatic affective posture recognition" in a game scenario and studies whether the basic information recorded by the game controller (for something like Wii or Kinect) is sufficient. D'Mello et al., (2008) believe the majority of affect detection systems are based on facial expression, vocal expression, and to a lesser extent, posture patterns. (Nunamaker Jr. et al., 2011) implement an interviewer for border security which can detect potential deception based on various sensors including video cameras, near infra-red, thermal cam, eye tracker, vocal pitch, and laser-dopler vibrometer). However the results of one of their studies showed that a single sensor (voice) was capable of detecting stress levels in the user when they were given the task of attempting to deceive the ECA. Kleinsmith et al. (2011) cites a number of references that suggest emotions can be reliably detected with some simple sets of data, citing an example where four basic emotions could be recognized by measuring the distance between body joints. Mao & Li (2009) use facial expressions, speech characteristics, and analysis of text. They discuss the importance of choosing an appropriate integration level, or "fusion technique", as a next step in their research. A sensory-level technique combines data from different sources (e.g. facial expression and speech) before making an affect classification. A decision-level technique would determine affect from each source first, and then make a final decision on the classification based on combining each of these in some way.

2.3.2 Analyzing the conversational record

The majority of affect detection methods rely on sensors and devices which are intrusive, expensive and may not be found outside a lab setting. There are benefits to exploring techniques that make use of text and linguistic features. These are readily available to a CA application and non-intrusive, and as such they are more suitable to detecting affect of students in learning situations. As D'Mello et al. (2008) state, the advantage of affect detection from discourse is that discourse is abundant and "inexpensive to collect".

Balahur, Hermida, & Montoyo (2012); Callejas et al., (2011); and Calvo & D'Mello (2010) all provide good overviews of different text-based AD methods. The user's text can be analyzed at several levels. A lexical approach attempts to identify words that have affective meaning associated with them. There are a number of software packages available to aid in this type of analysis. The Linguistic Inquiry and Word Count (LIWC) tool (Tausczik & Pennebaker, 2010) which attempts to predict the emotional content of a body of text based on the frequency of words that it classifies as positive or negative emotions (Kahn, Tobin, Massey, & Anderson, 2007). Liu, Lieberman, & Selker (2003) analyzes text in an email writing application using lexical approaches such as word spotting based on Ortony's Affective Lexicon (Ortony, Clore, & Collins, 1988), and lexical affinity to extract affect related knowledge from the Open Mind Commonsense (OMCS) knowledge-base (Singh et al., 2002).

Semantic analysis of the text goes a step further than matching words and looks at the meaning associated with text using techniques such as Latent Semantic Analysis (LSA) to compare how well text matches corpora containing emotional phrases. One study found

this method has shown some promise in detecting fear and joy but has had less success with other emotions (Calvo & D'Mello (2010). Liu et al. (2003) points out that semantic approaches such as LSA do not work well at the sentence level and are more useful working at a paragraph level.

Sentiment analysis is a more recently developed approach that is gaining acceptance in the field of natural language processing and affective computing (Calvo & D'Mello, 2010; Balahur et al., 2012). Rather than assigning specific emotions to text, the focus is on evaluating the overall affective valence or polarity (positive or negative). This is done based on models constructed from large knowledge-bases of real-world experience.

Emotional state of the user can also be detected or inferred without analyzing the content of the user's text. D'Mello et al. (2008) used discourse features associated with student interactions with the AutoTutor ITS to detect boredom, confusion, delight, flow, frustration, surprise. Discourse variables exemplified include speed of reply, length of response, last feedback from ITS, and appropriateness of response. Epp, Lippold, & Mandryk (2011) proposes the use of "keystroke dynamics", building on work using typing patterns as an authentication mechanism. Metrics include time between key presses, and length of time keys are pressed. Zimmermann (2003) proposed the use of keyboard and mouse activity to infer user affect. Jaques & Vicari (2007) use the OCC cognitive model of emotion to infer the student's state of mind by an appraisal of their actions and events surrounding them, such as failing to achieve a goal.

2.4 Engagement

2.4.1 Defining Engagement

Even among those researchers who agree that the success of an application depends on engaging the user, there appears to be no agreed upon definition of the term “user engagement”. Novielli (2010) describes engagement as a fuzzy concept, noting several different definitions from the literature. Boyle, Connolly, Hainey, & Boyle (2012) observe that while entertainment games appear to be engaging in nature, as evidenced by the amount of time users spend on that activity, explaining why remains difficult. Her review suggests that formal research is only starting to shift focus from usability issues to understanding nature of enjoyment.

The study carried out by O’Brien & Toms (2008) was dedicated to the purpose of providing an operational definition and attempts to supply a definitive list of attributes associated with engagement, broken down over the four stages in their model: point of engagement, sustained engagement, disengagement, and reengagement. They proposed this definition based on a survey of past and current research: “Engagement is a category of user experience characterized by attributes of challenge, positive affect, endurance, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control.” (O’Brien & Toms, 2008)

Ultimately, the difficulty in agreeing upon a definition may be due to the domain specific nature of the concept and a tendency for research to be focused on specific application areas. Nakano and Ishii (Nakano & Ishii, 2010) talk about conversational engagement as “the process by which two (or more) participants establish, maintain and

end their perceived connection”. Xu, Li, & Wang (2013) also define engagement in terms of how involved users are in a conversation.

Some researchers equate engagement to attention (Asteriadis, Karpouzis, & Kollias, 2009; Szafir & Mutlu, 2012), while O’Brien & Toms (2008) believe that attention is only one of many attributes associated with engagement. Sundar, Bellur, Oh, Xu, & Jia (2014) believe that an understanding of engagement must study both actions and subjective experience.

2.4.2 Applications and impact of engagement

The importance of user engagement and attention influences the application interface across many domains. Szafir & Mutlu (2012) are concerned with computer-based education (CBE) and the ability of an embodied virtual tutor to maintain student interest and attention. Students with learning difficulties can be monitored for engagement level while interacting with a computer (Asteriadis et al., 2009). User engagement is taken as a measure of social interaction for a game-playing robot companion in (Castellano et al., 2009). Advice-giving agents and systems, such as the one proposed by Novielli (2010), rely on a measure of engagement and social attitude to select appropriate strategies for persuading the user. Virtual assistants evaluate the buyer’s engagement during product descriptions and use this to determine level of interest (Nakano & Ishii, 2010).

Any of these scenarios may benefit from an engaging user interface, whether it is to ensure that the user is paying attention to improve the experience, or a matter of evaluating the level of engagement to adapt the interface to reengage the user. Questions arise as to what defines engagement and does it vary for each application domain

2.4.3 Detecting Engagement

Detection or estimation of levels of user engagement makes up the first of two components associated with building engagement-aware application interfaces (Nakano & Ishii, 2010; Xu et al., 2013). Techniques for detecting and identifying user engagement typically require some method of sensing various verbal and non-verbal behaviour cues (Szafir & Mutlu, 2012) as well as contextual information (Castellano et al., 2009). Eye-tracking headsets (Nakano & Ishii, 2010) are a common method for evaluating where the user's attention is focused. Electroencephalography (EEG) headsets that can measure brain activity associated with engagement are becoming more prevalent, though researchers continue to search for ways to collect this data in a non-intrusive way (Asteriadis et al., 2009), a requirement for making engagement-aware applications a practical reality. Stereo cameras can detect body positioning and movement as well as expressive gestures (Xu et al., 2013).

Perhaps one of the least intrusive approaches to evaluating engagement is the analysis of conversation in dialogue based systems. This is not an area that has been explored extensively in the literature, despite the substantial amount of research associated with text-based affect detection. Wen, Yang, & Rose (2014) describe a technique for measuring cognitive engagement based on Turney's level of word abstraction dictionary (Turney, Neuman, Assaf, & Cohen, 2011) to distinguish between forum posts which are more descriptive and those that are more interpretive. The 2015 version of LIWC provide a summary output variable 'Analytic' which appears to achieve a similar measure, based on research reported in Pennebaker, Chung, Frazee, Lavergne, & Beaver (2014).

This thesis attempts to break some new ground by proposing methods for detecting user engagement based on their contribution to the conversation. This is done using both a real-time analysis by the agent system, described in Chapter 4, and by an offline examination of the conversational record of student volunteers participating in a study, described in Chapter 6.

2.4.4 Engagement-aware Responses

The second component of engagement aware systems involves the modification of the behaviour of the application to maintain, increase, or to re-establish user engagement. An engagement-aware application makes use of user information to maintain engagement if it has been detected, trigger actions to re-engage the user if it is lost, or modify the interaction strategy to meet the personal needs or preferences of the user. Applications that attempt to mimic a human-human interaction experience may use this information to choose the correct behaviour (Szafir & Mutlu, 2012).

As expected, the way in which user engagement data is used is dependent on the application domain, the goals of the system, and role that user engagement plays in the success of the interaction. (Novielli, 2010) used engagement information to select the most persuasive advice-giving strategy. Engagement has been used to select appropriate strategies for reengaging inattentive customers (Nakano & Ishii, 2010) and students (Szafir & Mutlu, 2012). Engagement plays an important role in helping an embedded agent to be perceived as more human-like (Castellano et al., 2009; Xu et al., 2013).

2.5 Context

Detecting and understanding user states, such as affect and engagement information, is dependent to a degree on the context in which it is experienced. This is one

of the motivations for the approach adopted in this thesis. The system can adapt to the computing and environmental context of the student by activating the agents associated with available equipment. For example, an agent that detects facial expressions joins the system if a webcam is available.

Picard (1997) recognized the value of using detection of context (where, when, conditions, situation) to determine situation may be stressful, relaxing, etc. Vildjiounaite et al. (2009) also emphasizes the importance of context - expression of the same emotion different in different situations, interacting with different people. Epp, Lippold, & Mandryk (2011) identifies lack of user's context ("such as their location, expertise, or emotional state") as an underlying problem with interactive applications. In Kapoor & Picard (2005), information about the task being carried out is treated as one of the modalities of several affect detection measures combined to classify the user's emotion. D'Mello et al. (2008) acknowledges that a shortcoming in their 2008 study is that context is based on a single dialogue turn and suggests "Perhaps classification accuracies could be boosted by incorporating a broader scope of contextual information, including patterns of conversation that evolve over a series of turns leading up to an emotional experience." Liu et al. (2003) acknowledges the need to address the issue of not taking context into account preferring the term "affect understanding" as opposed to "affect sensing". Feidakis, Daradoumis, & Caballe (2011) proposes the use of social emotional learning (SEL) theory to address the problem that different students react differently to certain emotional state with respect to learning activities. They provide the example of how confusion can motivate some students to work harder to remove the confusion by understanding the subject better, and some can become frustrated and give up.

2.6 Pedagogy Related to Conversational Engagement

Yamashita, Kubota, and Nishida (2005) believe that comprehension is facilitated by the delivery of new information in a conversational form, when compared to simply reading the same information or hearing a monologue. "More specifically, our comprehension of a topic can be deepened if we ask questions and discuss the topic with others." (Yamashita et al., 2005, p. 126) One of the benefits of using the narrative approach and conversation is that the student is encouraged to stop, after receiving a short paragraph of information, and consider a conversational response. This may take the form of a question, which requires a level of comprehension. It would be similar to stopping reading after each chapter and thinking of a relevant question associated with the information. Although the CA may not know the answer, the student still goes through the exercise of thinking of the question. (In fact, if the CA does not understand the question, it may prompt the student to rephrase it, gaining potential benefit). This suggests the usefulness of encouraging conversational engagement.

Veletsianos & Russell, (2013) feel that social discourse is important to the effectiveness of pedagogical agents. They point out that learning is intended to be collaborative experience and that pedagogical agents need to add this social dimension to meet those goals. "The focus on task-oriented agents in the literature is in contrast to the vision of participatory, student-centered, and community-oriented learning experiences" (Veletsianos & Russell, 2013, p. 382).

Chapter III - Design Of Agent-Based Framework

3.1 Framework Objectives and Scope

A system that improves the interaction between a student and a CA should ideally be able to adapt to whatever student model data is accessible due to availability of cameras, physiological sensors, and other measurement devices. It should also be able to adapt to the capabilities of the CA, whatever student information it can recognize, and how it can respond to it.

The proposed approach is to address the dynamic nature of modeling both the student and the CA by using autonomous intelligent agents, which, by their nature, are designed to adapt to changes in the environment (Gonzalez-Sanchez, Chavez-Echeagaray, Atkinson, & Burleson, 2011). Each source of student data and CA behavior are to be represented by an agent, providing information to a central agent responsible for maintaining a model of the student and the CA.

The general goal is to improve a student's interaction with a CA by providing the CA with information about the student, and allowing the CA to improve the way it is perceived by the student. The intention of the framework is to define an agent-based approach to create this functionality. The purpose of the framework is to define the necessary roles and describe how agents will fulfill those roles. It also identifies the underlying communication and process protocols required to allow the agents to work together to achieve the overall goals. Template agents can be developed from this framework to execute these underlying protocols. These agents can be extended or

modified to meet the requirements of a specific CA and its associated learning objectives, as well as the goals for modeling the student, and integrating the relevant sources of data.

The framework does not provide a specification for some important functions. These include: the user interface and multiple session control, the interface to the CA, and specifics of using agents distributed across different systems. These functions were judged to be specific to the e-learning application and the choice of agent platform, and therefore best defined at implementation time. Examples of how each of these functions were realised are described in the proof of concept implementation in Chapter 5.

3.2 Overall System Analysis and Design

The system and architectural design were developed using the Prometheus agent design methodology and its associated design tool, PDT (Padgham & Winikoff, 2005).

3.2.1 System specification

The requirements of the system are described as follows:

- The system supports and enhances the interaction between a student and a CA-based learning system.
- The primary function is the communication between the student and the CA. This includes an interface for the student that accepts user input and displays CA responses, and a connection to the CA using whatever protocol is appropriate (e.g. HTTP is used in our case).
- Additional functions include the collection of user context data stored as a model of the student. This includes dynamic information such as affect, short-term

goals, and task performance, as well as semi-static information such as learning preferences, personality, long-term learning goals, and past performance.

- Similarly, the CA is modeled in terms of such parameters as teaching goals, personality, and emotional state.
- A CA administrator (or researcher) should be able to view and configure the model of the CA.
- A teacher (or researcher) should be able to modify the content of the learning system/CA.
- The behaviour of the CA should be based in part on the individual characteristics of the student. Those characteristics may vary between students, and change over time. What characteristics may be relevant to the CA may vary between CAs. What aspects of the CA's behaviour that can be modified based on student characteristics may vary between CAs.
- The system should be able to adapt to the addition or deletion of different student characteristic measures. The system should support the integration of available measures where appropriate to form a higher level description of the student's state.
- Information about the student's state is passed to the CA to allow it to make decisions and modify its behaviour based on student data.
- During an initial phase, available student characteristics are identified and the model is started. Student model information is sent to the CA. Negotiation regarding what information will be transmitted to the CA is carried out. Static information about the student is passed to the CA.

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- During the communication phase, user text is passed to the CA, and dynamic information about student state is passed to the CA. The CA may provide feedback regarding the observed behaviour of the student with respect to consistency with the model. For example, whether or not the student's choices are consistent with their learning goals.
- During the communication phase, the CA will modify its strategies to align with, or respond to, static information from the student model as well as dynamic state data.
- During either the initial or post-communication phase, a teacher (or researcher) should be able to monitor or call up reports about a student's progress and view details of their student model.

Based on the system description, the major goals were identified, and sub-goals defined to achieve them. Figure 1 shows the goal hierarchy that was developed.

The following actors (people, devices or software that interact with the system) were identified:

- Student – interacts with the system, converses with the CA, and provides personal data requested to support the student model (e.g. learning preferences and goals).
- CA (e.g. Freudbot) – converses with the student, via the system, using natural language on some subject domain.
- Instructor – monitors student performance
- CA Administrator – interacts with CA for the purpose of configuring and performance monitoring

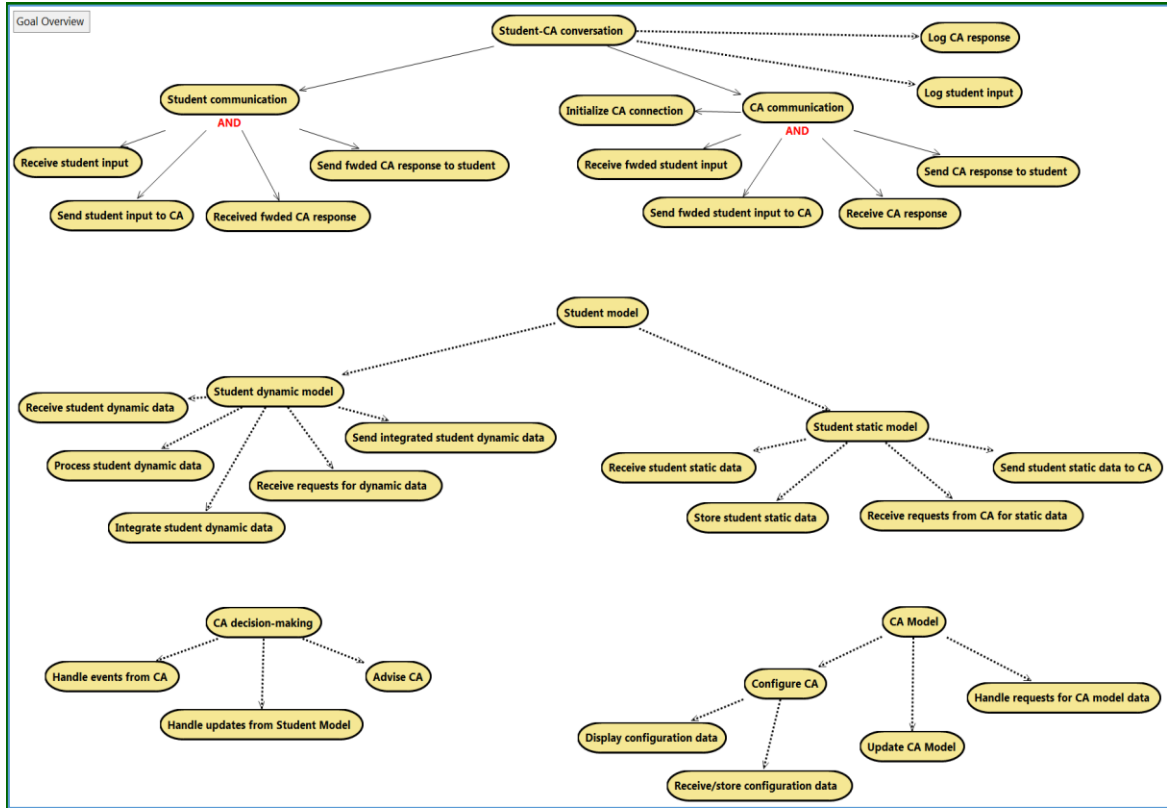


Figure 1: Goal overview

3.2.1.1 Analysis Overview

The Analysis Overview (Figure 2) is designed to show the interactions between the system and the environment. Once the actors were identified, the scenarios, percepts and actions involved in the system were added.

The scenarios anticipated are:

- Student Communication – the student interaction with the system, primarily conversing with the CA
- CA Communication – the interface to the (external) CA allowing it to receive student input and provide responses

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- Student Context Sensors – Data about the student is collected and processed.
- Student Model – all information about the student, static or dynamic, is maintained in one place
- CA Decision-making – the system aids the CA in how it interacts with the student taking into account the student and CA requirements
- CA Model – the configuration of the CA as well as any dynamic modeling of personality or emotion
- Monitor CA Performance – handles requests for performance data related to the CA
- Student Performance – handles requests for performance data related to the Student

Figure 3 provides the key to the icons used in this and other diagrams.

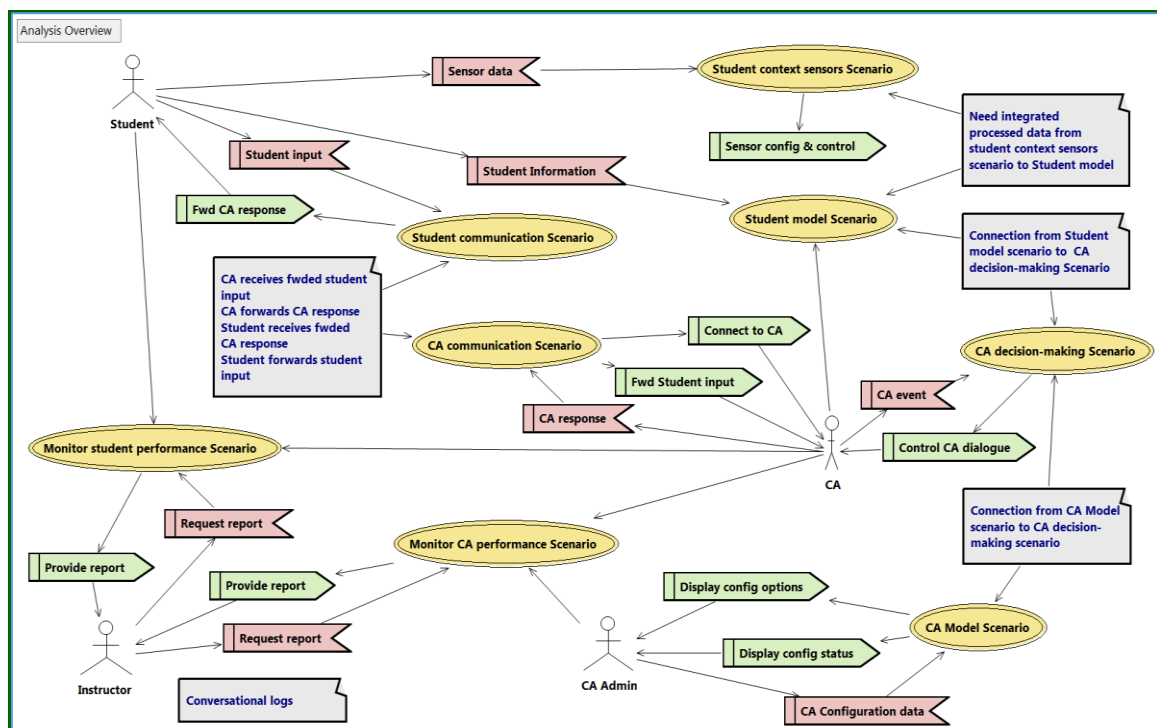


Figure 2: Analysis overview

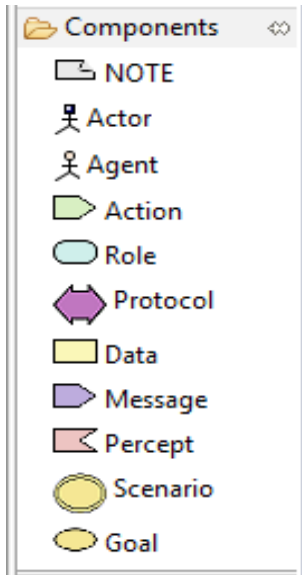


Figure 3: Key to PDT diagrams

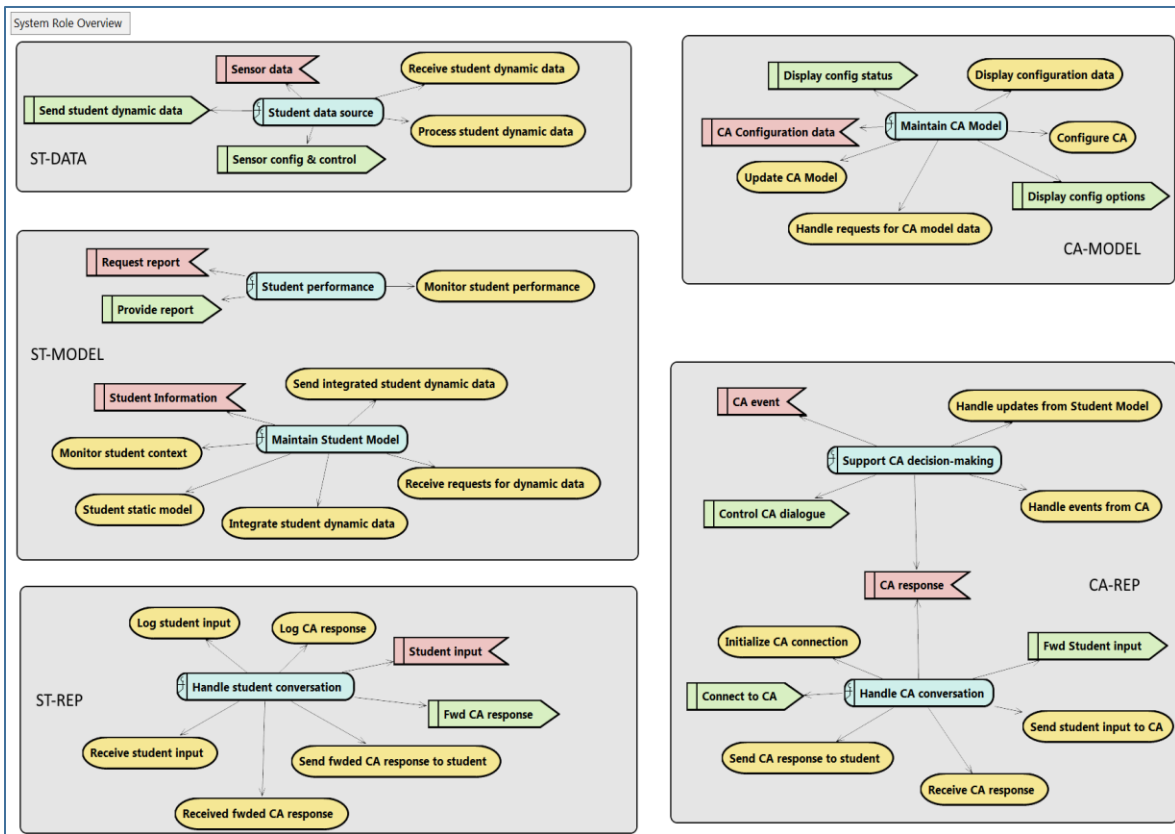


Figure 4: System roles and agent assignments

3.2.2 System roles and agent assignments

System roles are defined and goals, percepts and actions are associated with each role in Figure 4. Agent assignments are indicated by the boxes grouping the roles.

3.3 System Architecture

The system roles are grouped and assigned to agents as shown in Figure 5:

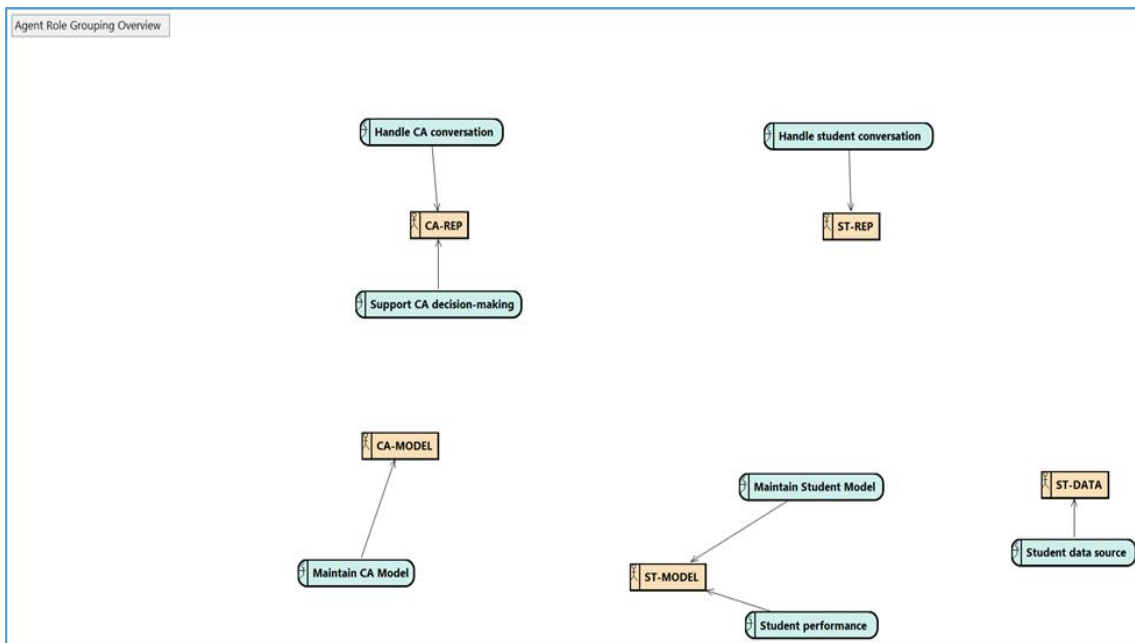


Figure 5: Agent role grouping overview

The decision to separate the modeling function from the representation function allows the task of integrating different student model data streams from the representation agents to be offloaded. This also facilitates implementing different modeling or integration schemes while leaving the conversational functions unchanged, allowing for easier maintenance of the system. A system that implements the two functions as separate agents will be more flexible and scalable, allowing the two agents to run on separate servers if appropriate. It may also be less vulnerable to bottlenecks as it is easier to run the process

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

of modeling in parallel with the function of providing an interface between the student and the CA, if they are implemented as separate agents.

Most importantly, a framework provides the most benefit to the developer if it defines the most general case. The protocols used to implement modeling and representation as separate agents are provided. If a decision is made to implement these functions as a single agent it is relatively simple to do so. If the framework only defined a single agent, a developer wishing to implement two agent would be left having to design these necessary communication protocols as they would not be provided by the framework.

Similarly, flexibility and adaptability are achieved by assigning responsibility for each student model characteristic to an individual agent. A default set of characteristics are assumed. These would rely only upon the text of the conversation, as this should

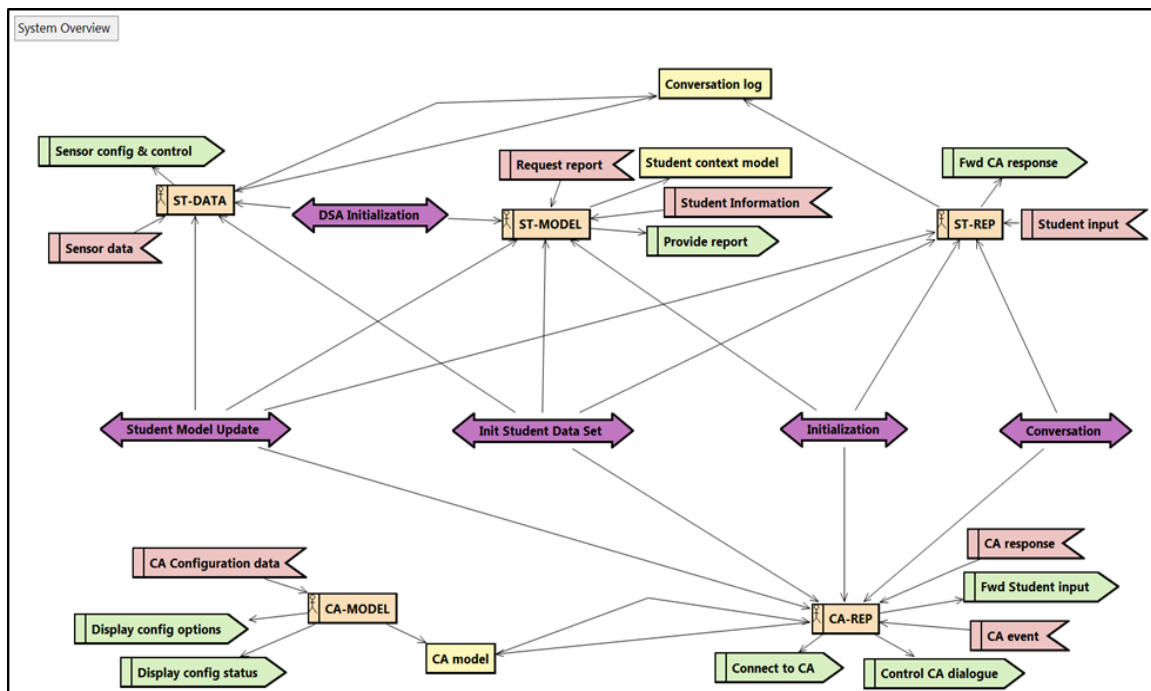


Figure 6: System overview

always be available, or previously defined data to operate. If external measurement devices (e.g. camera, BCI, biometric) are available, the agents associated with this measurement will register with the modeling agent.

The final system overview is shown in Figure 6.

3.3.1 Inter-agent message protocols

As Figure 6 shows, five messaging protocols were defined to cover interaction between agents.

An overview of the protocols follows.

3.3.1.1 *Initialization*

The initialization protocol (Figure 7) describes the messages which take place when the system starts up.

- All agents broadcast agent name and class
- Agents receive announcements and update their belief base with agent names
- Initiates further communication with assurance that target agent is up and running

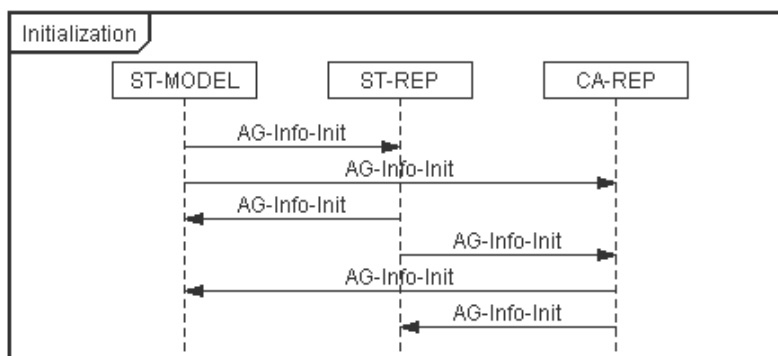


Figure 7: Initialization protocol

3.3.1.2 *Data Source Agent Initialization*

The details of initialization of the data source agents are shown in Figure 8. The steps are:

- Individual DSAs (shown as ST-DATA) respond to initialization broadcast message from ST-MODEL with agent name and types of data available (may be several for each agent)
- ST-MODEL subscribes to specific data channels as determined by negotiation with CA-REP – see Initialize Student Data Set protocol

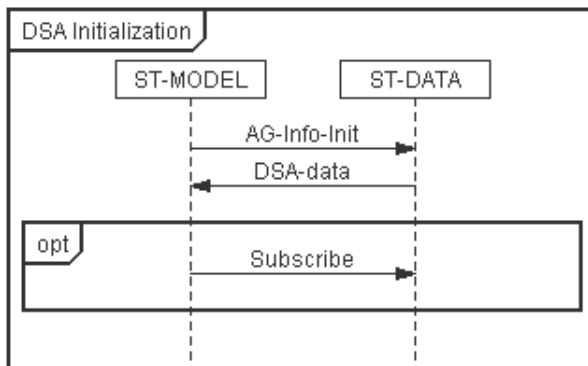


Figure 8: Initialization of data source agents

3.3.1.3 *Initialize Student Data Set*

Figure 9 provides the details of the protocol for establishing the student data set

- ST-REP queries CA-REP for list of supported data types upon receiving its Initialization announcement
- CA-REP provides list and ST-REP informs ST-MODEL it is Ready For Data

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- After receiving RFD from ST-REP, ST-MODEL forwards DSA data as it is received
- ST-REP informs ST-MODEL if data is supported, based on CA-REP list and its own needs
- ST-MODEL receives announcements from DSAs (ST-DATA) stating what data streams they provide
- ST-MODEL subscribes to DSA (ST-DATA) based on information from ST-REP

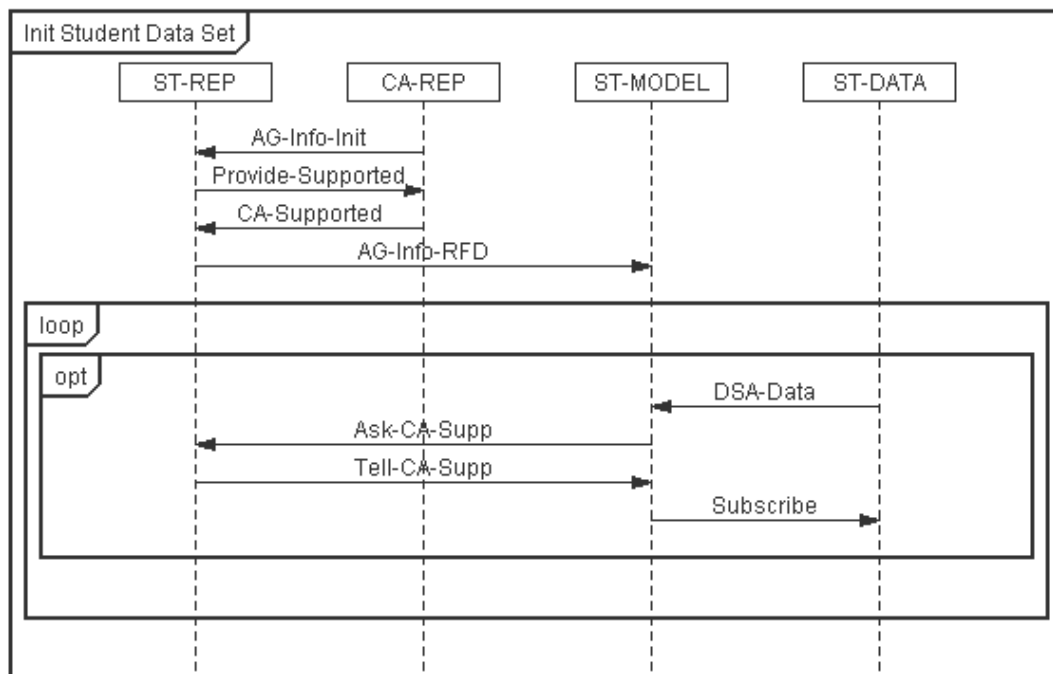


Figure 9: Protocol for establishing the student data set

3.3.1.4 Student model updates

Figure 10 shows the details of how updates to the student model are distributed to the appropriate agents.

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- The DSAs provide information at regular intervals. Data comes from an outside source, such as a sensor associated with the student or an update to the conversational log (only “Sensor data” is shown in AUML diagram)
- The DSA processes the raw data and provides the resulting information to the ST-MODEL (if subscribed)
- ST-MODEL integrates incoming data from different DSAs and forwards to ST-REP
- ST-REP may use this information, filter it, or pass it on unchanged to CA-REP
- CA-REP may use this information to modify the behaviour or provide decision support for the CA

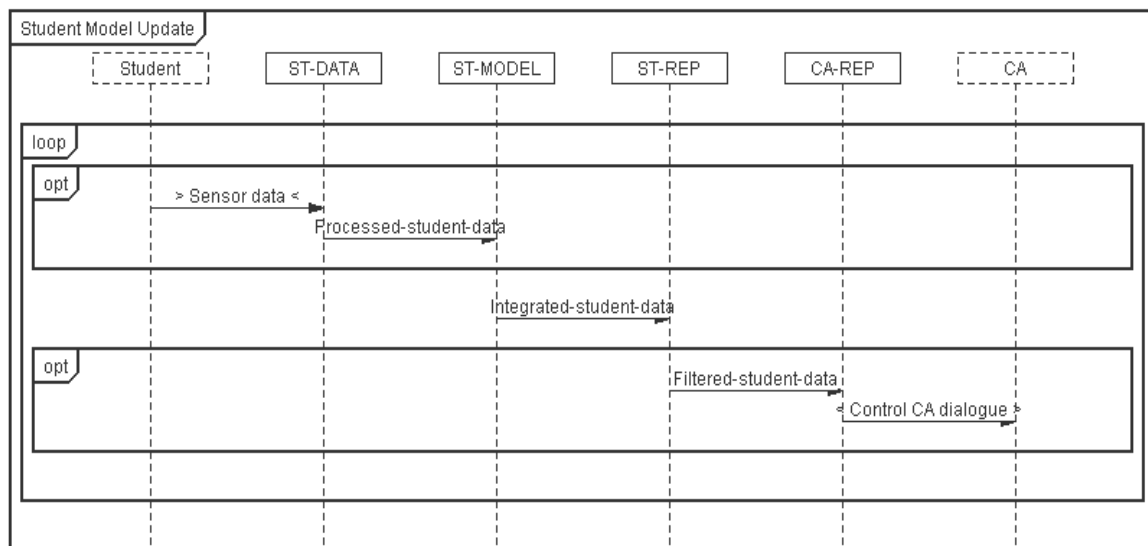


Figure 10: Dissemination of updates to student model

3.3.1.5 Conversation

The conversation protocol (Figure 11) details how information is passed between the user (student) and the CA.

- Student input is received by ST-REP and passed to CA-REP
- CA-REP sends the student input to the external CA
- CA-REP receives the CA-REP response and passes it to ST-REP
- ST-REP displays the CA response to the student

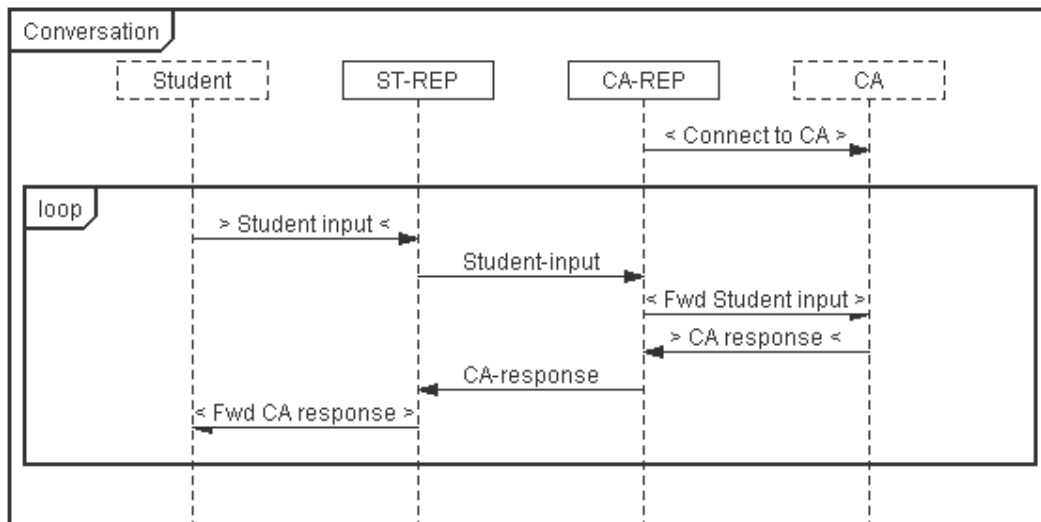







Figure 11: Conversation protocol

3.4 Detailed Design

The capabilities of agents can be described in terms of their beliefs (data) [], what they perceive about their environment (percepts) [], what actions [] they can take on their environment, and plans [] to achieve their goals. Communication between agents is carried out using messages [].

3.4.1 CA-REP agent

CA-REP is the agent responsible for representing the CA within the system. It provides an interface between the system and CA. It communicates primarily with the student representation agent, ST-REP, to provide the conversation channel between the student and the CA. It also receives information about the student, managed by the student model agent, ST-MODEL, via ST-REP, supporting the concept of what the CA perceives about the student. It communicates with the CA model agent, CA-MODEL, to support models of personality, affect, or embodiment for the CA, which aids in managing how the student perceives the CA.

The CA-MODEL function will only be minimally developed for this thesis and is therefore contained within the CA-REP agent. This includes selected knowledge stored about the student as it is delivered from the student agents ST-MODEL/ST-REP and about CA's own performance.

CA-REP is defined in terms of its messages, plans, data, and environment interaction in Figure 12.

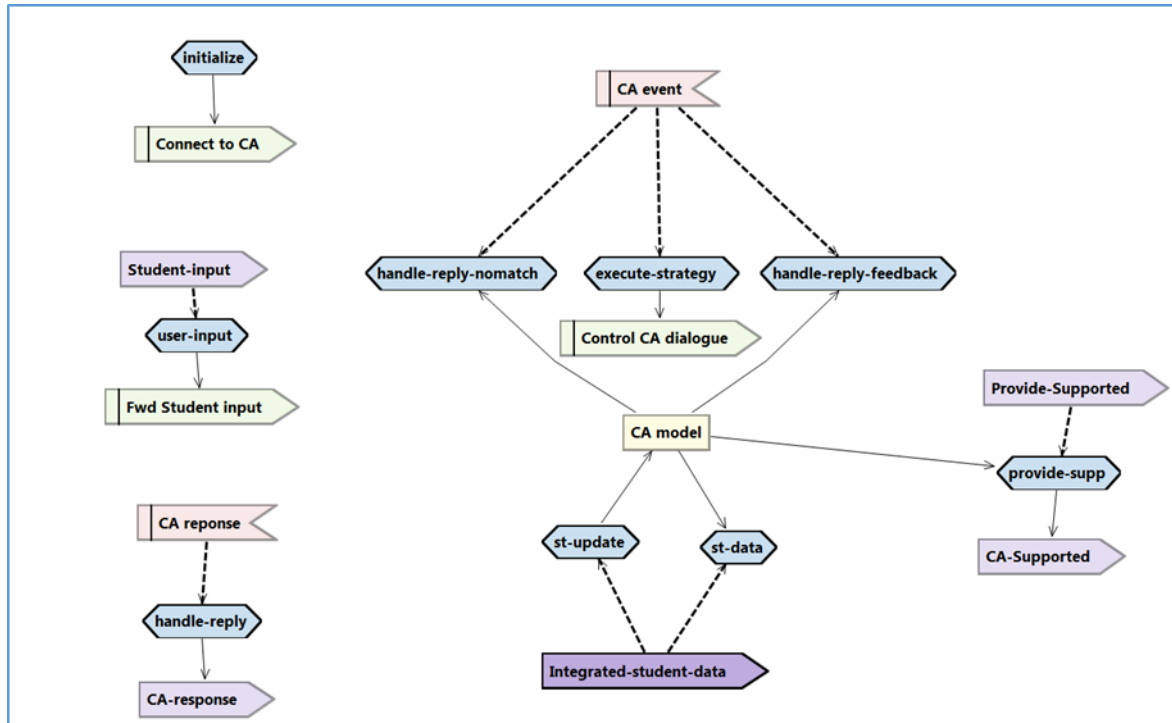


Figure 12: Agent overview - CA-REP

CA-REP plans cover the following capabilities:

- Basic initialization functions including the initial connection to the CA (*initialize*)
- Upon request, supplies ST-REP with list of student characteristics that it supports (*provide-supp*)
- Communication between the student and the CA, including accepting student input passed from ST-REP and sends to CA (*user-input*), and accepting responses from CA and forward to ST-REP (*handle-reply*)

- Handle events reported by the CA, such as not understanding student input (*handle-reply-nomatch*), decision support (*execute-strategy*), or feedback on student input (*handle-reply-feedback*)
- Handle student data updates from ST-REP, to be used in decision support (*st-update*, *st-data*)

3.4.2 ST-REP agent

The student representation agent (ST-REP) provides the interface between the student and the system, so in addition to providing the student with an connection to the CA via the agent-based system, agents that need to communicate with the student do so via ST-REP. ST-REP's responsibilities and capabilities are show in Figure 13, and described below:

- Provides user interface with student
- Works with CA-REP to provide communication channel between student and CA/ITS
- Logs conversation information to database
- Works with CA-REP and ST-MODEL to determine relevant student data set
- Accepts student data from ST-MODEL, potentially using this data to modify the interaction with the student, and/or filtering the data before forwarding to CA-REP

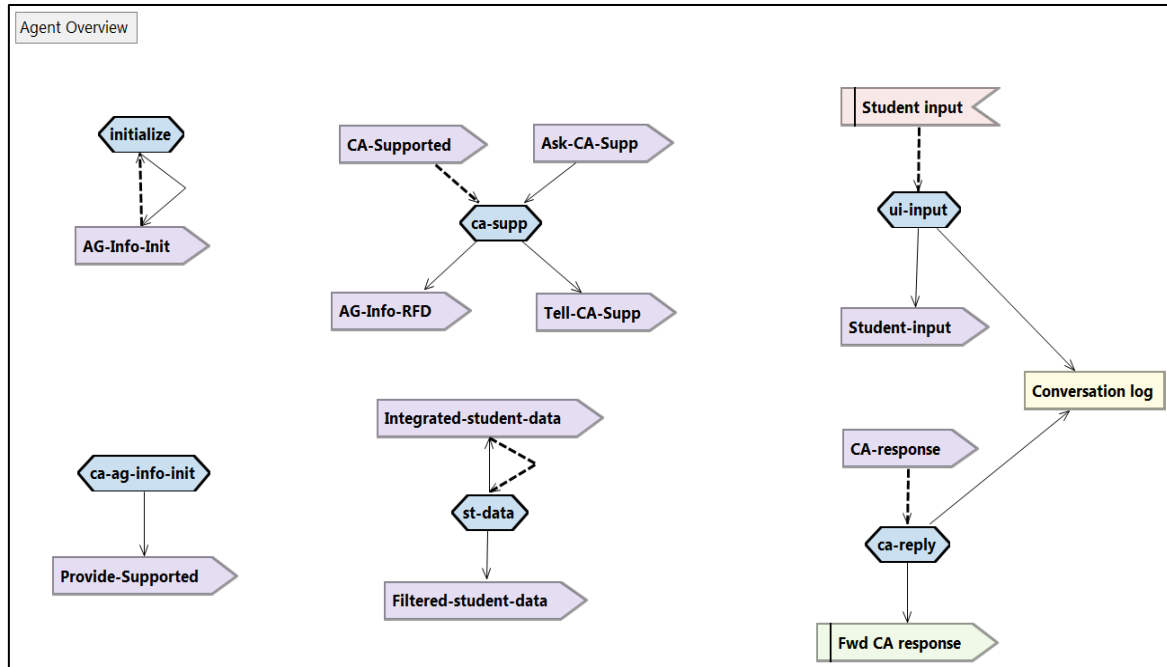


Figure 13: Agent overview - ST-REP

3.4.3 ST-MODEL agent

- Figure 14: Agent overview - ST-MODEL
- Works with ST-REP to define relevant student data set and subscribes to data streams supplied by Data Source Agents (DSAs) referred to here as ST-DATA.
- Integrates data and information from DSAs and sends to ST-REP
- Maintains ongoing model of student information

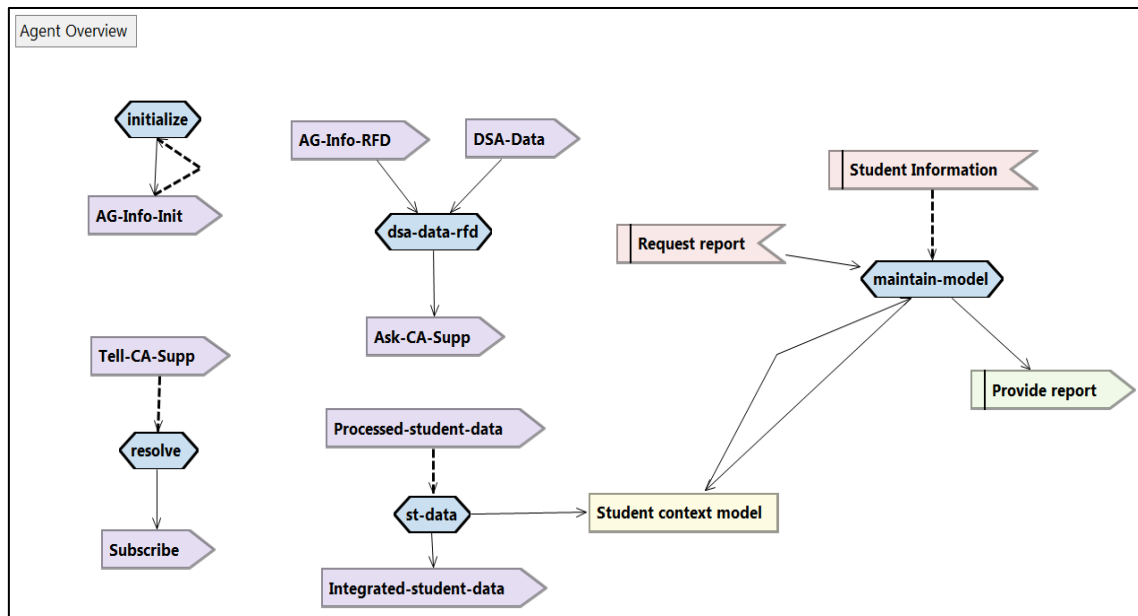


Figure 14: Agent overview - ST-MODEL

3.4.4 Data source agents

Regardless of the data that individual agents provide, all DSAs will conform to the same communication protocols, described briefly below and in the proceeding section (Figure 15).

- A DSA (ST-DATA) generally processes some raw data – from a sensor, camera, conversational log, etc. -- and uses it to provide some information about the student. It may provide multiple information channels. For example it may provide three channels: a measure of boredom, frustration, and confusion.
- At initialization the DSA announces to ST-MODEL what channels it has to offer. Using the example above this might be: `dsa_data(affect, boredom)`, `dsa_data(affect, frustration)`, `dsa_data(affect, confusion)`.

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- The DSA accepts subscription requests from ST_MODEL for any or all of these channels and proceeds to provide the relevant information at regular intervals to ST_MODEL.

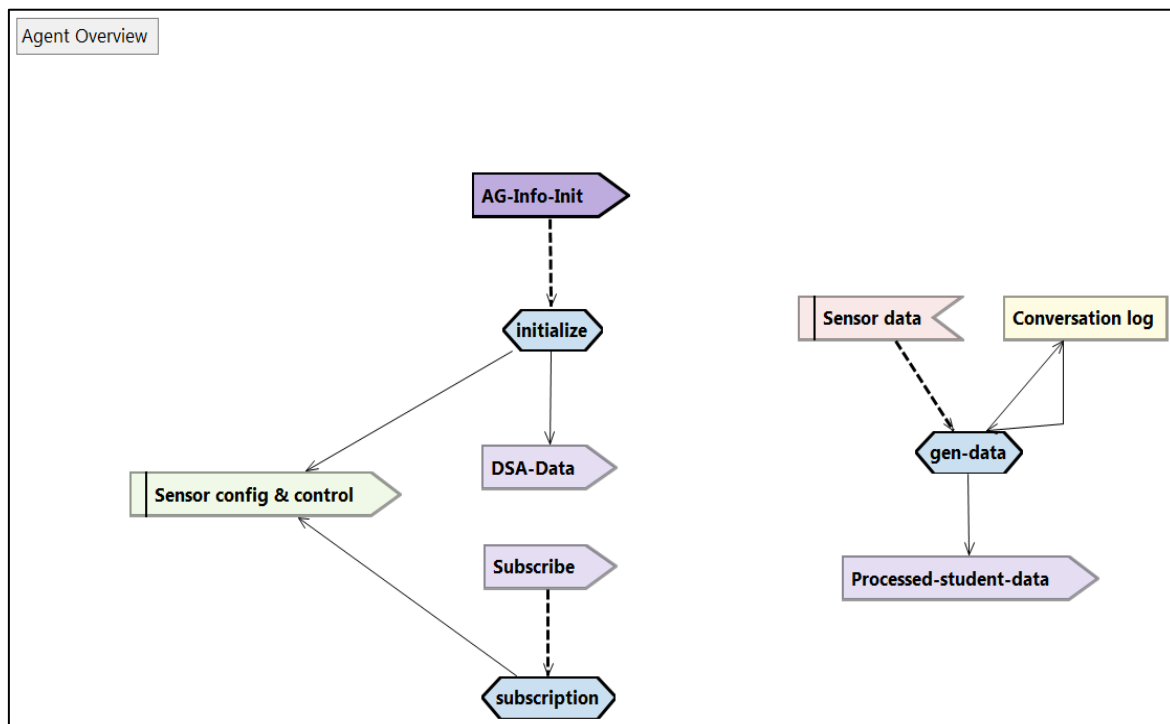


Figure 15: Agent overview - ST-DATA

Chapter IV – Analyzing Dialogue

4.1 Background

While the system is intended to be able to adapt to information sources about the student that are available, it is prudent to have some basic default measures that indicate the affective state of the student, with regard to the learning activity. An obvious source of data is the conversational record, referred to as the “chat log”, of the student interaction with the learning system. This information is always available and easily accessed, does not rely on special hardware or sensors, and is arguably of particular relevance to a conversation-based application.

Two measures were targeted: conversational quality and appropriateness; and user conversational behaviours.

4.1.1 Conversational quality and appropriateness

This section describes how discourse features in the conversation can be used to indicate the level of engagement, the possible causes of loss of engagement, and potential strategies for recovering user engagement. To do this we take a closer look at what engagement means in the context of the features of the conversational record. Using O’Brien and Toms’ (2008) description of engagement based on four stages: point of engagement, sustained engagement, dis-engagement, and reengagement, it is possible that the user will cycle through the last 3 stages multiple times during a conversation with a CA.

In this research, we designed a conversational agent Freudbot to simulate a historical figure, Sigmund Freud. Users are able to converse with Freudbot, using text input, as if in the role of interviewing him. Freudbot is designed to respond in first person to questions and comments about Freud's life, family, theories, and colleagues. In all over 90 topics, broken into multiple narrative chunks, are programmed to be delivered to the user following basic rules of conversation, such as greeting, turn-taking, and repairing misunderstandings. Responses are based on pattern matching using Artificial Intelligence Markup Language (AIML), a well-known XML-based language and platform for creating and serving chatbots (see <http://www.alicebot.org>). The AIML is supplemented by additional software to support the rules of conversation, track progress through each topic discussed, and support the narrative delivery.

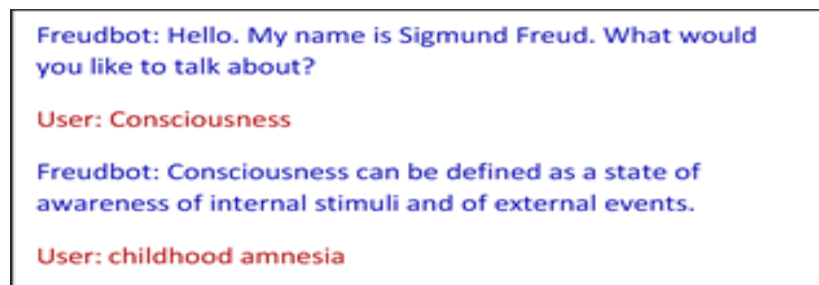
For this study, discourse features associated with engagement were developed as an extension of previous work in social presence (Heller et al., 2016). Engagement has been linked to social presence. Polhemus, Shih, Swan, & Richardson (2000) examined online discussion responses from 35 students. They found a positive correlation between affective language associated with social presence (e.g. reflection, personal address, and self-disclosure) and indicators of learning engagement.

4.1.2 Conversational behaviour

Detection of engagement focuses on user behaviors. We draw on the description of conversational engagement by Yu, Aoki, & Woodruff (2004) as a “commitment to interaction” and is expressed in terms of the degree to which the user is interested in, or attentive to, the conversation. Certain user behaviors were identified as indicating

engagement while interacting with the CA. In general terms, those indicators that the subject is cooperating in the role of conversant suggests engaged behavior.

Discourse features predicted to be associated with engagement include responding to statements and questions, and using complete sentences when appropriate. A user who is not engaged might ask single or keyword phrase type questions, as if using a search engine. Commenting on the performance of the CA may also indicate a break in engagement, depending on how it is done. “You are making a lot of mistakes” suggests a higher level of engagement than “This bot makes a lot of mistakes”, because in the first case the user is still cooperating in the role of pretending to speak with an intelligent being. Figure 16 shows an example of a participant’s use of keyword type phrases, demonstrating how this bears little resemblance to a human-human conversation. This would be rated as having a low conversation level.



```

Freudbot: Hello. My name is Sigmund Freud. What would
you like to talk about?

User: Consciousness

Freudbot: Consciousness can be defined as a state of
awareness of internal stimuli and of external events.

User: childhood amnesia

```

Figure 16: Excerpt from conversational log

If engagement with a role-playing CA is dependent on, or at least influenced by, the degree to which the user is willing to suspend disbelief and carry on a conversation with the software, it follows that anything that betrays the illusion that the CA is a capable conversational partner has the potential to disrupt engagement. In the case of Freudbot, the

types of errors that the authors predict will lead to loss of use engagement with a CA include:

- Inappropriate answers to questions related to Freud. These mistakes, often caused by a poor keyword match, expose the pattern matching nature of the interface.
- Stating non-understanding of something Freud would understand, particularly if stated simply, such as “Tell me about X”. This makes it difficult for the user to continue the pretense of speaking with Sigmund Freud. This is made worse when Freudbot asks the user to restate their question in another way, but still fails to recognize the topic.
- Repeating information. Although Freudbot is programmed to recognize when it is repeating content, and acknowledge the fact, this is a common sign of a programmed response.

In essence, these are failures which expose the programmatic nature of the CA, revealing, for example, a pattern matching mechanism. CA’s are often able to detect these types of failures and therefore self-evaluate their performance. Some failures may be supplied directly by the user’s feedback concerning the CA’s responses. For example, if the CA’s response does not make sense, it would be reasonable for the user to say so.

If we assume that such performance issues can lead to a breakdown of the illusion of intelligence, then this internal and external feedback has the potential to provide a predictor of loss of engagement as the conversation progresses. The goal is to recognize when CA is performing poorly, and to attempt to confirm that this is negatively affecting user conversational engagement by analyzing the user’s responses. The first part is relatively easy to achieve as the CA is aware of at least some failures, such as inability to

match user input. To address the second goal, assessing user input, machine learning techniques are used to train a classifier. The next section describes how this was accomplished, and how it fits within an agent-based framework which supports the interaction between the user and the CA.

4.2 Development

Conversational log data was collected from several years of interactions with publicly available role-playing CAs that simulate well-known historical figures. Conversational logs from two previous controlled studies (Heller et al., 2016; Heller & Procter, 2009) were also available. The 2009 conversational logs provided data from 10 minute conversations by 90 participants chatting with the historical figure CA, Freudbot.

A coding scheme was developed to classify user input and CA responses in the conversational logs. Key features for user input are associated with level of user engagement and include:

- Response Appropriateness: answering questions, responding to requests, addressing the topic under discussion, or changing to another domain related topic.
- Conversational quality: playing the role of conversant: using full sentences or phrases, not lone keywords, gibberish or random characters, non-repetitive utterances

The CA output was categorized by type of statement, such as a request for a topic of interest, content associated with different topics, or answers to questions. Statement types associated with non-understanding on the part of the CA are of particular interest. Each of these types was associated with a different strategy for addressing the inability to recognize the user's input. These include:

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- Ask the user if they have another topic they would like to discuss.
- Tell the user that their comment or question was not understood and ask them to restate it in another way. As a follow up to the user's response, if Freudbot still does not understand, then he says so and asks the user if he should continue talking about the current topic
- Ask the user a leading question, such as "Did you have a happy childhood?" After the user responds, Freudbot provides a follow up statement starting a new topic. "My childhood was pretty chaotic. We moved to Vienna before I was three."
- Make a "dead end" statement, such as "I am at a loss for words."

If Freudbot is unable to understand the user input in several consecutive exchanges, the user is asked if Freudbot should continue with the current topic, or Freudbot suggests a new topic at random. For example, "Would you like to talk about my cocaine habit?" or "We did not finish talking about my childhood. Would you like to talk about that?" The WEKA (version 3.7.2) data mining tool (Hall et al., 2009) was used to experiment with several machine learning algorithms for classification. The annotated logs were used to train two classifiers: one to target user response appropriateness, and another for user conversational quality. Support Vector Machine (SVM), Naive Bayes, and k-Nearest Neighbors (kNN) classifiers were tested. The training data was made up of 2716 CA/User turn pairs. Each instance consisted of a value representing the statement type of the CAs output (described earlier) and the text of the user's response, converted to word vectors. The classifiers were trained and tested using 10 fold cross validation.

As can be seen in Table 1 and Table 3, there is a large imbalance within the classes. Two approaches were employed in an attempt to compensate for this, as described in He & Garcia (2009): resampling, and the application of an adjusted cost vector. The goal of resampling is to select the data sets so as to balance the distribution of classes used for training. WEKA provides the Resample filter to do this. Cost-sensitive methods allow the cost of misclassifying each data example. Using an extreme case for an example, a trivial classifier can produce an accuracy of 82% for Conversational Quality (see Table 3) by identifying everything as class 1, albeit with a large number of false positives. By increasing the cost of false positives for this case, the accuracy of the classifier is lowered and the algorithm must adjust to optimize its accuracy. This was implemented in WEKA using the CostSensitiveClassifier meta classifier.

4.3 Evaluation of Algorithms

4.3.1 Conversational quality and appropriateness classifier

Table 1 and Table 2 show the confusion matrix and performance results, respectively, for the “response appropriateness” class SVM classifier. The SVM algorithm provided the best match for the priorities that were identified. Importantly, it is a fast algorithm that calculates results quickly enough for real-time classification of the conversation. Four classes are defined. Class 1 is assigned to user responses that address the preceding CA output, answering a question or making a comment on the same topic. Class 2 is a request to change topic while remaining “on-task”, or within the domain of the CA’s knowledge.

For this classifier, it is particularly important to identify classes 3 and 4, which would be associated with poor ratings, typically the result of comments or questions which

are not directly related to the task. Class 3 should be assigned to comments about Freudbot (“This is stupid”), while class 4 is associated with off-topic questions (“Who will win the Super Bowl?”) or even random characters. It is these cases where some sort of intervention is appropriate, but the preference is to err on the side of caution and accept some misses, i.e. false negatives are more acceptable than false positive for these two classes. It is expected that unnecessary interventions run the risk of confusing the user and reducing the credibility of the CA in their eyes.

Table 1. Response appropriateness confusion matrix

<i>Classified as</i> →	1	2	3	4	<i>Dist</i>
1	1894	134	19	6	0.76
2	271	164	9	8	0.17
3	77	21	25	4	0.05
4	38	11	4	31	0.03

Table 2. Response appropriateness classifier performance

<i>Class</i>	<i>FP Rate</i>	<i>Precision</i>	<i>Recall</i>	<i>F-score</i>	<i>ROC</i>
1	0.58	0.83	0.92	0.87	0.69
2	0.07	0.50	0.36	0.42	0.68
3	0.01	0.44	0.20	0.27	0.81
4	0.01	0.63	0.37	0.47	0.78

The confusion matrix provides details about the number true and false positives (TP and FP), and true and false negatives (TN and FN). Each row is associated with an actual class. Each column shows what class was assigned by the classifier. A perfect classifier would result in zero values in all cells except the diagonal from row 1, column 1 to row n, column n, where the classified value = actual class. Values in the same row, not

on the diagonal, indicate the number of false negatives for that class. Similarly, values in the same column, other than the diagonal, indicate false positives for a given class.

The performance metrics shown in Table 2 are based on statistical analysis of the confusion matrix data. With an unbalanced distribution of classes, a simple accuracy measure can be misleading. A trivial classifier that assigned the most common class to all cases would have an accuracy equivalent to the proportion of test cases that have that class. For example, if 80% of test cases are class A, a classifier could assign class A to all cases and still achieve an accuracy of 80%.

The measures reported are intended to address this issue by taking into account incorrectly assigned classes. FP rate is the rate of false positives, i.e. instances falsely classified as a given class in a ratio to the number of false negatives for that class. Precision is the proportion of true positives for a class divided by the total instances classified as that class (TP+FP), while Recall is the proportion of instances classified as a given class divided by the actual total in that class (TP+FN). F-Measure is a combined measure for precision and recall calculated as $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$. The area under the ROC curve, which plots the true positive rate (equals Recall) to the false positive rate, provides another popular measure of how well the classifier identifies classes while reducing false positives. A value of 0.5 represents chance, while a perfect classification would have a value of 1.

Table 3 and Table 4 show the confusion matrix and performance results for our SVM classifier for conversational quality. Class 1 is assigned to user input that uses full sentences, where class 3 is associated with “keyword” type input, similar to abbreviated text one might use with a search engine. Class 2 was assigned to a “not sure” condition.

In general the “problem” class (3) is of interest, and for similar reasons as those stated for the appropriateness classifier. Again the preference is to err on the side of caution and avoid false positives for class 3.

Both classifiers have much room for improvement. However it’s important to remember how they will be used. Classification data from the agent will be processed by a student model agent which can choose to act on that information based on a confidence rating for the associated class and other heuristics, such as the frequency of the rating. The student model agent is responsible for accepting data from other similar data source agents as well. This additional data would be used in combination with that from the student response data to corroborate the classification. For example, a cognitive model based agent may also report signs that the student may be frustrated. This combined with a low conversational quality rating would provide greater confidence that some sort of intervention is required. This may in tern trigger a suggestion for a new topic, or asking questions designed to re-engage the user.

Table 3. Conversation quality confusion matrix

<i>Classified as →</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>Dist</i>
<i>1</i>	2098	4	112	0.82
<i>2</i>	68	2	45	0.04
<i>3</i>	112	3	272	0.14

Table 4. Conversation quality classifier performance

<i>Class</i>	<i>FP Rate</i>	<i>Precision</i>	<i>Recall</i>	<i>F-score</i>	<i>ROC</i>
<i>1</i>	0.36	0.92	0.95	0.93	0.80
<i>2</i>	0.00	0.22	0.02	0.03	0.39
<i>3</i>	0.07	0.63	0.70	0.67	0.85

These are initial results, based primarily on default configuration values for filters and classifiers in WEKA. It is hoped that performance will improve as more of the CA logs are annotated and added to the training data.

4.3.2 User behaviour detection

Evaluation of the conversational behaviour identification algorithms was achieved by manually rating 26 conversations (613 turn pairs) from the chat logs of a previous experiment (Heller & Procter, 2009). Each conversation was assigned a rating for each of the three types of behaviour: *trying*, *keywording*, and *moreing*. A standalone version of the algorithm used by the DSA was created for testing purposes. Results from comparing the manual and automated ratings are shown in Table 5. Four versions of the *tryer* algorithm were tested, producing very different results, depending on how what type of Freudbot response was taken into account and other adjustments to the counting. False positives were judged to have a negative effect since they are likely to trigger inappropriate interventions. This can be confusing to the user, and undermine the perception of intelligence that plays a large part in engaging the student.

Table 5: Behaviour algorithm testing

	<i>True Pos</i>	<i>True Neg</i>	<i>False Pos</i>	<i>False Neg</i>	<i>Total</i>
<i>Tryer1</i>	8	8	1	9	26
<i>Tryer2</i>	17	3	3	3	26
<i>Tryer3</i>	3	9	0	14	26
<i>Tryer4</i>	12	7	1	6	26
<i>KW</i>	3	17	6	0	26
<i>More</i>	2	23	0	1	26
<i>Total</i>	30	51	10	13	104

The fourth *tryer* algorithm (highlighted) was selected as having the best balance between catching the behaviour and not accidentally triggering a false intervention. The next best alternative would be the second algorithm which is more aggressive in identifying the behaviour but also has a higher risk of a false positive. The *keyworder* and *morer* algorithms produced satisfactory results but more conversations need to be manually coded to produce enough examples of these two behaviours to have confidence in the results.

Chapter V – Implementation

This chapter describes the architecture of the agent-based framework and how a proof-of-concept implementation was developed, which demonstrates the process of integrating an existing CA to the framework. The implemented system provides a platform to test and evaluate the performance of the system, as well as a platform to collect live data from students interacting with the system.

5.1 Overview of System Architecture

A high level overview of the agent framework is shown in Figure 17. The student and the CA, shown at the bottom, communicate through the agent framework above them. The architecture consists of three layers. The *Representation* layer is responsible for providing an interface between the student and the CA, with an agent representing each. The *Model* layer maintains information about the state of each of the participants, again with agents assigned to each. The *Model Sources* layer provides information to the Model layer agents. Multiple data source agents (DSAs) process data from devices and provide one or more information channels to the model.

5.1.1 Student Representation (ST-REP)

Goal: Represent the student by: Communicating student input to CA (via CA-REP); Communicating CA response to student; Provides feedback to student based on data provided by CA-REP; Provides student model data from ST-MODEL to CA-REP.

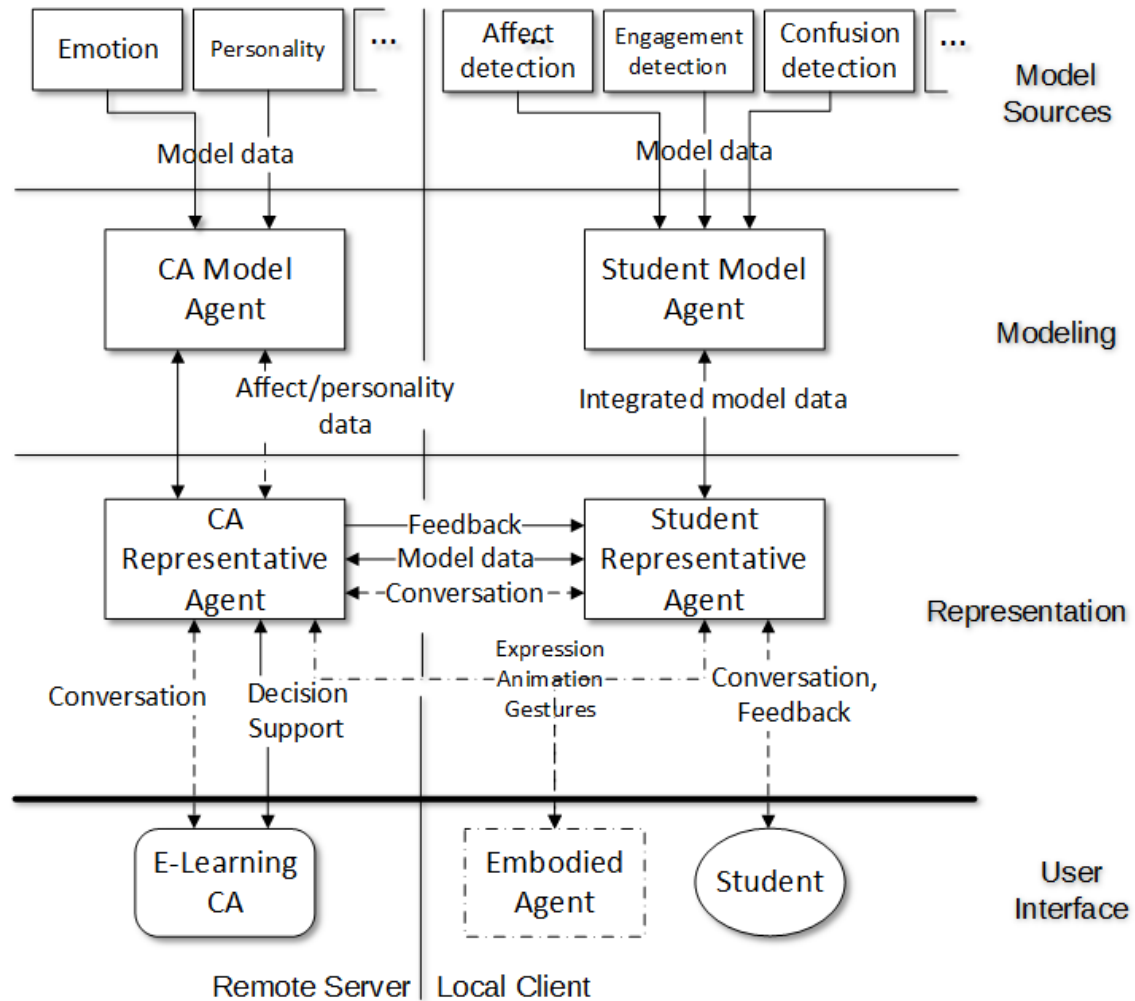


Figure 17: System architecture

5.1.2 CA Representation (CA-REP)

Goal: Represent the CA by: Sending student input (from ST-REP) to CA; Sending CA response to student (via ST-REP); Recommending conversation strategies based on student data from ST-MODEL; Analyze student responses and provide feedback to ST-REP; Send data to embodied agents to support lip-sync, gestures, expressions.

5.1.3 Student Modeling (ST-MODEL)

Goal: Maintain data about student by: Subscribing to data source agents (DSAs) based on CA requirements; Receiving data from DSAs (affect, personality, engagement, goals); Sending updates to representation layer agents integrating data where appropriate.

5.1.4 CA Modeling (CA-MODEL)

Goal: Maintain information about CA by: Subscribing to data source agents (DSAs) based on CA requirements; Updating CA-REP from data provided by DSA models (e.g. emotion, goals, personality, CA performance assessment) to support embodied agents and conversation strategy recommendations.

5.1.5 Student and CA Data Source Agents (DSA)

Goal: Provide a source of data about the student (or CA) by: Provides information channel using a publish/subscribe mechanism (e.g. affect detected from facial expressions, engagement detected from conversational analysis).

5.2 Proof of Concept Implementation

A version of the described system has been implemented using Jason, a Java-based interpreter of an improved version of AgentSpeak(L), which supports multi-agent systems (MAS) based on the BDI (Belief-Desire-Intention) agent paradigm (Bordini, Hübner, & Wooldridge, 2007). The BDI architecture is commonly used in the development of cognitive agents and has been used successfully in agent-based pedagogical applications (Soliman & Guetl, 2012).

5.2.1 Scope of implementation

This system implements the two representation layer agents (ST-REP and CA-REP) and the student model agent (ST-MODEL) as BDI agents. ST-MODEL employs a simple integration strategy that combines multiple sources of a data type using a weighted average based on the accuracy rating provided by the DSA. The CA-MODEL agent and associated DSAs will be developed as part of a future research phase.

5.2.2 Goals of implementation

5.2.2.1 Demonstrate the process of integrating an existing CA to the agent-based framework

The Freudbot CA described in Chapter 4 was used as the test case for the proof-of-concept implementation of the framework. The tasks required to modify the CA, and to adapt the CA representation agent, are described in section 5.4

5.2.2.2 Provide a platform to test and evaluate the performance of an implementation

The resulting implementation was tested for latency and computing resource impact as described in section 5.6.

5.2.2.3 Provide a test platform to collect live data from students interacting with the system

The completed implementation was used as the basis of a short study using volunteer Psychology students as participants. Students chatted with the CA via the agent-based system and rated the experience by completing a short questionnaire. In addition to collecting student feedback, this also served to prove that the system could be applied practically.

5.2.2.4 *Demonstrate the ability to add or add intelligence and change functionality of the CA*

A common event that must be handled by the CA is the case when the user input cannot be understood, possibly because of poor grammar, spelling errors, nonsensical input, or shortcomings in the CAs programming. The Freudbot CA has several possible responses and selects ones of these randomly. To demonstrate the decision support role of the CA-REP agent, and to show how these the agent-based system can provide additional intelligence to the CA, a new strategy for selecting an appropriate response to unmatched user input was implemented. The new strategy was based on how many “misses” had taken place and is loosely based on the method used by Silvervarg & Jönsson (2011). They describe a simple but effective method of selecting a “repair strategy” by increasing the level of conversational control with each successive attempt to handle unmatched input. First a clarification is requested. If that fails, the user is requested to provide a topic. If the response is still not understood, then the CA proposes of a new topic. The final approach is to ask a question or start a new narrative. Freudbot’s strategies for unmatched user input vary in a similar way in terms of how much control the CA takes in the conversation.

CA-REP also directs the CA to deliver conversational interventions under certain conditions. These are discussed in further detail sections 5.4 and 5.5.

5.2.3 Core DSAs - Conversation text classifiers

The DSAs are the key components to providing a dynamically configured system capable of adapting to whatever student information is available. However, as this is a CA application, a core set of conversation-based DSAs should always be available. Currently two DSAs have been created which process the conversational log to provide measures of

user engagement. They rate user input on whether it is conversational in nature, and appropriate to the CA’s statements. These agents are described in detail in section 5.5.

5.2.4 External communication – connecting users and devices to the system

5.2.4.1 *Support for distributed agents*

Jason supports a JADE (jade.tilelab.com) environment to provide distributed MAS. This was used to provide remote access for users and devices. A Java servlet connected to a JADE agent supplies the student interface (Figure 18). The top part of the diagram represents the agent framework shown in Figure 17. User input through a web page is sent from the Jade agent to the ST-REP agent, using Agent Communication Language (ACL) messages. This allows for a relatively thin client and a mechanism for devices to connect

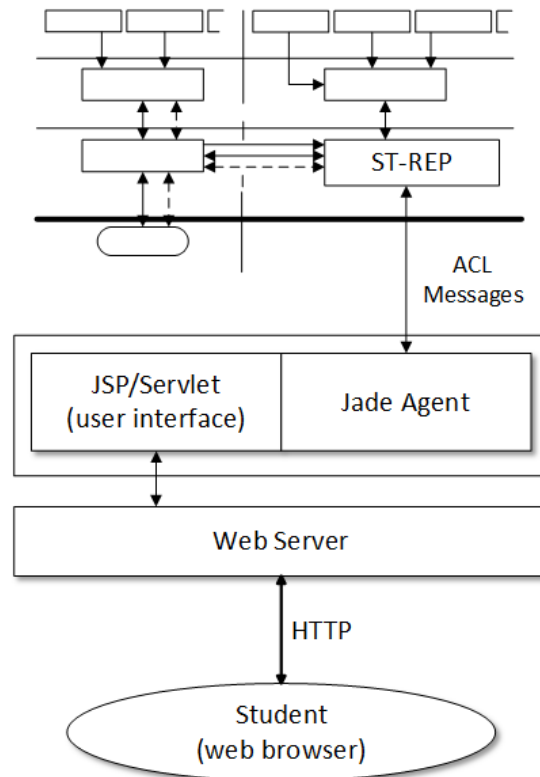


Figure 18: Remote user interface communication

to DSAs through a Jade agent installed on the user's system. Most importantly, the JADE agent environment for Jason makes it easy to distribute the agents across different servers, allowing for greater flexibility and scalability.

5.2.4.2 *Multiple user session support*

In order to allow multiple students to access the CA concurrently, it was decided that a set of agents would be invoked for each user session. This means that each student has their own CA-REP, ST-REP, ST-MODEL, and DSA agents assigned to them. This approach has several advantages. The code for the agents is less complex and more streamlined as it does not have to manage the allocation of data and messages for multiple students. Although multiple sets of agents may use more computing resources, this solution is more scalable, as the agents can be run across multiple servers when the JADE environment is used.

5.3 System Data and Execution

5.3.1 System data

Student and CA data are represented as a tuple $\langle C,S,V,A \rangle$ (Category, Sub-category, Value, Accuracy). For example $\langle \text{Affect, Engaged, 2, 0.6} \rangle$. ST-MODEL maintains a list of known data types in the form $\langle C,S,_,_ \rangle$ and current student data.

5.3.2 System execution

Agent interactions and system behavior are described in terms of six activity phases in Table 6. Bold text identifies components to be provided or customized by the developer to meet the needs of the CA or student data to be used. With the exception of Initialization, activity phases occur asynchronously. CA-REP and ST-REP are responsible for combining resulting messages for the user into the conversation.

Table 6: Agent activity phases.

<i>Initialization</i>
<ol style="list-style-type: none"> 1. User connects to JSP agent which requests new start from Invoker agent. 2. Invoker creates a set of agents (ST-REP, CA-REP, ST-MODEL, core DSAs). 3. DSAs announce data types available. CA-REP announces data types needed. 4. ST-MODEL resolves needed and available data, subscribes to DSA data streams. 5. ST-MODEL records cases of more than one source of a data type (typically from two or more DSAs) and stores these for integration.
<i>Communication</i>
<ol style="list-style-type: none"> 1. ST-REP receives user input (text). Sends to CA-REP. 2. CA-REP receives user input. Sends to CA. 3. CA-REP receives CA response (text). 4. CA response may trigger a Decision Support action. 5. Student data update may trigger an Intervention action. 6. CA-REP sends CA response to ST-REP. ST-REP display response to student.
<i>Intervention (CA-REP initiated)</i>
<ol style="list-style-type: none"> 1. CA-REP receives student data update, executes Student Update action. 2. Determines if an Intervention plan should be executed. 3. Gets CA output from intervention (e.g. “Would you like to change topics?”). 4. CA-REP combines CA intervention output with response to student input.
<i>Decision Support (DS) (CA initiated)</i>
<ol style="list-style-type: none"> 1. CA-REP receives CA response indicating decision request, triggering a DS plan. 2. CA-REP requests student data from ST-MODEL if needed. 3. CA-REP executes DS plan (e.g. select topic), sends input to CA. 4. CA sends new response. CA-REP sends CA response to ST-REP.
<i>Student data update</i>
<ol style="list-style-type: none"> 1. DSA processes incoming data (e.g. wearable device, camera, new text log entry) 2. DSA sends data update to subscribers for each information streams <C,S,V,A> 3. ST-MODEL receives student data message and executes integration plan if other sources for the same data type are received. 4. ST-MODEL sends data update to CA-REP (via ST-REP)

5.4 Implementation Tasks

The template agents provide the underlying communication protocols and some default functions. To adapt the framework to a specific CA and student model use case, some functions (plans) and data (beliefs) must be provided. Functions which can be extended are also identified.

5.4.1 CA data set

The following process is used to define a CA's ability to use student information. This is used to determine potential candidates for student information detection (e.g. specific measures of affect) based on what aspects of the CA behavior can be modified, and with what student information might be associated with those behaviors.

1. **Identify strategies:** What aspects or behaviors can the CA control (or be able to control with modification to the CA). (E.g. topic selection, hint giving, interventions)
2. **Identify triggers, characteristics, or signals:** What student information is relevant to these strategies and should signal a change or action (E.g. confusion, boredom)
3. **Update initial belief base in CA-REP** to identify information that can be used (E.g, `st_data(affect, boredom, 0)`). This data will be delivered to ST-MODEL during the initialization phase.
4. **Write plans for CA-REP to handle student model data updates**, i.e. changes to `st_data(C,S,V)`, initiating actions if necessary.

5.4.2 CA communication and decision support

1. **Handle communication between the CA and CA-REP** by modifying the internal action `.ca_send()`. For Freudbot, a simple HTTP protocol is used. `Ca_send()` passes the user input as it's parameter and initiates a `handle_reply` goal event to provide the CA's response.
2. **Associate triggers with strategies** and implement strategy algorithm. (E.g. provide hint if confused). In Jason this may be implemented as an internal action. This is done by writing `handle_reply` plans to process responses from the CA. The default `handle_reply` plan sends the CA's response on to ST-REP. Additional plans can be

implemented to handle different conditions flagged by the CA, or changes to the student data.

3. **Modify CA as needed** to interface with the strategy algorithm in CA-REP. For example, with Freudbot, the CA was modified to create a “NOMATCH” handle_reply goal event rather than handle the condition where the user input cannot be recognized. The associated handle_reply plan executes a strategy process to determine which of the selection known to the CA should be used, and sends the result to the CA. The CA is modified to accept this choice and execute the corresponding strategy.
4. **Intervention plans** are used to address issues that have been identified from data coming from the student model agent, the CA model agent (e.g. monitoring CA performance), or flagged by the CA, if it has that capability. Generally these plans direct the CA to do something, or the CA representation agent may inject a response on behalf of the CA. It may take immediately or when then next CA response is generated, depending on what is appropriate in terms of the conventions of normal conversation. An example might be an observation about how the user is conversing combined with some conversational cues to make better use of the CA’s capabilities.

5.4.3 Student model plans

Model plans provide the logic to process data updates from the DSAs. A model goal event is created for each update. The plans to handle these are unique to the data type and the DSA that produced them. The default plan simply stores the data in the belief base. The plans used for Freudbot to manage conversational engagement keep a complete record of the classifications provided by the DSA over time, as well as an average response over

a predefined window. This is a simple example of ensuring that spurious changes in the data do not initiate actions by the CA.

5.4.4 Integration strategies

An integration goal event is triggered for each data update from a DSA. Integration plans combine data from different sources. There are several reasons why this might be done to compare different sources of the same data to determine confidence in a rating. For example, the system can place more confidence in a rating for user affect if two DSAs provide similar value. Similarly it may choose not to act if the two sources disagree. Integration plans can also be used to augment one type of model data with data from another. In the case of Freudbot, conversational engagement classifications provided by one DSA are used the algorithms which determine user behaviours reported by another DSA.

5.5 Data Source Agents - DSAs

5.5.1 Constructing a DSA

A template DSA has been created which handles announcing available data streams, accepting subscriptions and publishing data. It also supports connecting to an HTTP listener built into the environment (see below) as an interface to external devices, if this is required. To use this template the developer needs to provide the following as a minimum:

- Write the data processing plan which buffers and converts incoming data and provides one or more information channels

- Write data interface plan to get read external data source (reading database, device data, etc.). If using HTTP listener, modify the listener registration plan to provide data type information.
- List data streams available for subscription in the agent's belief base.

5.5.2 Conversation-based DSAs

Implementation of the DSAs described in Chapter 4 followed the outline of steps described in the previous section to adapt a template DSA to provide the required information stream to the student modeling agent. In both cases, the data source is the conversational record, rather than a device. For a full production version of the system, the plan for receiving and processing source data would use a function to read a database. In the proof-of-concept system, this is simulated by messages sent from student representation agent, which handles all input and output to the user. Conversation log updates therefore arrive as messages which trigger a processing plan in the DSA.

5.5.2.1 *Conversational quality and appropriateness*

The input data processing plan accepts log data, CA and user utterances. The CA data is first processed using a custom internal action written to identify dialogue acts associated with Freudbot's output (e.g. greeting, various repair strategies when not understanding, educational content). The identified dialogue act, along with the student's input text, are passed to another internal action which instantiates the conversational quality and appropriateness classifier and returns a simple value for each. As described in section 4.3.1 a value of '1' is considered good quality or appropriateness. A value of 3 or 4 indicates low quality or appropriateness.

These two information streams are listed in the belief base of the agent. The template functions automatically announce the availability of these information streams to the student model agent, handle any resulting subscriptions requests, and publishes data to subscribers as it is processed.

5.5.2.2 *Conversational behaviour*

Similar to the conversation quality DSA, this agent accepts incoming messages containing conversation log data. In this case, the adjacency pairs are the user input and the CA response to that input, so that the agent can judge whether the user's behaviour is resulting in "good" responses from the CA, i.e. information about Sigmund Freud.

Instead of a machine learning classifier, this agent relies on a series of tests and rules to identify different common user behaviours. This is similar to an expert system approach where the rules have been programmed in and tweaked manually to produce the desired result.

An examination of the logs of past studies using Freudbot revealed three recurring patterns of user behaviour:

1. ***Tryer***: The user attempts to ask questions exactly as one would hope they would, using full sentences (or close) on topics related to Freud. They continue to do this despite little or no success in getting Freud-related information from the CA. This *trying* behaviour is characterized by relatively long sentences, high number of no-match cases per inputs and possibly input words with high abstractness value, a measure of cognitive engagement (Wen et al., 2014).

2. **Keyworder**: The user answers questions or responds to bot output with single words or phrases associated with Freud or psychoanalysis. E.g. "ego", "psychoanalysis", "anxiety". Typically jumping from one topic to the next. This *keywording* behaviour could be detected by short inputs (number of words input), non-repetition, low number of no-match cases per inputs, and possibly low abstractness value of input words.
3. **Morer**: The user discovers a word that leads to advancement through the narrative and repeats that word. For example, just keeps saying "ok". *Moreing* behaviour could be detected by recognizing backchannel type words and phrases ("more", "ok", "I see"), and high consecutive repetition of those words.

Detection

Users may exhibit more than one of these behaviours. They may start off *trying* and eventually give up and start *moreing*. Or they might just stick with one strategy, like *keywording* and never experience a proper conversation. Often, these behaviours come about as a result of poor performance on the part of the CA, and the student attempting to find a strategy that results in useful information being returned.

Again custom internal actions have been programmed to identify certain user dialogue acts, such as *backchannel* comments, which are used in conversation to indicate that one is following along and encouraging the other conversational partner to continue (e.g. "Okay", "I see", "uh huh"). Freudbot is programmed to recognize these phrases and continue the narrative associated with the current topic. The agent keeps a history of the

use of these words and determines if consecutive repeated use of the same term has been used.

In a similar way, tests are carried out to see if the user is a *tryer*, indicated by the use of longer sentences, suggesting complex questions or comments, followed by repair statements from the CA indicating it does not understand the user input. The poor performance of the CA is an important aspect because an intervention is not required if the CA is successfully responding to the user input with appropriate educational content. Again if occurrences of this situation exceed a threshold, the associated data is published by the DSA and received by the model agent. Another set of tests detect potential *keyworder* behaviour.

In each case, if the behaviour is detected enough times to exceed predetermined thresholds, the appropriate user label – *tryer*, *morer*, or *keyworder* – is applied and this determination is published to the information stream, for the student model agent to collect, possibly integrate with other data, such as the conversation quality, and determine if it should be passed on to the CA representation agent.

5.5.2.3 Interventions

Each of the different behaviour types described above has an issue associated which determines the appropriate type of intervention to be applied in each case.

The problem, simply put, is that the student is either not managing to get to the CA content, as in the case of the *tryer*, or is not doing so through a conversational approach (*morer* and *keyworder*). The first type of problem is very bad, the second simply doesn't make use of the conversational capabilities of the CA. Although *morer* behaviour does

expose significant Freud content, it is not much different than reading a book. *Keywording* is similar to using a search engine. Both cases leave little motivation for the user to interact again. Both cases would likely result in a poor rating of the CA.

If the user is able to obtain content through a conversational approach, then there is no need to change anything. The user is left in control of the conversation. If the system can recognize that the user is having trouble obtaining content through a conversational approach, i.e. a *tryer*, the CA representation agent can address this by taking some control of the conversation in an attempt to introduce relevant topics. This may not be as good as when the user can drive things, but is preferable to the student having to resort to other behaviours to obtain useful information, such as just saying 'yes' (*moreing*), using non-conversational input like keywords, or simply doesn't get content.

In the case of *trying* behaviour, when a no-match condition is signaled by the CA to CA-REP, the BB information will cause CA-REP to trigger a plan which will output a new default responses. (E.g. "I don't seem to be doing very well in trying to understand your comments and questions. If I can ask, are you more interested in my theories, or in my life?"). If the user responds to this then the CA will recognize the user response to the question and suggest an appropriate topic (theories, life/people, or both depending on stated preference). Additionally, future "no-match" responses will favour repair strategies that suggest topics related to the user's interest, or ask leading questions related to the user's interest. These are repair strategies that take away some of the control of the conversation from the student, but are more likely to result in information being delivered.

In the case of *keywording* behaviour, a normal response from the CA will, because of the information in the belief base, trigger a plan in CA-REP that will output the response,

but also append some information, instruction, or questions. ("I can't help noticing you have a somewhat abrupt conversational style. In any case, you can ask me to tell you more about a topic if you'd like to go into more depth."). The intention is to at least encourage the student to use conversational directives to experience the narrative structure and appreciate the depth of the content, rather than just seeing the first section of each topic.

In the case of *moreing* behaviour, the process of triggering an intervention is the same as for *keywording*, i.e. it adds to a normal response, rather than wait for a "no match" response. It informs the student "You seem to be advancing the conversation by repeating the same word. This does allow you to cover a topic thoroughly, but remember that you can branch off to other topics ('Tell me about...') and come back to a topic ('Tell me more about...')." Again the intention is to provide the student with other ways to interact and encourage them to do so in a conversational way.

A secondary potential benefit of the interventions is to suggest that the CA has some level of awareness (of the user's behaviour) and therefore promote a sense of social presence.

5.6 System Evaluation

5.6.1 Achievement of proof-of-concept goals

5.6.1.1 Demonstrate the process of integrating an existing CA to the agent-based framework

The CA representation agent was successfully integrated with the existing Freudbot CA. The connection was tested to be reliable (no messages lost) whether or not the CA was running on the same server as the agent system, or on a remote system. The agent

system provided a conduit between the user and the CA which was transparent to the student. The system was able to intercept events reported by the CA and react in a timely manner to modify the CA's dialogue when appropriate.

5.6.1.2 Provide a platform to test and evaluate the performance of an implementation

The implemented system was used to test latency and computer resource impact as described in section 5.6.2.

5.6.1.3 Provide a test platform to collect live data from students interacting with the system

The system was used successfully to conduct the study and collect data from 56 participants. Log and survey data was captured reliably. Chapter 6 describes the methodology and results of the study.

5.6.1.4 Demonstrate the ability to add or add intelligence and change functionality of the CA

The new strategy was implemented successfully. When an unrecognizable input was detected by the CA the representative agent employed a new algorithm to select an appropriate response based on the history of the non-matches maintained by the agent. This required minimal changes to the CA as the logic was implemented in the agent. This demonstrates the power of the agent-based approach and the potential for extending CA capability without having to modify the CA heavily.

5.6.2 Performance

The primary concerns for performance are that 1) the system does not introduce a time delay that will negatively impact the user experience, and 2) the system is scalable in

terms of load on CPU and memory resources. Additional delay was measured by introducing code to report timestamps at user input and CA response events and subtracting the response times reported by the CA server.

Though an insufficient number of samples to report statistics, observed time delays added by the system did not exceed 500 msec with conservative testing (verbose diagnostic output, testing on a low-powered desktop computer). It is expected this will improve with further optimization of the agents and removal of diagnostic output.

Chapter VI – User Testing and Discussion

6.1 Description of User Testing Methodology

To validate the system, and collect feedback for future development, 56 volunteer student participants chatted with an existing CA, Freudbot, enhanced with the proof-of-concept implementation of the agent-based framework described in Chapter 5. Immediately after chatting the students completed an online survey designed to collect their feedback regarding the chatting experience. For practical purposes, all interactions

Chatting with Freudbot

Once you start the session, you can chat with Freudbot by typing your question or comment in the appropriate text window and pressing return, or clicking the 'Say' button.

Freudbot is programmed to respond to most natural language questions and comments about his concepts and theories, significant people in his life, and autobiographical events.

He is also capable of discussing many of these topics at greater lengths if you try to engage him.

Please pay attention to **spelling** and **typos**.

Freudbot responds best if your responses are restricted to a **single sentence without punctuation**.

Please enter your token and click 'begin'

Do not close your browser window.

You will be given the option to start the questionnaire after 10 minutes has elapsed.

Figure 19: Freudbot start page

were carried out remotely via a web interface. For this reason, only the text-based DSAs were used.

A total of 56 participants were recruited from a pool of undergraduate-level students enrolled in PSYC 289 “Psychology as a Natural Science”, an introductory Psychology course at Athabasca University. PSYC 289 students are given the option of participating in a research study for course credit, for the purpose of providing an experiential learning task associated with research methods.

Participants were required to chat with Freudbot for at least 10 minutes. No direction was provided in terms of what to talk about, though some basic instructions were provided to optimize the interaction, as shown in Figure 19. After 10 minutes a message was displayed below the chat window informing them that they could end the chat at any time by clicking the “End Conversation” button, which was displayed at the same time

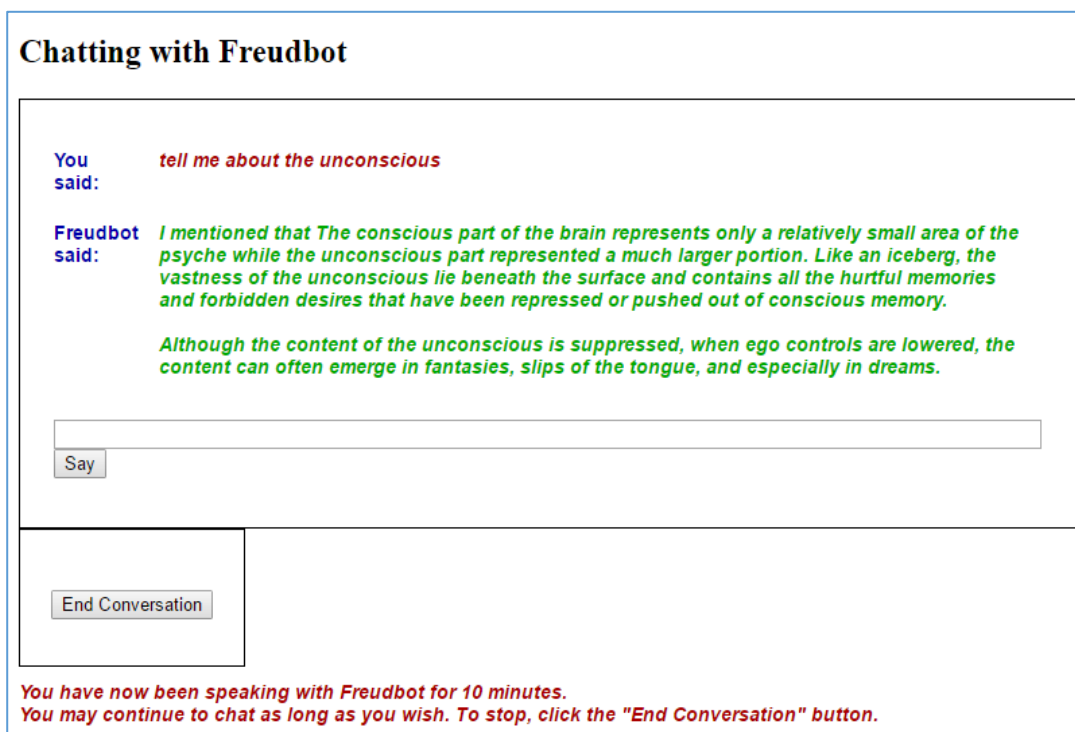


Figure 20: Freudbot interface with End Conversation button

(Figure 20). Clicking this button closed the chat interface and took the student to the online survey.

6.1.1 Questionnaire

The online survey is based on previously used questionnaires (Heller et al., 2005; Heller & Procter, 2009) that are designed to collect student feedback, attitudes, and advice about the experience of chatting with Freudbot, and CAs in general. The questions from the previous surveys were designed to measure user experience (“How engaging was this activity?”), and social presence (“How easy was it to pretend you were talking with Freudbot?”). For this study, questions were added to measure student reaction to the features specific to the agent-based version of Freudbot, i.e. the use of interventions and the modified strategy for handling cases where the CA does not understand the user input, including suggesting a new topic. These questions attempt to capture whether the student was aware of these features when they occurred, and how they rated them in terms of value and effectiveness. These features are all associated with how the CA controls the conversation and attempt to measure whether the feature was applied appropriately, and effectively. The participant was only asked to rate a feature if they first indicated that they had experienced it during their chat with Freudbot (e.g. “Did Freudbot suggest any topics to discuss?”). More detail is provided in section 6.3.2.2 and Appendix B.

There were a total of 42 questions, including 12 questions profiling the participants’ educational background, computer-related experience, and general demographic questions. The complete survey is shown in Appendix B.

The survey was conducted using LimeSurvey (www.limesurvey.org), an open-source tool for running online questionnaires. LimeSurvey was run on an AU Faculty of

Health and Social Sciences secure server. The recruitment system randomly assigned a unique token number for each user, which was used to identify participants' data within the CA chat logs (described below) and survey results. No personal information, e.g. names, student ID, email address, was provided to the agent test platform, or available to the researcher.

6.1.2 Chat logs

A log of the conversations between the participants and the CA was captured for analysis. This contains timestamped pairs of exchanges, arranged by conversation and identified by the same token that is used in the survey, allowing the chat log and survey data to be matched by user while maintaining anonymity. The logs used were automatically collected by the CA and stored on the same server which hosted the CA and the agent-based system.

6.2 Analysis

6.2.1 Purpose and expected outcomes

For the purpose of system evaluation, the study provided a means to assess the two decision support mechanisms provided by the agent system, which modify the behaviour of the CA so it could respond appropriately to detected user conversational behaviours. The first expected outcome was a measure of the actual and perceived effectiveness of the interventions which are triggered by the agent system to be carried out by the CA. A second important outcome was to measure the effect of no-match conditions, where the CA is unable to classify the user input to produce a meaningful response, and the perceived appropriateness of the strategy recommended by the agent system for the CA to employ.

The other important goal was to gain some insight into what factors are related to students perceiving the system as useful for learning, and which factors may drive the motivation to use it again.

The two sources of data for analysis are the conversational record, or chat logs, and the responses to survey questions.

6.2.2 Chatlog analysis

An analysis of the conversational record (chat logs) was carried out to

- Provide insight into the way students interact with a conversational agent in a learning context.
- Compare observed interaction and conversational behaviour from the chat logs (objective data) with user report data about the experience (subjective data).
- Test predictions about how student report outcomes are associated with certain conversational behaviours.
- Test predictions about how certain actions by the conversational agent are associated with student impressions of the exercise.
- Provide performance and usage data for developing strategies for conversation, improving existing ones, and future direction for improving the system.

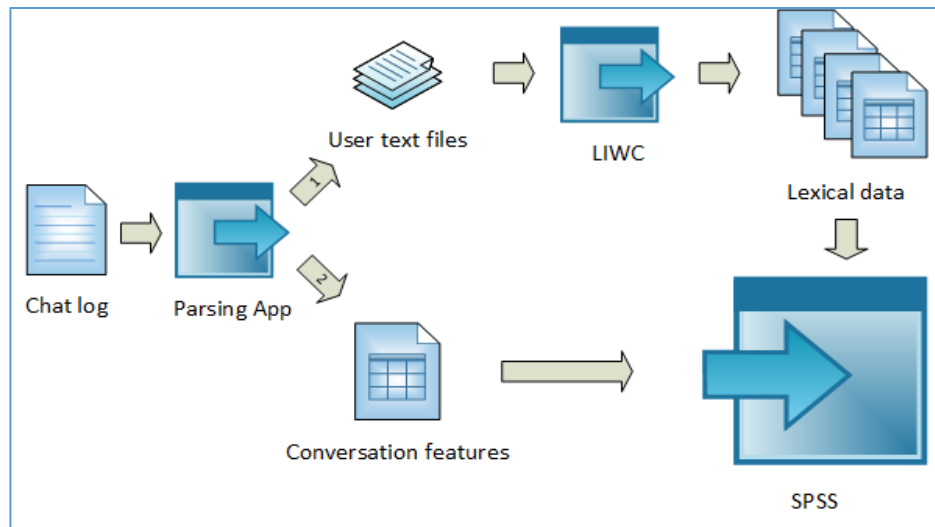


Figure 21: Conversation log processing

6.2.2.1 Processing chat logs

To prepare the log data for statistical analysis the text of the conversations was parsed by a custom application which created two sets of output, each of which would then be further analyzed (Figure 21).

1) LIWC input files

For each participant, a text file was generated by the parsing application, containing only the user's input, removing timestamps, CA output, and other extra text contained in the chat log. These files were further processed by the LIWC (Linguistic Inquiry and Word Count) tool, producing a lexical analysis of the user's text. The parsing application made some simple modifications to prepare the user text for input to this tool, in order to comply with the guidelines for LIWC. For example, because LIWC delimits sentences by ending punctuation, the application ensured that each sentence ended in a period, if punctuation had not already been applied by the user. The parsing application also inserted special delimiters in the text to identify where interventions made by the CA had taken place in

response to detected user behaviours. These delimiters used to segment the files for LIWC analysis, described later.

2) Conversational features

The second output provided by the chat parsing application consisted of summary data of some important features of the conversation. For each participant the following data was provided

- User token
- Number of exchanges before and after intervention type 1 (tryer)
- Number of exchanges before and after intervention type 2 (keyworder)
- Number of exchanges before and after intervention type 3 (morer)
- Number of no-match cases before and after intervention type 1
- Number of no-match cases before and after intervention type 2
- Number of no-match cases before and after intervention type 3
- Ratio of no-match cases before/after each intervention, to number of exchanges
- No. CA content deliveries before/after each intervention type
- Ratio CA content deliveries before/after each intervention to number of exchanges
- The order that interventions were delivered

The intention of generating these numbers is to explore whether interventions can affect either user behaviour, or change the user's experience interacting with the CA. The number of no-match cases, where the CA could not recognize the user input, and how many times the CA provided educational content, were predicted to be two of the most

important measures affecting the participant's satisfaction rating of the exercise. The numbers collected were expected to help measure what effect the interventions had on these values, if any. Measuring the number of exchanges – user input/CA response pairs – before and after the intervention allows for the analysis of the measures as a ratio to number of exchanges. It also provides an idea of where in the conversation the intervention took place.

The conversational feature data was formatted for input to a statistical package, with a row for each participant and columns for each of the data types listed above.

6.2.2.2 *Lexical analysis of user input*

The LIWC tool was used to analyze the individual files of user text provided by the log file parsing application. Information was provided about user text occurring before and after each type of intervention that took place, using the delimiters embedded by the chat parsing application. LIWC parses the text into sentences and words, identifying and counting different components, including words per sentence, number of words over 6 letters, and different parts of speech (nouns, verbs, articles, pronouns, etc), as well as linguistic dimensions, such as interrogatives, comparisons, and quantifiers. LIWC also uses proprietary dictionaries to classify words into different categories associated with psychological constructs, including affective, social, and cognition processes. Finally, the latest version, LIWC2015, generates counts of “summary language variables”: *analytical thinking*, *clout*, *authentic*, and *emotional tone*. Overall, LIWC generates over 90 output variables. A full list can be found in (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

LIWC outputs one line of 90+ variables for each segment of each input file. An input file represents one participant's input to the conversation, segmented at the point(s)

in the conversation where interventions from the CA took place. The first column identifies the input file name, which is the token id used to identify the student, allowing this data to be matched with student questionnaire and other chat log data from the log parsing application described previously.

Table 7 lists the LIWC output variables that were selected for analysis of the chat logs. These variables are intended to provide a measure of social presence and are based on those used in Heller & Procter (2014). The variable names changed slightly with the current (2015) version of LIWC used in this thesis.

Table 7: LIWC variables for chat log analysis

LIWC Variable	Description
WC	Word count
WPS	Words per sentence
Sixltr	Words>6 letters
Analytic	Analytical thinking (Pennebaker et al., 2014)
ppron	Personal pronouns
posemo	Positive emotion
negemo	Negative emotion
social	Social words (collection)
cogproc	Cognitive processes (collection)
percept	Perceptual processes (collection)
bio	Biological processes (collection)
relativ	Relativity
focuspast	Past focus
focuspresent	Present focus

LIWC was used to process the files four times, resulting in 4 sets of data formatted for use with a statistical package. One data file was produced for each intervention by configuring LIWC to segment the files by the delimiter associated with that intervention. This means that for participants who had multiple interventions, each intervention is treated separately. Therefore a participant who had two interventions ended up with two

sets of data, segmented before and after the intervention (two segments), rather than one set of data segmented into three parts (before first intervention, between first and second intervention, and after third intervention). Finally, a fourth data set was created by having LIWC segment conversations in half. This fourth file was generated to address the concern that users may modify their approach to interacting with the system over time, regardless of the presence of an intervention. Using this data, selecting those users that did not receive an intervention, provides a simple baseline for comparison with those participants that did receive intervention(s).

6.2.3 Survey data analysis

This section provides an overview of the nature of the survey data. Section 6.3 looks at the relationships between key survey and chatlog data. 13 male and 43 female students participated in the study. The modal value for distance education courses completed was 0 (22 participants) with the most student having completed no more than 5, and for undergraduate psychology courses completed was 0 (25 participants), the majority having complete 2 or less. The data collected from the questionnaires are primarily responses in the form of five-point Likert scale ratings (e.g. poor to excellent, not useful to very useful) and are ordinal or categorical in nature. These are broken down into categories associated with the perception of the experience, social presence, CA performance, and pedagogical utility,

6.2.3.1 *Overview – User experience measures*

Figure 22 shows the frequency distributions of a selection of variables representing a summary of user experience ratings. In general, ratings ranged from 1 to 5, with 1 representing the poorest rating, and 5 the highest rating. For example, for the question *How*

engaging was this activity? the ratings range from 1="Not engaging" to 5="Very engaging".

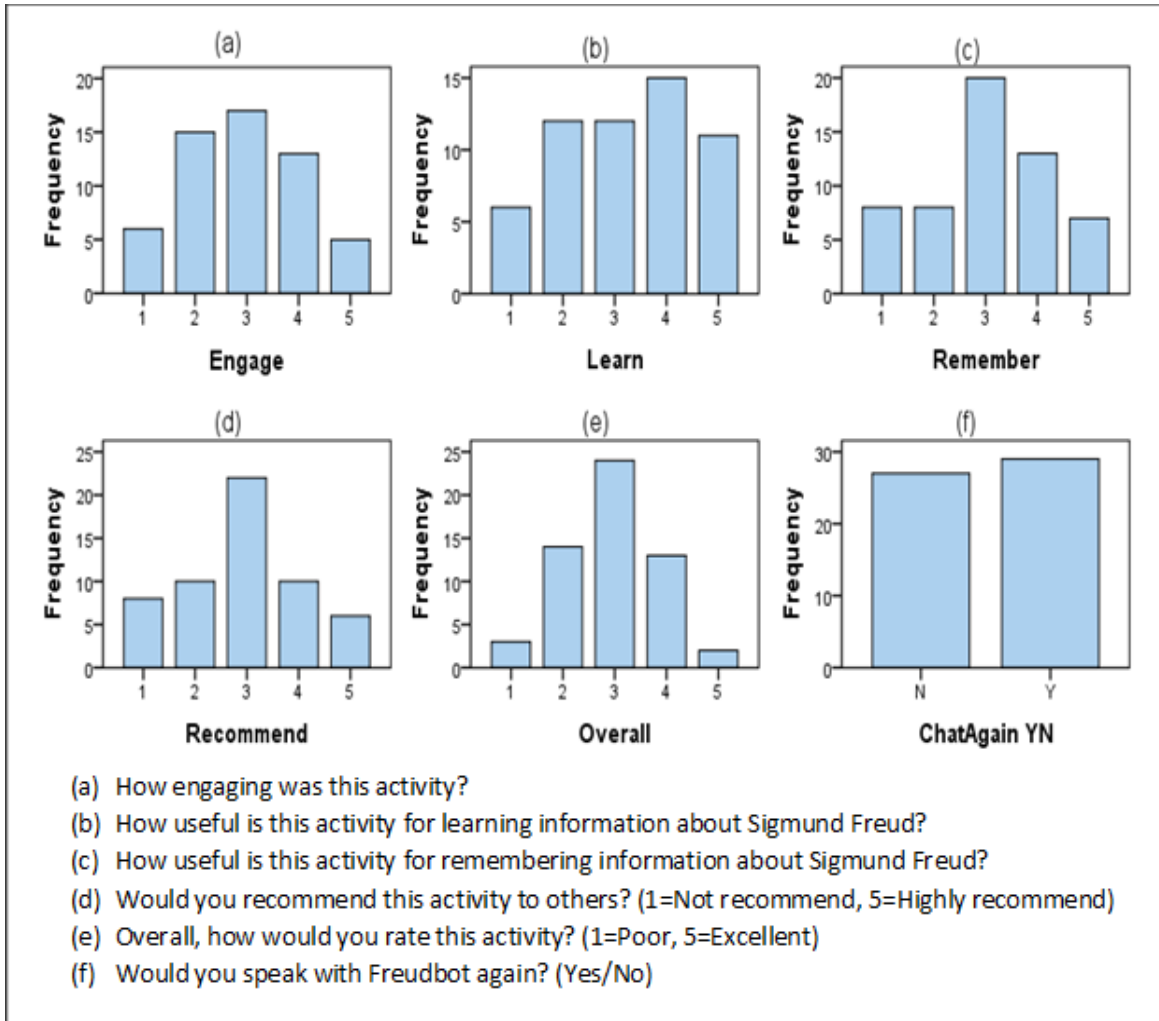


Figure 22: User experience frequency data

80% of participants rated the experience between 2 and 4 (mode=3) for *How engaging was this activity?* (Figure 22a). Participants appear to acknowledge the potential value of the CA as a learning tool with 46 of participants gave a rating of 4 or higher (mode=4) when asked *How useful is this activity for learning information about Sigmund*

Freud? but were less sure about the potential to retain the knowledge learned from this tool however (Figure 22b and c).

Three measures of overall satisfaction with the exercise yielded similar results. Most participants did not feel strongly one way or the other when asked *Would you recommend this activity to others?*, *Overall, how would you rate this activity?*, and *Would you speak with Freudbot again?* (Figure 22d,e,f). The dichotomous variable ChatAgain shows that the population is evenly split between positive and negative opinion towards the experience. A chi-square test between the Recommend and Overall variables with the ChatAgain variable yielded a Fisher’s Exact Test score of 24.756 ($p < .001$) and 16.719 ($p = 0.001$) respectively. Therefore the ChatAgain variable will be used as an overall satisfaction rating for further testing. The frequency tables for ChatAgain vs Recommend (Table 8a) and ChatAgain vs Overall (Table 8b) show how the counts are distributed.

Table 8: ChatAgain/Recommend/Overall frequency tables

<i>a) ChatAgain - Recommend</i>				<i>b) ChatAgain – Overall rating</i>					
	ChatAgain				ChatAgain				
	0	1	Total		0	1	Total		
Recommend	1	8	0	8	Overall	1	3	0	3
	2	8	2	10		2	10	4	14
	3	10	12	22		3	13	11	24
	4	1	9	10		4	1	12	13
	5	0	6	6		5	0	2	2
Total	27	29	56	Total	27	27	29		

6.2.3.2 Overview – Social presence measures

Social presence related measures are considered important because it was reasoned that there is a relationship between the degree to which the user feels that they are

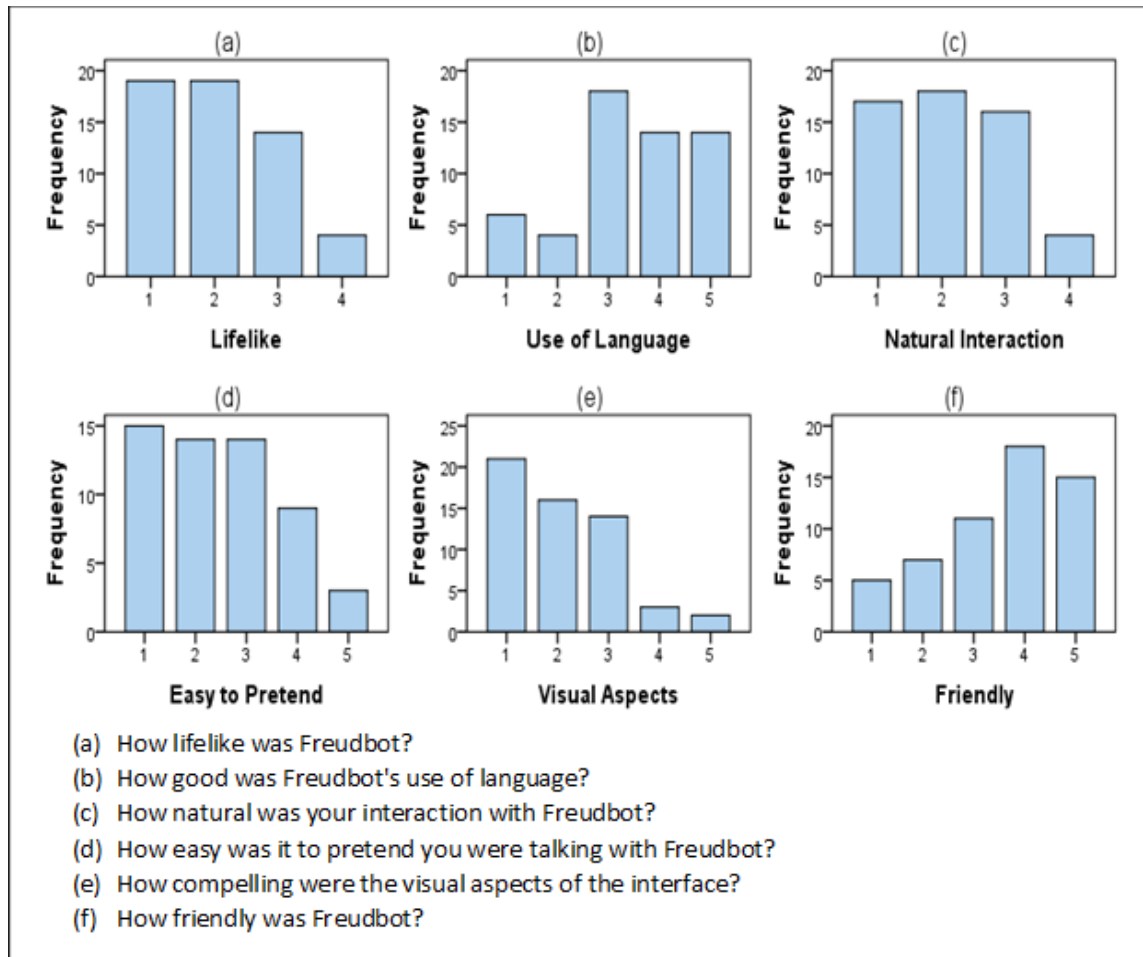


Figure 23: Social presence frequency data

interacting with a person and the likelihood of conversing in a human-like way, or what is referred to as *conversational engagement* in this thesis. Frequency data for results of questions associated with social presence are shown in Figure 23.

Participants did not judge Freudbot to be “lifelike” with 68% giving a rating below 3 and no one rating it at 5 (*How lifelike was Freudbot?* mode=2 median=2, 1=“not lifelike”, 5=“very lifelike”) (Figure 23a). This is not particularly surprising given the very

plain text-only interface that was used. Similar results (mode=1 median=2) were reported for *How compelling were the visual aspects of the interface?* (1="Not compelling", 5="Very compelling") (Figure 23e). More encouraging were the results for *How good was Freudbot's use of language?* (Figure 23b) with 82% of participants giving a rating of 3 or higher, and 25% giving a score of 5 (Mode=3 Median=3.5, 1="Poor", 5="Very good").

64% gave Freudbot a rating below 3 (Mode=2 Median=2, 1="Not natural", 5="Very natural") for *How natural was your interaction with Freudbot?* (Figure 23c). More disappointing, however, was that only 22% scored above 3 for *How easy was it to pretend they were talking with Freudbot?* (mode=1, median=2, 1="Not easy", 5="Very easy") while nearly twice as many rated it below 3 (Figure 23d). At odds with the lack of evidence for social presence indicated by these results, a significant number of participants attributed Freudbot with a friendly nature (Mode=4 Median=4, 1="Not friendly", 5="Very friendly") in response to *How friendly was Freudbot?* (Figure 23f).

6.2.3.3 Overview – Conversational performance measures

Performance measures focused on two factors: the CA's ability to handle cases where the user input was not understood, and the perceived usefulness of the interventions supplied by the CA when certain conversational behaviours were detected by the system. The frequency charts are shown in Figure 24.

The perception of the CA's ability to understand (*How well would you rate Freudbot's ability to understand?*) was judged to be below average by most participant with 52% giving a rating of 2 (1=Understood nothing, 5=Understood everything, mode=2, median=2). No participants gave a rating of 5 (Figure 24a). When asked *Overall, how*

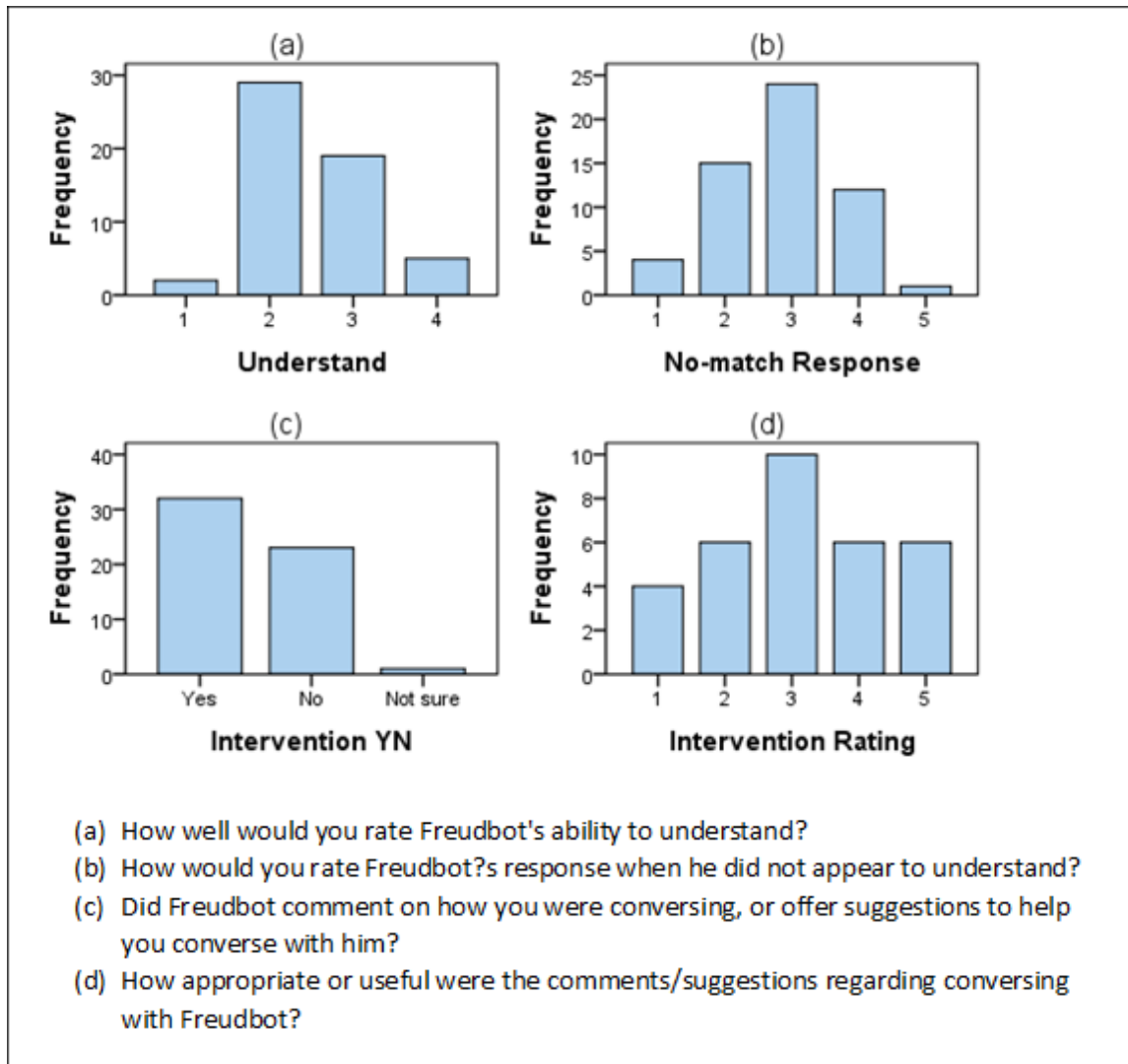


Figure 24: Performance measure frequency data

would you rate Freudbot's response when he did not appear to understand? (1=Poor, 5=Excellent), the median and mode response was 3 (43% of participants) as shown in Figure 24b. This is an important performance measure because it is both difficult to improve the ability of the CA to understand user input, and virtually impossible to completely remove all inability to understand. It is therefore important to handle these cases in a way that still suggests intelligence and supports the continuation of the conversation.

Figure 24c shows that 32 of 56 participants (57%) reported having received at least one of the three conversational interventions (*Did Freudbot comment on how you were conversing, or offer suggestions to help you converse with him?*). Those who did, rated the intervention, answering the question *How appropriate or useful were the comments/suggestions regarding conversing with Freudbot? (1=poor, 5=excellent)*. Ratings were evenly split between positive and negative with 10 ratings below, and 12 above the median, and mode, of 3 (Figure 24d). It's worth noting that chat log analysis shows that 43 participants actually received at least one intervention.

6.3 Findings

6.3.1 The efficacy of interventions

Three conversational behaviours are detected by the agent-based system. Each can result in an appropriate conversational intervention being triggered by the CA representational agent, which is responsible for providing decision support to the CA. The three behaviours and associated interventions are described briefly in Table 9.

Table 9: Behaviour types and interventions

Behaviour	Description	Intervention Number / Description
<i>Tryer</i>	Attempts to use proper conversation but CA does not match most input	1 CA apologizes and suggests topics based on user's area of interest (Freud's life or theories)
<i>Keyworder</i>	Does not attempt to converse. Enters single words or short phrases	2 Suggests conversational phrases to advance further into topics ("Tell me more about...")
<i>Morer</i>	Advances through topics by repeating the same "more" type word ("ok", "more", "go on")	3 Reminds user they can branch to other topics ("Tell me about") or come back to a topic ("Tell me more about...")

Table 10 gives a breakdown of how many of each type of intervention was delivered. Note that some participants had more than one intervention, and some did not receive any.

Table 10: Intervention frequencies

	Intervention 1 (tryer)	Intervention 2 (keyworder)	Intervention 3 (morer)	No Intervention
Participants	37	7	8	13

To answer the question whether or not the interventions provided by the system invoke a change in user conversational behaviour, or experience, the chat logs are examined before and after each intervention, testing for differences in key variables.

6.3.1.1 *Intervention 1 – ‘trying’*

For Intervention 1, the expected outcome is an improvement in the pedagogical utility of the experience, i.e. the delivery of more educational (Freud related) content. There is no attempt to change the user's behaviour because they are already conversationally engaged. The issue is that, generally due to the CA's shortcomings, the student is not being rewarded for their efforts with useful information. The strategy is to, as naturally as possible, take some control of the conversation and provide content, while maintaining a conversational approach. Ideally this approach should be modeled after the way a human would handle the same situation. Faced with questions from an interviewer but not understanding what they are getting at, one solution is to attempt to narrow down the area of interest, and then suggest a topic.

A paired-samples T-Test was carried out to compare Freud content and no-match counts before and after each intervention. The expected outcome is an increase in content

delivered and possibly a decrease in number of no-match conditions. Measures of this are provided by the conversational features dataset and the results (Table 11) show that there is a significant increase ($p < .001$) in content and a significant decrease ($p = .009$) in non-matches after the delivery of Intervention 1 (Pairs 1 and 4 respectively). The effect size for Freud content (Cohen's $d = .801$) suggests a large effect (Cohen, 1988). However, because the effect size for no-matches ($d = .453$) is small to medium, this result should be treated with some caution.

Table 11: Freud content and no-match counts before and after interventions

<i>Paired Samples Test</i>			Paired Differences				
	Mean	Std Dev	Mean	Std Dev	t	df	Sig (2-tailed)
Pair 1 frC1Bexch1B	.3789	.1422	-.1485	.1855	-4.867	36	.000
frC1Aexch1A	.5274	.1372					
Pair 4 nm1Bexch1B	.2720	.1778	.0770	.1698	2.758	36	.009
nm1Aexch1A	.1950	.1126					

6.3.1.2 Intervention 2 – 'keywording'

For Intervention 2, the desired outcome is to affect a change in behavior in the way the user is interacting with the CA, in order to boost conversational engagement, or at least provide the user with the feedback to allow them to try a more conversational approach. The motivation for this is that, while students who exhibit *keywording* behavior may enjoy some success in accessing the domain content, they are not taking full advantage of the capabilities of the interface, including the option to delve down deeper into topics, change topics, or ask analytical questions.

To measure potential changes in behavior, a lexical analysis of the user's input before and after the intervention is examined using the LIWC dataset. A paired-samples T-test was carried out to test for differences between social presence and other factors measured before and after the intervention. Two common traits of increased conversational engagement are the length of the sentences (WPS) and greater use of longer words (6ltr). The latest version of LIWC also provides a variable called “analytic”, a summary variable which indicates use of analytical words, based on research carried out by the authors of LIWC (Pennebaker et al., 2014). Social presence measures are also expected to increase

Table 12: LIWC social presence measures before/after intervention 2

		<i>Paired Samples Test</i>				
		Paired Differences		t	df	Sig (2-tailed)
	LIWC Variable	Mean	Std Dev			
Pair 1	Per-pronoun	-5.0529	8.3544	-1.600	6	.161
Pair 2	Pos-emotion	-2.1100	7.9004	-.707	6	.506
Pair 3	Neg-emotion	2.3671	6.6544	.941	6	.383
Pair 4	Social	-1.1957	7.1964	-.440	6	.676
Pair 5	Cog-proc	4.4886	4.0608	2.924	6	.026
Pair 6	Perception	.6500	1.3737	1.252	6	.257
Pair 7	Bio	1.6300	3.8934	1.108	6	.310
Pair 8	Relativity	-1.0700	8.9956	-.315	6	.764
Pair 9	Focus-past	-.3600	2.6206	-.363	6	.729
Pair 10	Focus-present	-.7757	8.6712	-.237	6	.821
Pair 11	Authentic	-5.5986	26.3346	-.562	6	.594
Pair 12	WPS	.3614	.7926	1.206	6	.273
Pair 13	Six-letter	1.3400	12.3409	.287	6	.784

with conversational engagement because the higher the degree to which the student views the CA as an intelligent presence, the more they are expected to converse with it in a human-like way. A low social presence rating would suggest that the student does not differentiate the CA from a data base query application or search engine. The selection of LIWC output variables considered to be associated with an increased perception of social presence is based on Heller & Procter (2014) and Kramer, Oh, & Fussell (2006).

Table 12 shows no significant changes, with the exception of cognitive processes (Pair 5), which actually increased (not shown). There is no evidence that the interventions for *keywording* behaviour had an effect on measures of social presence.

6.3.1.3 Intervention 3 – ‘moreing’

As with Intervention 2, the intervention for *morer* behaviour is expected to result in a slight modification to the way the student interacts with the CA. Specifically, it encourages the student to add some conversational acts in order to drive how the content is delivered rather than relying on the systematic, ordered output of the narratives of each topic. The justification for this is that it involves a higher cognitive process to consider different branches in the structure of the topics. *Morer* behaviour, the simple repetition of the same backchannel word, such as "okay", is not very different from simply reading a text book, document, or web content that does not contain hyperlinks. If this is the preferred way to receive information about a topic, there is little incentive for the student to use this tool.

Table 13: LIWC social presence measures before/after intervention 3

Paired Samples Test

	LIWC Variable	Paired Differences		t	df	Sig (2-tailed)
		Mean	Std Dev			
Pair 1	Per-pronoun	-8.3888	5.5038	-4.311	7	.004
Pair 2	Pos-emotion	-1.8563	4.8861	-1.075	7	.318
Pair 3	Neg-emotion	1.9150	5.5921	.969	7	.365
Pair 4	Social	-2.7788	10.9271	-.719	7	.495
Pair 5	Cog-proc	-1.0063	10.1694	-.280	7	.788
Pair 6	Perception	.6825	.8967	2.153	7	.068
Pair 7	Bio	1.0650	2.0265	1.486	7	.181
Pair 8	Relativity	2.1888	12.8729	.481	7	.645
Pair 9	Focus-past	.3138	5.8913	.151	7	.885
Pair 10	Focus-present	1.6538	6.7253	.696	7	.509
Pair 11	Authentic	-28.1250	54.2775	-1.466	7	.186
Pair 12	WPS	-.6075	2.0909	-.822	7	.438
Pair 13	Six-letter	.5163	7.7920	.187	7	.857

The same LIWC variables as those for Intervention 2 were used to detect if the intervention was successful in encouraging the desired change in conversational behaviour. Again, paired-samples T-test did not reveal statistically significant differences between social presence and other metrics collected before and after the intervention (Table 13), with the exception of personal pronouns, which showed a significant increase (means not shown, $p=.004$). There is insufficient evidence to reject the null hypothesis.

6.3.1.4 Examination of the conversational record

LIWC analysis did not find support for the hypothesis that the interventions for *keyworders* and *morers* would have a positive influence on social presence, as a dimension of conversational engagement. Unfortunately there were fewer cases of each (8 *morers*, 7 *keyworders*) which made statistical analysis challenging. One possible explanation for the relatively small number of cases of these types of behaviour is that the participants are Psychology students who are taking part in a study for the purpose of learning about research methods. As such, they may be more inclined to take the task seriously and attempt to ask meaningful questions. This is more in line with the profile of *tryer* behaviour.

There are a few explanations for why no changes were detected in user behaviour as a result of Intervention 2 and 3. These are:

- The interventions were not effective and changes did not take place
- The interventions were inappropriately applied, i.e. *keyworder* and *morer* detection was inaccurate
- The LIWC metrics selected were not sufficiently sensitive to detect changes, particularly given the small sample sizes for these two intervention types

A manual examination of the log files was carried out to confirm whether or not changes in user behaviour were observed, but not detected, as a result of the interventions. For each participant who received either Intervention 2 or Intervention 3, a simple qualitative analysis was carried out to

- identify the type of conversational behaviour that preceded the intervention

- verify whether the user's conversational behaviour justified the intervention
- determine if the intervention resulted in an appropriate change in behaviour

The results of the qualitative analysis were generally encouraging. Of those participants receiving Intervention 2, all but one were judged to be showing *keywording* behaviour, meaning that the intervention was justified. Five (of 7) participants responded to the intervention by attempting at least a few full, though sometimes short, sentences. One of these reverted back to non-conversational behavior. This appeared to be a result of poor performance on the part of the CA. This suggests that the intervention is successful in motivating the user to explore the option to converse with the CA, but still requires the CA to do its part by rewarding the user with an improved experience, i.e. provide domain information.

Similar results were found with participants who received Intervention 3. All users exhibited *moreing* behaviour, justifying the intervention. Seven out of 8 cases resulted in a change of behaviour. One was inconclusive as it occurred at the end of the conversation. Of those that did switch to full sentences after the intervention, 2 reverted back to *moreing* after attempts were met with no-match responses.

In addition to providing some support for the hypotheses, the qualitative analysis of the chat logs suggests some additional directions in terms of analysis of this data. These are discussed in the Future Work section of Chapter 7.

6.3.2 Factors associated with perceived usefulness

As stated earlier, the ChatAgain variable correlated well with the other two overall measures of user satisfaction (Overall and Recommend). ChatAgain was selected as a

dependant measure of user satisfaction which provides a dichotomous measure of the important, definitive question: will the student use the CA again?

6.3.2.1 *Interaction experience and social presence*

To understand which interaction experience and social presence variables may be associated with the decision to chat again, a Mann Whitney U test was carried out for the experience-related questions using ChatAgain as a grouping variable. The Mann-Whitney U test is a non-parametric evaluation of the similarity of the distributions of two groups. The required assumptions for this test are met: the data is ordinal, the observations from both groups are independent, and the null hypothesis that the two distributions are equal is reasonably assumed. The two groups are based on the response (yes=1 or no=0) to the question: *Would you speak with Freudbot again? (ChatAgain)*.

It is expected that all the social presence variables that will associate with ChatAgain (mean rank will be higher for “yes” than for “no”): how lifelike Freudbot was perceived to be, how well the CA used language, how natural the interaction seemed, how easy it was to pretend they were talking to Freud, and how friendly Freudbot was seen to be. A possible exception would be “how compelling were the visual aspects” because a simple text interface was used. This variable is typically more relevant to an embodied conversational agent.

Table 14 shows that the all social presence variables, with the exception of Visual Aspects, do show a significant increase ($p < .05$ 1-tailed) in mean rank, suggesting that students who rate these measures more highly are more likely to choose to speak with Freudbot again. Surprisingly, Use of Language had the highest p-value, and was not

Table 14: Social presence ratings association to Chat Again

Ranks – Social Presence

	ChatAgain	N	Mean Rank	Sum of Ranks
Lifelike	0	27	19.81	535.00
	1	29	36.59	1061.00
	Total	56		
Use of Language	0	27	24.76	668.50
	1	29	31.98	927.50
	Total	56		
Natural Interaction	0	26	20.38	530.00
	1	29	34.83	1010.00
	Total	55		
Easy to Pretend	0	26	21.23	552.00
	1	29	34.07	988.00
	Total	55		
Visual Aspects	0	27	25.07	677.00
	1	29	31.69	919.00
	Total	56		
Friendly	0	27	22.19	599.00
	1	29	34.38	997.00
	Total	56		

Test Statistics^a- Social Presence

	Lifelike	Use of Language	Natural Interaction	Easy to Pretend	Visual Aspects	Friendly
Mann-Whitney U	157.000	290.500	179.000	201.000	299.000	221.000
Exact Sig. (2-tailed)	.000	.087	.000	.002	.113	.003
Exact Sig. (1-tailed)	.000	.043	.000	.001	.057	.002

a. Grouping Variable: ChatAgain

significant for $p < .05$ for the 2-tailed test. The researcher’s intuition is that language use would have one of the stronger effects in determining user acceptance. Frequency information for language use reveals that most participants rated Freudbot high for this

category (82% rated 3 or higher, with 25% each giving a rating of 4 and 5). This means that even people who respond no to chatting again rating Freudbot high for language use, which explains the barely significant difference in mean rank.

For experience ratings, again it is expected that all variables will increase in mean rank for the case where ChatAgain = Yes. Table 15 confirms this is true ($p < .0001$) for all

Table 15: Usage experience ratings association with Chat Again

Ranks – Experience Measures

	ChatAgain	N	Mean Rank	Sum of Ranks
Enjoyable	0	27	20.59	556.00
	1	29	35.86	1040.00
	Total	56		
Engage	0	27	19.43	524.50
	1	29	36.95	1071.50
	Total	56		
Easy Activity	0	27	26.17	706.50
	1	29	30.67	889.50
	Total	56		
Learn	0	27	20.67	558.00
	1	29	35.79	1038.00
	Total	56		
Remember	0	27	20.70	559.00
	1	29	35.76	1037.00
	Total	56		

Test Statistics^a – Experience Measures

	Enjoyable	Engage	Easy Activity	Learn	Remember
Mann-Whitney U	178.000	146.500	328.500	180.000	181.000
Exact Sig. (2-tailed)	.000	.000	.287	.000	.000
Exact Sig. (1-tailed)	.000	.000	.145	.000	.000

a. Grouping Variable: ChatAgain

measures except Easy Activity. Again, frequency data show that only 10% of participants rated the activity below 3 in terms of how easy it was. Most people found it easy, regardless of whether they would chat again. As with Use of Language, this does not mean that this metric is not important in general. If the activity was deemed to be difficult it could affect where people would chat again. But for this exercise it is not a discriminating factor.

6.3.2.2 *Conversational control ratings (Intervention, No-match response, Topic suggestions)*

Three survey questions were classified as *conversational control* measures. That is, these questions relate to the users' perceptions of how well the CA does along these three dimensions of affecting the flow of the conversation.

Intervention strategies have been discussed and are an important part of encouraging the student to interact appropriately to gain the full benefit of the conversational interface, or to help ensure that educational content is delivered. How effective these strategies are depends on how appropriate they are perceived to be.

No-match response ratings – answers to the question *Overall, how would you rate Freudbot's response when he did not appear to understand?* – are expected to influence the likelihood of the user choosing to chat again. Strategies for handling the all too common case of the CA unable to understand the user's input are key to reducing frustration, and helping to ensure that domain content is still imparted to the student.

Ratings for *How would you rate Freudbot's choice of topics?* were expected to be important to the overall rating of the CA because this an important both in helping the student with ideas of what to talk about, facilitating the conversation, and in handling

situations where a change in topic is an appropriate repair strategy i.e. getting a conversation back on track.

Table 16 confirms that users who choose to chat again also have a tendency to rate these measures highly. Mean ranks for these ratings are all significantly ($p < .005$) higher for the ChatAgain = Yes response. Further research would have to be carried out to determine if there is a causal relationship but result is reassuring.

Table 16: Conversation control ratings association with Chat Again

<i>Ranks</i>				
	ChatAgain	N	Mean Rank	Sum of Ranks
Intervention Rating	0	15	10.43	156.50
	1	17	21.85	371.50
	Total	32		
No-match Response	0	27	21.67	585.00
	1	29	34.86	1011.00
	Total	56		
Topic Suggest Rating	0	23	15.67	360.50
	1	25	32.62	815.50
	Total	48		

<i>Test Statistics^a</i>			
	Intervention Rating	No-match Response	Topic Suggest Rating
Mann-Whitney U	36.500	207.000	84.500
Exact Sig. (2-tailed)	.000	.001	.000
Exact Sig. (1-tailed)	.000	.001	.000

a. Grouping Variable: ChatAgain

b. Not corrected for ties.

6.3.2.3 Chatlog observations (content, no-match count)

Chatlog observations on the amount of content delivered and the number of no-match situations (both as a ratio to the number of exchanges) were compared participants' responses to the ChatAgain question. The hypothesis is that the users who experience a lot Freud-related conversation from the CA will be more likely to respond Yes to chatting again. Similarly, the users who experience the most cases where Freudbot cannot understand their input are the most likely to answer No to chatting again.

Since the no-match and content data are both continuous variables, and ChatAgain is a dichotomous ordinal variable, a point-biserial Pearson's correlation was used to test the hypothesis. Table 17 shows that while total no-match cases (nmTexchT) are significantly negatively correlated ($p < .01$) with ChatAgain (No=0, Yes=1), delivery of Freud content (frCTexchT) is not significant. It would appear that the frustration of no-match cases is particularly significant to user satisfaction. It may also be that the amount of content delivered is not as important to user satisfaction as the patterns in which it is delivered. For example, a series of successful exchanges with the CA – several questions and/or statements in a row all resulting in Freudbot providing useful information – may have a larger positive impact on the user than the same amount of content dispersed between multiple non-match exchanges. Testing for consecutive content or other patterns may reveal something.

Table 17: Content and no-match measures association with Chat Again

	<i>Correlation with ChatAgain (Pearson's r)</i>		
	r	Sig. (2-tailed)	N
nmTexchT	-.347**	.009	56
frCTexchT	.115	.397	56

** . Correlation is significant at the 0.01 level (2-tailed).

Chapter VII – Conclusions and Future Work

7.1 Discussion and Conclusions

The overall goal of this research is to propose an approach to improve the way that students interact with conversation-based e-learning applications. It attempts to do this through three contributions:

Design an adaptable agent-based framework for the purpose of improving interactions with conversation-based learning applications. Agents are used to represent each of the partners in the conversation: the student and the CA. Other agents model relevant characteristics of the student so that the CA can manage the conversation appropriately, and model characteristics of the CA to provide a more engaging virtual character. Data source agents represent the different information sources to maintain these models, allowing the system to adapt to whatever hardware or software is present for these purposes. There has been a significant amount of research over the last two decades into how to make pedagogical agents more effective by modeling human characteristics. This includes affective computing (Picard, 1997), embodiment (Cassell, 2001), and modeling personality (Mairesse et al., 2007). This framework extends this research, with the goal of providing an adaptable system, using agents, to implement appropriate models for a variety of types of CA's and differing student learning environments.

The framework presented in the thesis can be used as a guide for implementing systems that provide pedagogical conversational agents the ability to collect and use any

kind of information about the student that is relevant to that CA, Agent roles and functions have been defined along with the protocols required for communication between the agents. The proof-of-concept implementation acts as an extensible example that supports remote access of multiple concurrent users. The implemented agents can be used as templates that employ the framework protocols.

Introduce a new approach to detecting user engagement based on judging conversational behaviour detected from the ongoing transcript of the interaction. A machine-learning classification of conversational quality was investigated, as well as a heuristic-based algorithm for detecting certain user behaviour types. Each of these was implemented as an agent that provided real-time analysis of the interaction between the student and the CA. Conversational interventions were designed to apply to each of these behaviour types, and successfully triggered by the appropriate agent when they detected. This proposed use of conversational quality is a new approach to detecting engagement which is particularly well suited to conversation-based applications. This overlaps somewhat with the area of conversational analysis and the study of adjacency pairs to evaluate conversational responses (Beun, 2001; Boyer et al., 2009). It may be possible to apply these techniques to real-time evaluation of engagement in conversations. The idea of comparing user responses to CA output is discussed further under Future Development in Section 7.2.3.

The implemented data source agents provide a default measure of engagement that relies only on the conversational record. These also demonstrate the implementation of DSAs as defined in the framework. The implemented interventions provide an example of

how the CA representation agent can be integrated with and provide intelligence and decision making support to an existing CA.

Collect observational and self-report data from user testing of a proof-of-concept implementation of the framework. 56 volunteer student participants chatted with the CA using the agent-based framework, followed by a short survey to gauge users' evaluation of experience.

The evidence for the effectiveness of the interventions, collected from the conversational log files, was encouraging. In the majority of cases, the intervention resulted in at least a temporary change of behaviour by the user to encourage better use of conversation with the CA. In other cases the system successfully modified the conversational strategy of the CA to address an underlying issue such as poor delivery of the domain content due to a failure on the part of the CA to interpret the user input.

Overall, the self-report data collected from the surveys provided mixed results. The overall ratings were split with 50% positive and negative. User ratings for three conversational control mechanisms were examined. A correlation was found between participants who rated the interventions as useful and/or appropriate and those who stated they would be willing to chat with the CA again. Similar correlations to chat again were found for ratings of how the CA responded to no-match cases, and the selection of topics suggested by the CA to keep the conversation going.

Other data collected from the user testing is discussed in the next sections.

7.2 Future Directions for Research and Development

7.2.1 Further analysis

By design, the survey collected information outside of the scope of this thesis with the intention of providing data for future research and analysis. There were several opportunities for participants to provide free form comments about best and worst features of the exercise, suggested enhancements, and other possible applications of CA technology. They are also invited to provide any final comments. A qualitative analysis of these responses may provide some understanding of the priorities that students have in evaluating these systems, and helps in distinguishing the degree to which ratings are based on the agent-based component or the underlying CA, Freudbot.

There are several other dimensions of the survey data which can be explored. Examples include understanding what factors are associated with ratings for how useful the system is for learning and remembering information about Sigmund Freud, and influence of participants' experience with chatbots or other CAs on performance ratings.

In addition there is much more that can be mined from the chatlogs. Motivated by the partial examination of the logs described in Chapter 6, it's expected that an in-depth qualitative analysis of the conversational record may reveal more sophisticated patterns in how student interact with the CA than can be found using automated lexical analysis tools such as LIWC. Veletsianos in particular (Veletsianos & Miller, 2008; Veletsianos & Russell, 2013) has explored qualitative methods for examining conversational records of pedagogical CAs to understand the nature of the interactions that take place. Further investigation of the approaches outlined in their research may provide some guidelines for analyzing Freudbot's chat logs.

7.2.2 Further research – Learning outcomes

The pilot study described here is useful for collecting data on how students interact with a CA and feedback on how it is perceived. The design of a controlled experiment is required to test what benefits, if any, are realized in terms of learning and retaining the subject matter. One possible direction for such a study would be to test the pedagogical value of conversation, particularly the effect of asking questions on comprehension and retention. A brief examination of the conversational record from this study shows that many students ask questions in response to the narrative output of Freudbot. This is natural behaviour during a conversation, displaying interest, and contributing to the dialogue. It's possible that a different level of cognitive processing associated with forming questions may have an effect on learning outcomes. Graesser, Person, & Hu (2002) promote the value of discourse for increasing comprehension. They point to past studies and reviews (e.g. King, 1992; Rosenshine, Meister, & Chapman, 1996) which have linked the act of students asking questions, while reading or attending lectures, to an increase in objective comprehension scores. It would be interesting to know if asking questions while conversing provides similar benefits. If so, it would help to justify interventions to improve the quality of the conversation when interacting with the CA.

7.2.3 Further development – Data source agents

Future development will be focused on improving the conversation based DSAs created for this study. The machine learning technique for determining conversational quality and appropriateness would benefit from additional annotated log data for training the classifier. Additional attributes could be explored using lexical analysis tools such as LIWC to preprocess the utterances, if it can be done in real-time. O'Shea's work with

dialogue act identification techniques and semantic processing to determine sentence similarity (O'Shea, 2012) could be implemented in a new DSA for processing adjacency pairs for testing similarity between CA output and user input. This would allow the CA to estimate whether the user's responses are related to what the CA is saying, allowing another method of judging the level of conversational engagement. This would supplement the techniques described in this thesis.

References

- Afzal, S., & Robinson, P. (2011). Designing for Automatic Affect Inference in Learning Environments. *Journal of Educational Technology & Society*, 14(4), 21–34.
- Asteriadis, S., Karpouzis, K., & Kollias, S. (2009). Feature Extraction and Selection for Inferring User Engagement in an HCI Environment. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (pp. 22–29). Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.aupac.lib.athabascau.ca/chapter/10.1007/978-3-642-02574-7_3
- Baker, R. S. J. d., D’Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223–241. <https://doi.org/10.1016/j.ijhcs.2009.12.003>
- Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4), 742–753. <https://doi.org/10.1016/j.dss.2012.05.024>
- Battaglino, C., & Bickmore, T. (2015). Increasing the engagement of conversational agents through co-constructed storytelling. In *Eighth Workshop on Intelligent Narrative Technologies*. Retrieved from <http://relationalagents.com/publications/int8-2015.pdf>

- Baylor, A. L., & Kim, Y. (2005). Simulating Instructional Roles through Pedagogical Agents. *International Journal of Artificial Intelligence in Education*, 15(2), 95–115.
- Becker, C., Kopp, S., & Wachsmuth, I. (2007). Why emotions should be integrated into conversational agents. *Conversational Informatics: An Engineering Approach*, 49–68.
- Bellotti, F., Berta, R., De Gloria, A., & Lavagnino, E. (2011). Towards a conversational agent architecture to favor knowledge discovery in serious games. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (p. 17:1–17:7). New York, NY, USA: ACM. <https://doi.org/10.1145/2071423.2071444>
- Beun, R.-J. (2001). On the generation of coherent dialogue: A computational approach. *Pragmatics & Cognition*, 9(1), 37–68. <https://doi.org/10.1075/pc.9.1.03beu>
- Bogdanovych, A., Trescak, T., & Simoff, S. (2016). What makes virtual agents believable? *Connection Science*, 28(1), 83–108. <https://doi.org/10.1080/09540091.2015.1130021>
- Bordini, R. H., Hübner, J. F., & Wooldridge, M. (2007). *Programming Multi-Agent Systems in AgentSpeak using Jason*. John Wiley & Sons.
- Boyer, K. E., Phillips, R., Ha, E. Y., Wallis, M. D., Vouk, M. A., & Lester, J. C. (2009). Modeling Dialogue Structure with Adjacency Pair Analysis and Hidden Markov Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (pp. 49–52). Stroudsburg, PA,

- USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1620853.1620869>
- Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, 28(3), 771–780. <https://doi.org/10.1016/j.chb.2011.11.020>
- Callejas, Z., López-Cózar, R., Ábalos, N., & Griol, D. (2011). Affective Conversational Agents: The Role of Personality and Emotion in Spoken Interactions. In D. Perez-Marin & I. Pascual-Nieto (Eds.), *Conversational Agents and Natural Language Interaction* (pp. 203–222). IGI Global. Retrieved from <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-617-6>
- Calvo, R. A., & D’Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 18–37.
- Cassell, J. (2001). Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine*, 22(4), 67.
- Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2004). BEAT: the Behavior Expression Animation Toolkit. In H. Prendinger & M. Ishizuka (Eds.), *Life-Like Characters* (pp. 163–185). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-662-08373-4_8
- Castellano, G., Pereira, A., Leite, I., Paiva, A., & McOwan, P. W. (2009). Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features. In *Proceedings of the 2009 International Conference on Multimodal*

- Interfaces* (pp. 119–126). New York, NY, USA: ACM.
<https://doi.org/10.1145/1647314.1647336>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 edition). Hillsdale, N.J: Routledge.
- Danforth, D. R., Procter, M., Chen, R., Johnson, M., & Heller, R. (2009). Development of virtual patient simulations for medical education. *Journal For Virtual Worlds Research*, 2(2). Retrieved from <https://jvwr-ojs-utexas.tdl.org/jvwr/index.php/jvwr/article/view/707>
- Dehn, D. M., & Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, 52(1), 1–22. <https://doi.org/10.1006/ijhc.1999.0325>
- Desmarais, M. C., & Baker, R. S. J. d. (2011). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1–2), 9–38. <https://doi.org/10.1007/s11257-011-9106-8>
- D’Mello, S., Craig, S., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner’s affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1), 45–80. <https://doi.org/10.1007/s11257-007-9037-6>
- D’mello, S., & Graesser, A. (2013). AutoTutor and Affective Autotutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers That Talk Back. *ACM Trans. Interact. Intell. Syst.*, 2(4), 23:1–23:39. <https://doi.org/10.1145/2395123.2395128>

- Ekman, P., & Rosenberg, E. L. (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.
- Endrass, B., Klimmt, C., Mehlmann, G., André, E., & Roth, C. (2014). Designing User-Character Dialog in Interactive Narratives: An Exploratory Experiment. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2), 166–173. <https://doi.org/10.1109/TCIAIG.2013.2290509>
- Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 715–724). New York, NY, USA: ACM. <https://doi.org/10.1145/1978942.1979046>
- Feidakis, M., Daradoumis, T., & Caballe, S. (2011). Endowing e-Learning Systems with Emotion Awareness. In *2011 Third International Conference on Intelligent Networking and Collaborative Systems (INCoS)* (pp. 68–75). <https://doi.org/10.1109/INCoS.2011.83>
- Gonzalez-Sanchez, J., Chavez-Echeagaray, M. E., Atkinson, R., & Burleson, W. (2011). ABE: An Agent-Based Software Architecture for a Multimodal Emotion Recognition Framework. In *2011 9th Working IEEE/IFIP Conference on Software Architecture (WICSA)* (pp. 187–193). <https://doi.org/10.1109/WICSA.2011.32>
- Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, Vol 3: Application to learning and teaching*. (pp. 451–473). Washington, DC, US: American Psychological Association.

- Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by Communicating in Natural Language With Conversational Agents. *Current Directions in Psychological Science*, 23(5), 374–380. <https://doi.org/10.1177/0963721414540680>
- Graesser, A. C., Person, N. K., & Hu, X. (2002). Improving Comprehension Through Discourse Processing. *New Directions for Teaching and Learning*, 2002(89), 33–44. <https://doi.org/10.1002/tl.45>
- Grant, M., Sandeep, V., & Fuhua, L. (2013). Integrating Multiagent Systems into Virtual Worlds. In *3rd International Conference on Multimedia Technology (ICMT-13)*. Atlantis Press. Retrieved from <http://www.atlantis-press.com/php/paper-details.php?id=10423>
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1), 10–18. <https://doi.org/10.1145/1656274.1656278>
- He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Heller, B. (2016). Conversational Agents as Historical Figures: Individual Differences and Perceptions of Agent and Social Presence (Vol. 2016, pp. 1368–1374). Presented at the EdMedia: World Conference on Educational Media and Technology. Retrieved from <https://0-www.learntechlib.org.aupac.lib.athabascau.ca/p/173133/>
- Heller, B., & Procter, M. (2011). Embodied and Embedded Intelligence: Actor Agents on Virtual Stages. *Intelligent and Adaptive Learning Systems: Technology Enhanced Support for Learners and Teachers*, 280–290.

- Heller, B., & Procter, M. (2014). Conversational Agents in Virtual Worlds: Immersion and the Conversational Record. *Inter-Disciplinary. Net. Retrieved from*. Retrieved from http://www.inter-disciplinary.net/at-the-interface/wp-content/uploads/2014/02/heller-procter_elvw4.pdf
- Heller, B., Procter, M., & Rose, C. (2016). Conversational Agents in Second Life. In S. Gregory, M. J. W. Lee, B. Dalgarno, & B. Tynan (Eds.), *Learning in Virtual Worlds* (pp. 153–166). Athabasca University Press.
- Heller, B., Procter, M., Mah, D., Jewell, L., & Cheung, B. (2005). Freudbot: An Investigation of Chatbot Technology in Distance Education. *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005*, 2005(1), 3913–3918.
- Heller, R., & Procter, M. (2009). Animated Pedagogical Agents: The Effect of Visual Information on a Historical Figure Application. *International Journal of Web-Based Learning and Teaching Technologies*, 4(1), 54–65. <https://doi.org/10.4018/jwltd.2009010104>
- Jaques, P. A., & Vicari, R. M. (2007). A BDI approach to infer student's emotions in an intelligent learning environment. *Computers & Education*, 49(2), 360–384. <https://doi.org/10.1016/j.compedu.2005.09.002>
- Johnson, W. L., Rickel, J. W., Lester, J. C., & others. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11(1), 47–78.

- Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263–286. <https://doi.org/10.2307/20445398>
- Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 677–682). New York, NY, USA: ACM. <https://doi.org/10.1145/1101149.1101300>
- King, A. (1992). Comparison of Self-Questioning, Summarizing, and Notetaking-Review as Strategies for Learning From Lectures. *American Educational Research Journal*, *29*(2), 303–323. <https://doi.org/10.3102/00028312029002303>
- Kirakowski, J., O'Donnell, P., & Yiu, A. (2007). The Perception of Artificial Intelligence as “Human” by Computer Users. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments* (pp. 376–384). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73110-8_40
- Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic Recognition of Non-Acted Affective Postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *41*(4), 1027–1038. <https://doi.org/10.1109/TSMCB.2010.2103557>
- Kramer, A. D. I., Oh, L. M., & Fussell, S. R. (2006). Using Linguistic Features to Measure Presence in Computer-mediated Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 913–916). New York, NY, USA: ACM. <https://doi.org/10.1145/1124772.1124907>

- Kumar, R., & Rosé, C. P. (2011). Architecture for Building Conversational Agents that Support Collaborative Learning. *IEEE Transactions on Learning Technologies*, 4(1), 21–34. <https://doi.org/10.1109/TLT.2010.41>
- Lester, J., Branting, K., & Mott, B. (2004). Conversational agents. *The Practical Handbook of Internet Computing*, 220–240.
- Lester, J. C., & Stone, B. A. (1997). Increasing believability in animated pedagogical agents. In *Proceedings of the first international conference on Autonomous agents* (pp. 16–21). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=269943>
- Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 125–132). New York, NY, USA: ACM. <https://doi.org/10.1145/604045.604067>
- Löckelt, M. (2011). Design and Implementation Issues for Convincing Conversational Agents. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*, 156.
- Luger, E., & Sellen, A. (2016). “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* (pp. 5286–5297). New York, NY, USA: ACM. <https://doi.org/10.1145/2858036.2858288>
- Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1), 457–500.

- Mao, X., & Li, Z. (2009). Implementing emotion-based user-aware e-learning. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (pp. 3787–3792). New York, NY, USA: ACM. <https://doi.org/10.1145/1520340.1520572>
- McClure, G., Chang, M., & Lin, F. (2013). MAS controlled NPCs in 3D virtual learning environment. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on* (pp. 1026–1033). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6727316
- Nakano, Y. I., & Ishii, R. (2010). Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (pp. 139–148). New York, NY, USA: ACM. <https://doi.org/10.1145/1719970.1719990>
- Norman, D. A. (1994). How Might People Interact with Agents. *Commun. ACM*, 37(7), 68–71. <https://doi.org/10.1145/176789.176796>
- Novielli, N. (2010). HMM modeling of user engagement in advice-giving dialogues. *Journal on Multimodal User Interfaces*, 3(1–2), 131–140. <https://doi.org/10.1007/s12193-009-0026-4>
- Nunamaker Jr., J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., & Patton, M. W. (2011). Embodied Conversational Agent--Based Kiosk for Automated Interviewing. *Journal of Management Information Systems*, 28(1), 17–48.
- O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6), 938–955. <https://doi.org/10.1002/asi.20801>

- Olney, A., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. (2003). Utterance Classification in AutoTutor. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2* (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1118894.1118895>
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions* (Vol. xi). New York, NY, US: Cambridge University Press.
- O'Shea, K. (2012). An approach to conversational agent design using semantic sentence similarity. *Applied Intelligence*, 37(4), 558–568. <https://doi.org/10.1007/s10489-012-0349-9>
- Padgham, L., & Winikoff, M. (2005). Prometheus: A Practical Agent-Oriented Methodology. In B. Henderson-Sellers & P. Giorgini (Eds.), *Agent-Oriented Methodologies*: IGI Global. Retrieved from <http://0-www.igi-global.com.aupac.lib.athabascau.ca/gateway/chapter/full-text-pdf/5057>
- Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring Emotions in Students' Learning and Performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, 36(1), 36–48.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/31333>
- Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, 9(12), e115844. <https://doi.org/10.1371/journal.pone.0115844>

- Perez-Marin, D., & Pascual-Nieto, I. (2011). Future Trends for Conversational Agents. In *Conversational Agents and Natural Language Interaction* (pp. 395–400). IGI Global. Retrieved from <http://0-www.igi-global.com.aupac.lib.athabascau.ca/gateway/chapter/54648>
- Picard, R. W. (1997). *Affective computing*. Cambridge, Mass.: MIT Press, c1997. Retrieved from <http://0-search.ebscohost.com.aupac.lib.athabascau.ca/login.aspx?direct=true&db=cat01422a&AN=aucat.b1147198&site=eds-live>
- Polhemus, L., Shih, L.-F., Swan, K., & Richardson, J. (2000). Building Affective Learning Community: Social Presence and Learning Engagement (pp. 800–802). Presented at the WebNet World Conference on the WWW and Internet, Association for the Advancement of Computing in Education (AACE). Retrieved from <https://www.learntechlib.org/p/6511/>
- Reeves, B., & Nass, C. (1996). The Media Equation: How people treat computers, television, and new media like real people and places. *CSLI Publications and Cambridge*. Retrieved from <http://www.humanityonline.com/docs/the%20media%20equation.pdf>
- Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching Students to Generate Questions: A Review of the Intervention Studies. *Review of Educational Research*, 66(2), 181–221. <https://doi.org/10.3102/00346543066002181>
- Rus, V., D’Mello, S., Hu, X., & Graesser, A. (2013). Recent Advances in Conversational Intelligent Tutoring Systems. *AI Magazine*, 34(3), 42–54. <https://doi.org/10.1609/aimag.v34i3.2485>

- Silvervarg, A., & Jönsson, A. (2011). Subjective and objective evaluation of conversational agents in learning environments for young teenagers. In *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Retrieved from https://www.researchgate.net/profile/Arne_Joensson/publication/266895634_Subjective_and_Objective_Evaluation_of_Conversational_Agents_in_Learning_Environments_for_Young_Teenagers/links/546329510cf2cb7e9da67e88.pdf
- Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 1223–1237). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-36124-3_77
- Smith, C., Rumbell, T., Barnden, J., Hendley, B., Lee, M., & Wallington, A. (2007). Don't worry about metaphor: affect extraction for conversational agents. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 37–40). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://0-dl.acm.org.aupac.lib.athabascau.ca/citation.cfm?id=1557769.1557782>
- Soliman, M., & Guetl, C. (2012). Experiences with BDI-based design and implementation of Intelligent Pedagogical Agents. In *2012 15th International Conference on Interactive Collaborative Learning (ICL)* (pp. 1–5). <https://doi.org/10.1109/ICL.2012.6402046>

Sundar, S. S., Bellur, S., Oh, J., Xu, Q., & Jia, H. (2014). User Experience of On-Screen Interaction Techniques: An Experimental Investigation of Clicking, Sliding, Zooming, Hovering, Dragging, and Flipping. *Human-Computer Interaction*, 29(2), 109–152. <https://doi.org/10.1080/07370024.2013.789347>

Szafir, D., & Mutlu, B. (2012). Pay Attention!: Designing Adaptive Agents That Monitor and Improve User Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 11–20). New York, NY, USA: ACM. <https://doi.org/10.1145/2207676.2207679>

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>

Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing* (pp. 680–690). Retrieved from <http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rt doc&an=18533382>

Veletsianos, G., & Miller, C. (2008). Conversing with pedagogical agents: A phenomenological exploration of interacting with digital entities. *British Journal of Educational Technology*, 39(6), 969–986. <https://doi.org/10.1111/j.1467-8535.2007.00797.x>

Veletsianos, G., & Russell, G. S. (2013). What Do Learners and Pedagogical Agents Discuss When Given Opportunities for Open-Ended Dialogue? *Journal of*

- Educational Computing Research*, 48(3), 381–401.
<https://doi.org/10.2190/EC.48.3.e>
- Veletsianos, G., & Russell, G. S. (2014). Pedagogical Agents. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 759–769). Springer New York. Retrieved from http://0-link.springer.com.aupac.lib.athabascau.ca/chapter/10.1007/978-1-4614-3185-5_61
- Vildjiounaite, E., Kyllonen, V., Vuorinen, O., Makela, S.-M., Keranen, T., Niiranen, M., ... Peltola, J. (2009). Requirements and software framework for adaptive multimodal affect recognition. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009* (pp. 1–7). <https://doi.org/10.1109/ACII.2009.5349393>
- Wallace, R. S. (2009). The Anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test* (pp. 181–210). Springer Netherlands. Retrieved from http://0-link.springer.com.aupac.lib.athabascau.ca/chapter/10.1007/978-1-4020-6710-5_13
- Warren, M. (2006). *Features of naturalness in conversation* (Vol. 152). John Benjamins Publishing. Retrieved from <https://books.google.ca/books?hl=en&lr=&id=nX46AAAAQBAJ&oi=fnd&pg=PR1&dq=Features+of+Naturalness+in+Conversation&ots=8SCA-01rwr&sig=zOqnfzKOO2NiXNjpoIMf-b-22AU>
- Wen, M., Yang, D., & Rose, C. P. (2014). Linguistic Reflections of Student Engagement in Massive Open Online Courses. In *Eighth International AAAI Conference on*

- Weblogs and Social Media*. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8057>
- Wong, W., Cavedon, L., Thangarajah, J., & Padgham, L. (2012). Flexible Conversation Management Using a BDI Agent Approach. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents* (Vol. 7502, pp. 464–470). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.aupac.lib.athabascau.ca/chapter/10.1007/978-3-642-33197-8_48
- Xiang Yuan, & Yam San Chee. (2005). Design and evaluation of Elva: an embodied tour guide in an interactive virtual art gallery. *Computer Animation & Virtual Worlds*, 16(2), 109–119. <https://doi.org/10.1002/cav.65>
- Xu, Q., Li, L., & Wang, G. (2013). Designing Engagement-aware Agents for Multiparty Conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2233–2242). New York, NY, USA: ACM. <https://doi.org/10.1145/2470654.2481308>
- Yamashita, K., Kubota, H., & Nishida, T. (2005). Designing conversational agents: effect of conversational form on our comprehension. *AI & SOCIETY*, 20(2), 125–137. <https://doi.org/10.1007/s00146-005-0011-8>
- Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A Meta-analysis of the Impact of the Inclusion and Realism of Human-like Faces on User Experiences in Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1–10). New York, NY, USA: ACM. <https://doi.org/10.1145/1240624.1240626>

Yu, C., Aoki, P. M., & Woodruff, A. (2004). Detecting User Engagement in Everyday Conversations. *arXiv:cs/0410027*. Retrieved from <http://arxiv.org/abs/cs/0410027>

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39–58. <https://doi.org/10.1109/TPAMI.2008.52>

Zimmermann, P., Guttormsen, S., Danuser, B., & Gomez, P. (2003). Affective computing—a rationale for measuring mood with mouse and keyboard. *International Journal of Occupational Safety and Ergonomics*, 9(4), 539–551.

Appendix A – Ethics Review Letter of Approval



June 22, 2016

Mr. Michael Procter
Faculty of Science & Technology\School of Computing & Information Systems
Athabasca University

File No: 22170

Expiry Date: June 21, 2017

Dear Michael Procter,

The School of Computing and Information Systems Departmental Ethics Review Committee, acting under authority of the Athabasca University Research Ethics Board to provide an expedited process of review for minimal risk student researcher projects, has reviewed the revisions made to your project, 'A Multi-agent Framework to support User-Aware Conversational Agents in an E-learning Environment'.

Your application has been **Approved on ethical grounds** and this memorandum constitutes a **Certification of Ethics Approval**. It is noted that you require AU Institutional Permission to access university systems, staff or students to conduct your research project. As such, a request for this permission from the Associate Vice-President, Research has been initiated on your behalf. As per University Policy, if you are proposing to access information or assistance or recruit participants from a particular faculty or department, written support from the Dean (or designate) or Departmental Head is required. As you are targeting a specific course to recruit participants, support/approval from the course coordinator to post your recruitment invitation is required. Please forward this written support once received that it may be added to your file.

Participant recruitment and/or data collection **may not proceed** until this institutional permission has been granted. You will be notified in writing of the outcome of this request for access.

AUREB approval, dated June 22, 2016, is valid for one year less a day.

As you progress with the research, all requests for changes or modifications, ethics approval renewals and serious adverse event reports must be reported to the Athabasca University Research Ethics Board via the Research Portal.

To continue your proposed research beyond June 21, 2017, you must apply for renewal by completing and submitting an Ethics Renewal Request form. Failure to apply for **annual renewal** before the expiry date of the current certification of ethics approval may result in the discontinuation of the ethics approval and formal closure of the REB ethics file. Reactivation of the project will normally require a new Application for Ethical Approval and internal and external funding administrators in the Office of Research Services will be advised that ethical approval

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

has expired and the REB file closed.

When your research is concluded, you must submit a Project Completion (Final) Report to close out REB approval monitoring efforts. Failure to submit the required final report may mean that a future application for ethical approval will not be reviewed by the Research Ethics Board until such time as the outstanding reporting has been submitted.

At any time, you can login to the Research Portal to monitor the workflow status of your application.

If you encounter any issues when working in the Research Portal, please contact the system administrator at research_portal@athabascau.ca.

If you have any questions about the REB review & approval process, please contact the AUREB Office at (780) 675-6718 or rebsec@athabascau.ca.

Sincerely,

Ali Akber Dewan
Chair, School of Computing and Information Systems Departmental Ethics Review Committee
Athabasca University Research Ethics Board

Appendix B – Questionnaire

Freudbot chat study

In this study you will chat with Freudbot, a conversational agent that simulates conversing with the famous psychologist Sigmund Freud about his theories, life, and family. After your conversation you will complete a questionnaire about your experience with Freudbot. This should not take more than 30 minutes, but should be completed without interruption. Note that you will be redirected to a website maintained by the researcher, for the purpose of chatting with Freudbot. You are being asked to participate in a study that will investigate the use of conversational agents, or chatbots, for educational purposes. The chatbot used in the study is designed to emulate Sigmund Freud. Freudbot is programmed to talk extensively about Freud's concepts, theories and biographical events.

If you agree to participate, you will be given instructions to chat with Freudbot for a minimum of 10 minutes, and connected to the chatbot. After 10 minutes, a link will appear which will take you to a set of questionnaires about your Freudbot experience and relevant demographic variables. In addition, the researcher and research associates will have access to the anonymous transcript of your conversation. The entire procedure should take approximately 30 minutes and can be done anytime at your convenience. Please note that once you start, the questionnaire must be completed immediately after the chat session and should be done without interruption.

Potential benefits to participants:

Participants will likely learn more about Freud's concepts, theories and biographical events. This information will be helpful in achieving the course objectives in courses that cover Freud and his theories. In addition to learning more about Sigmund Freud, participation will lead to a better understanding of the mechanics of psychological research over the internet and the ethical principles that govern all research with human participants. Finally, the results of the study will provide important information on the effective use of conversational agents in online distance education and the best strategies for providing information.

Potential risks and discomforts:

There are no risks from participation. However, if you experience any type of discomfort during the study, you may discontinue your participation with no consequences of any kind.

There are 36 questions in this survey

Start

Thank you for agreeing to take part in this study.

There are two parts to the study. First you will chat with Freudbot, a conversational agent which simulates having a conversation with Sigmund Freud, the famous psychologist. Freudbot can talk about his life, family and colleagues, and theories, so you may think of this as carrying out an interview with Sigmund Freud. Freudbot does not perform psychoanalysis or provide therapy.

You will chat for a minimum of 10 minutes. After 10 minutes you may end your conversation at any time and fill out a short set of questionnaires.

Remember, you need at least 30 minutes of uninterrupted time to complete this study. If possible, please work in a quiet location where you can be alone and will not be interrupted by other people. If you are ready, click here to continue.

Click the link below to start chatting with Freudbot. After finishing you will automatically return here to complete the questionnaire.

[Start Chat](#)

Only answer this question if the following conditions are met:

Answer was 'No' at question ' [h1]' (Have to have something here even though it is hidden)

Do not use your browser's forward and back buttons.

When you have completed the questions on a page, click the Next button. The Next button can be found at the bottom of each page. You may need to scroll down to see it. Please do not use the enter key to advance.

Please rate your experience with Freudbot based on the following questions

Please choose the appropriate response for each item:

	Not very				Very
	1	2	3	4	5
How lifelike was Freudbot?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

	Not very 1	2	3	4	Very 5
How good was Freudbot's use of language?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How natural was your interaction with Freudbot?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How easy was it to pretend you were talking with Freudbot?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How compelling were the visual aspects of the interface?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How friendly was Freudbot?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please rate your experience with Freudbot based on the following questions

Please choose the appropriate response for each item:

	Not very 1	2	3	4	Very 5
How enjoyable was this activity?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How engaging was this activity?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How easy was this activity?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How useful is this activity for learning information about Sigmund Freud?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How useful is this activity for remembering information about Sigmund Freud?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Would you recommend this activity to others? (1=Not recommend, 5=Highly recommend)

Please choose **only one** of the following:

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

1 2 3 4 5

Overall, how would you rate this activity? (1=Poor, 5=Excellent)

Please choose **only one** of the following:

1 2 3 4 5

What was the best feature about this activity?

Please write your answer here:

What was the worst feature about this activity?

Please write your answer here:

Would you speak with Freudbot again?

Please choose **only one** of the following:

- Yes
- No

How well would you rate Freudbot's ability to understand?
(1=Understood nothing, 5=Understood all)

Please choose **only one** of the following:

1 2 3 4 5

Which of the following actions did Freudbot take when he did not appear to understand your input? (select all that apply) *

Please choose **all** that apply:

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- Asked me what topic I would like to discuss
- Suggested a topic to discuss
- Asked me to clarify or restate my response
- Asked me a question
- Asked if he should continue
- Not sure
- Freudbot understood all my input

Overall, how would you rate Freudbot's response when he did not appear to understand? (1=Poor, 5=Excellent)

Only answer this question if the following conditions are met:

`((is_empty(a13_6.NAOK) and (is_empty(a13_7.NAOK)))`

Please choose **only one** of the following:

- 1 2 3 4 5

Did Freudbot suggest any topics to discuss? *

Please choose **only one** of the following:

- Yes
- No
- Not sure

How would you rate Freudbot's choice of topics? (1=Poor, 5=Excellent)

Only answer this question if the following conditions are met:

`Answer to previous question is 'Yes'`

Please choose **only one** of the following:

- 1 2 3 4 5

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

Did Freudbot comment on how you were conversing, or offer suggestions to help you converse with him? *

Please choose **only one** of the following:

- Yes
- No
- Not sure

How appropriate or useful were the comments/suggestions regarding conversing with Freudbot? (1=Poor, 5=Excellent)

Only answer this question if the following conditions are met:

Answer to previous question is 'Yes'

Please choose **only one** of the following:

- 1 2 3 4 5

Rate each of the following features in terms of how important they are to you

Please choose the appropriate response for each item:

	Not important				Very important
	1	2	3	4	5
The conversational approach to interacting with Freudbot	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The organization using topics made up as stories or narratives	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The ability to change topics and return to where you left off	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Conversing with Freudbot in the first person, as if interviewing Freud	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot suggesting topic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot asking questions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

If Freudbot were to be enhanced, please rate each of the following areas of improvement in terms of how important they are to you

Please choose the appropriate response for each item:

	Not important 1	2	3	4	Very important 5
Freudbot should be capable of responding to audio input (i.e. voice recognition)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot should be capable of providing an audio response (text-to-speech)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot should be animated or capable of displaying facial expressions and lip movements	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot should be able to chat about non-Freud topics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot should be able to display emotions and respond to different emotions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Freudbot should be able to respond more accurately to user input (i.e. improved chat behaviour)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Are there other enhancements to Freudbot that you would consider important?

Please write your answer here:

Rate the following chatbot applications in terms of how importance they are to you

Please choose the appropriate response for each item:

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

	Least important 1	2	3	4	Most important 5
Create chatbots to populate or inhabit chatrooms	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Create chatbots as guides to course administration (i.e. a FAQbot)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Create chatbots as guides to course content	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Create chatbots to assist students with practice quizzes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Create chatbots to represent other famous psychologists	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Create chatbots to help with learning English as a second language	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Are there other chatbot applications that you would consider important?

Please write your answer here:

Please answer the following questions based on your experience with computers.

Overall, how would you rate your general skill level in working with computers? (1=Poor, 5=Excellent)

Please choose **only one** of the following:

- 1
 2
 3
 4
 5

Have you interacted with a chatbot or virtual assistant before (select all that apply)? (select all that apply) *

Please choose **all** that apply:

- Website assistant
- Shopping assistant

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- Special topic assistant (e.g. healthcare, legal)
- Education-related
- Celebrity or historical figure
- Siri, Google Assistant (“Okay, Google”), or Cortana
- Other chatbot or virtual assistant
- Not sure

Please answer the following questions based on your experience with computers.

How would you compare your experience chatting with Freudbot with other chatbots? (1=Worse than most, 5=Better than most)

Only answer this question if the following conditions are met:

Answer to previous question is not “Not sure”

Please choose **only one** of the following:

- 1 2 3 4 5

Please answer the following questions based on your academic background.

How many undergraduate psychology courses have you completed?

Please choose **only one** of the following:

- None
- 1-2 courses
- 3-5 courses
- 6-9 courses
- 10-14 courses
- 15 or more

How many distance education university courses have you completed?

Please choose **only one** of the following:

- None

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

- 1-2 courses
- 3-5 courses
- 6-9 courses
- 10-14 courses
- 15-20 courses
- 21-29 courses
- 30 or more

How many university courses have you completed?

Please choose **only one** of the following:

- None
- 1-5 courses
- 6-10 courses
- 11-20 courses
- 21-30 courses
- 31-40 courses
- 41-50 courses
- 51 or more

Prior to your chat, how would you rate your knowledge of Freudian concepts and theories? (1=Poor, 5=Excellent)

Please choose **only one** of the following:

- 1 2 3 4 5

In your own opinion, how important do you think Freud's theories were to our understanding of human behavior? (1=Not important, 5=Very Important)

MULTI-AGENT FRAMEWORK FOR USER-AWARE CONVERSATIONAL AGENTS

Please choose **only one** of the following:

- 1
- 2
- 3
- 4
- 5

Demographic Information

Gender identification

Please choose **only one** of the following:

- Male
- Female

Age category

Please choose **only one** of the following:

- 18-22
- 23-27
- 28-32
- 33-37
- 38-42
- 43-47
- 48-52
- 53-57
- 58-62
- 63+

Current student status

Please choose **only one** of the following:

- Full-time
- Part-time
- Non-student

Academic ability

Please choose **only one** of the following:

- Under 50th percentile
- Between the 50th & 65th percentile
- Between the 66th & 79th percentile
- Between the 80th & 89th percentile
- Above the 90th percentile

Please check all that apply

Please choose **all** that apply:

- This questionnaire was easy to complete
- This questionnaire was too long
- This questionnaire captured my opinion adequately
- The questions were relevant to me
- Completing the questionnaire was a useful learning experience

Please provide any additional comments (optional)

Please write your answer here:

You must click the **SUBMIT** button at the bottom to receive credit for participating

Debriefing

The purpose of this study is to assess a conversational agent (CA) that is capable of detecting and responding to user engagement and conversational behaviour. Conversational agents or chatbots refer to web-based programs that are designed to emulate human conversationalists and are increasingly used in commercial roles as virtual company representatives. The use of chat bots in distance education and online education, however, remains largely unexplored. For the purpose of this study we are using Freudbot, a CA

developed to help students learn more about Sigmund Freud in a more natural and engaging setting. Additional software agents have been developed and are used to monitor how the student converses with the Freudbot, attempting to identify some common conversational behaviours, and to determine the level of engagement by analyzing the conversational in real time. This information is used to provide feedback to the student or change the way that Freudbot converses.

In this study we examine and compare objective data, from an analysis of the chat log, and subjective data, collected by the questionnaire. We are interested in how well the agents-based enhancements to Freudbot perform, and how they are perceived by the participants in the study. Our hypothesis is that timely use of conversational interventions can increase the student's level of engagement by assisting them in interacting with the CA when it is determined that they may be experiencing problems. We also hypothesize that increasing the student's level of engagement will improve their overall impression of the interaction with the CA.

You might be wondering why, in the beginning, we didn't explain to you exactly what our hypotheses were at the beginning of the study. If we told you our hypothesis, you might have felt pressure to react in the way you thought we expected you to on the basis of our theory rather than reacting the way you normally would. The possibility that some participants might react to the manipulations based on what the experimenters expect is called the *demand awareness effect*. This can be a problem in research because our results could reflect nothing having to do with the psychological processes that we're interested in studying, but could simply reflect demand awareness. If this was the case, scientific progress would be slowed and inappropriate avenues of research could be followed. I hope you can see why it was necessary to conceal this aspect of the study from you.

Thank you very much for participating. Without the help of people like you, we couldn't answer some of the important and interesting practical questions in psychology and education.

Mike Procter

Submit your survey.
Thank you for completing this survey.