Evaluating the Perception of Personality and Naturalness in Computer Generated Utterances

by

Christopher McCormick, B.Sc.(Hons)

Dissertation

Presented to the

University of Dublin, Trinity College

in fulfillment

of the requirements

for the Degree of

Master of Science in Computer Science

University of Dublin, Trinity College

August 2012

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Christopher McCormick

August 31, 2012

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Christopher McCormick

August 31, 2012

Acknowledgments

I would like to thank my supervisor Dr. Saturnino Luz for his helpful advice and direction throughout the course of the research. I would also like to express my gratitude to Teresa McCormick for her many dedicated proof reading sessions, without which the dissertation would look nothing like it is today. My thanks also go out to Conor McCormick for helping with the final proof read and pointing out the final mistakes that I had overlooked. Many of my thanks go out to Lisa Dollard for all of her support throughout the year, and for putting up with my insistent tendency to think out loud.

Christopher McCormick

University of Dublin, Trinity College
August 2012

Evaluating the Perception of Personality and Naturalness in Computer Generated Utterances

Christopher McCormick
University of Dublin, Trinity College, 2012

Supervisor: Saturnino Luz

The aim of this research is to assess how well a natural language generation system could perform as the natural language generation component of a dialogue system for non-player characters in a computer game. Primarily, the system will be evaluated with regard to its ability to produce natural sounding utterances that can portray personality profiles. The PERSONAGE natural language generation system is chosen to be the target of this research, as it is highly parameterizable and able to produce utterances that portray personality, with respect to the Five Factor Model. A study is designed, conducted and evaluated to determine if PERSONAGE can consistently produce natural sounding utterances that manifest recognizable personalities. A second study is then performed in a more game specific scenario, to determine if the results of the first study translate to the computer game domain.

Contents

Acknowledgments **iv**

Abstract **v**

List of Tables **ix**

List of Figures **xi**

Chapter 1 Introduction **1**

1.1 Motivation 2

1.2 This Report 3

Chapter 2 State of the Art **5**

2.1 Introduction 5

2.2 Personality Traits 7

2.2.1 Trait Correlations with Linguistic Style 8

2.3 Ten-Item Personality Inventory 8

2.4 PERSONAGE 9

2.4.1 Content Planning 11

2.4.2 Sentence Planning 11

2.4.3 Realization 14

2.4.4 Evaluation 14

2.5 Statistical Analysis 16

2.5.1 *t*-test 17

2.5.2 Inter-Rater Agreement 18

Chapter 3 Designing the Study **19**
3.1 Aim . . . 19
3.2 Study Design . . . 20
3.3 Control . . . 21
3.4 Medium . . . 22
3.5 Gathering Participants . . . 24
3.6 Utterances . . . 25
3.6.1 PERSONAGE Utterances . . . 25
3.6.2 Handcrafted Utterances . . . 26
3.7 Questionnaire . . . 27
3.8 Avoiding Study Problems . . . 28
3.8.1 Order Effects . . . 28
3.8.2 Experimenter Effects . . . 29
3.8.3 Demand Characteristics . . . 30
3.9 Ethically Sound . . . 30

Chapter 4 Implementing the Study **32**
4.1 Original Implementation . . . 32
4.2 Modifications . . . 34
4.3 Textures . . . 35
4.4 Double Buffering . . . 35
4.5 ScreenManager . . . 36
4.5.1 MenuScreen . . . 36
4.5.2 PopupScreen . . . 36
4.5.3 UtteranceScreen . . . 37
4.6 DatabaseManager . . . 38
4.7 UtteranceManager . . . 39
4.7.1 PERSONAGE Implementation . . . 40

Chapter 5 Analyzing the Study **41**
5.1 Raw Data . . . 41
5.2 *t*-Test . . . 43
5.3 Inter-Rater Agreement . . . 44

5.4 Discussion 45

Chapter 6 A Second Study **49**

6.1 Study Modifications 49

6.1.1 Modifying PERSONAGE's Domain 50

6.2 Analysis 52

6.2.1 Raw Data 52

6.2.2 *t*-Test 53

6.2.3 Inter-Rater Agreement 54

6.2.4 Discussion 55

Chapter 7 Conclusions and Future Work **57**

7.1 Summary 57

7.2 Conclusion 59

7.3 Future Work 60

Appendices **62**

Bibliography **66**

List of Tables

2.1 PERSONAGE Content Planning Parameters 11
2.2 PERSONAGE Syntactic Template Selection Parameters 12
2.3 PERSONAGE Aggregation Parameters 12
2.4 PERSONAGE Pragmatic Marker Insertion Parameters 13
2.5 PERSONAGE Lexical Choice Parameters 14
2.6 Example adjectives associated with the Big Five traits 15
2.7 PERSONAGE Average Personality Ratings 16
2.8 PERSONAGE Generation Accuracy 16
2.9 PERSONAGE Utterance Naturalness Ratings 16

3.1 Handcrafted Utterances 27

5.1 Study 1 Results: Percentages of Correctly Recognized Personality Profiles 42
5.2 Study 1 Results: Average Naturalness of Utterances 42
5.3 Study 1 Results: Profile Ratings t-test 43
5.4 Study 1 Results: Naturalness t-test 43
5.5 Study 1 Results: Inter-Rater Agreement 44
5.6 Comparing Percentages of Correctly Recognized Personality Profiles with the Original Study 46
5.7 Comparing Average Naturalness with the Original Study 46

6.1 Modified Handcrafted Utterances 50
6.2 Study 2 Results: Percentages of Correctly Recognized Personality Profiles 52
6.3 Study 2 Results: Average Naturalness of Utterances 52
6.4 Study 2 Results: Profile Ratings t-test 54

6.5 Study 2 Results: Naturalness t-test 54
6.6 Study 2 Results: Inter-Rater Agreement 54

List of Figures

2.1 PERSONAGE Architecture 10
2.2 Equation to calculate t (t-test) 17

3.1 Generated Extrovert Utterance Example 26

4.1 Menu Flow Chart 33
4.2 The Main Menu Screen 37
4.3 An Utterance Screen 38

Chapter 1

Introduction

Computer games have progressed at an astounding rate over the last decade. Much of this progress is due to the great attention that has been given to researching and improving the areas that can propel a game forward and make it stand out among its peers. Physics, rendering, animation, artificial intelligence of in-game agents and, more recently, procedural content generation have all played a part. Breakthroughs in the aforementioned areas have progressively led to new generations and styles of computer games. These interactive, visually appealing and immediately gratifying techniques have held the spotlight in the game research domain to date. Other, less dramatic, yet equally important, areas of research have been forced to take a backseat in the rush to develop the "next big thing". One such area, which is the focus of this report, is the procedural generation of written dialogue for non-player characters, using natural language generation techniques.

Natural language generation, hereafter referred to as NLG, refers to the subfield of artificial intelligence that is concerned with the conversion of an abstract notion, that a computer can understand, into an utterance which is comprehensible to a human being, in a human language [1, p. 1].

NLG systems have seen great use in areas such as document planning, and expert systems. However, there has yet to be much attention given to the systems concerning use in mainstream computer games.

In recent years, there has been increasing interest in the use of NLG in interactive entertainment technology, such as games. In particular, with the advent of procedural

quest generation systems, role-playing computer games stand to benefit greatly from NLG techniques.

A class of systems that use NLG to produce comprehensible utterances, which seem the most applicable to mainstream games, is dialogue systems. A dialogue system produces short utterances, responding to user inputs, to imitate a conversation. This type of system could translate across to the role of non-player character dialogue generator. However, for the dialogue generator to be adequate, it is important that the character's utterances reflect their respective personalities. Therefore, the research in this paper is directed at evaluating the performance of an NLG system, with respect to its ability to portray personality in natural sounding utterances, in the hopes of it being used as the NLG component of a dialogue system.

Recent research in affective NLG, NLG which deliberately influences emotions of the recipient [2], has led to the development of systems that can produce utterances that match certain personality profiles. In other words, the system is provided with a personality and the resulting utterances are modified. Examples of these modifications are adding stutters or repetitions to portray nervousness or lengthening the utterance to portray a sociable demeanor. PERSONAGE is such a system and, given its parameterizable nature, seems well suited for the role of NLG component in a procedural non-player character dialogue generator. PERSONAGE will be the system under scrutiny in this study.

1.1 Motivation

By introducing a dialogue system into a game engine to procedurally handle most, if not all, non-critical, non-player character dialogue interactions with a player, a game studio could theoretically save themselves a lot of time and money.

An example of the use of a dialogue system in a game could be for a non-player character to recommend a particular quest over a more difficult one, when a player is presented with a choice of how they wish to progress. Normally a script writer would be required to script dialogue in a situation such as this.

However, imagine a scenario where 100 different non-player characters, with diverse personalities, are required to recommend the player's next quest in accordance with their particular personality. This would require a great deal of effort on the part of the

script writer. In contrast, the dialogue system used in the first example could produce the desired results, given that the NLG component of the dialogue system can handle personalities and that the personalities were passed to the system. The benefits of procedural dialogue generation become apparent under such circumstances.

The next step towards confirming PERSONAGE's potential to be used as a sophisticated NLG component for an in-game dialogue system is to demonstrate whether the character's generated way of speaking is natural and consistent with the personality intended. In the paper which PERSONAGE was originally presented, a study was undertaken to determine if users could identify the intended personality of a speaker, using the Ten-Item Personality Inventory [3], when presented with a PERSONAGE generated utterance [4]. The study returned mainly positive results, which are explained in more detail in Chapter 2.

Although the study was well structured and thoroughly researched, the participant pool was somewhat limited. A group of 3 judges evaluated the personality trait Extraversion, thought by Goldberg to be the most important trait [5], while a group of 2 judges evaluated the rest. All of the judges were professional researchers in various fields and were familiarized as to the nature of the Big Five personality traits prior to taking part in the study. This population may not be a representative sample, and may also be too small a sample to derive any definitive conclusions. Therefore, in order to determine whether the personality in a PERSONAGE generated utterance is perceived correctly, it would be useful to repeat this study, or conduct a similar study, targeted at a larger and more representative sample.

1.2 This Report

The aim of this report is to provide substantial, empirical evidence pertaining to the performance of the rule-based PERSONAGE system, with regard to the portrayal of personality profiles in, natural sounding, generated utterances. The study will be performed online in an attempt to evaluate the system on a larger scale, with a more representative sample, than the original study. The study design will be largely based on the original efforts presented in [4].

As this report is primarily interested in the application of NLG to interactive entertainment technology, it would be of interest to determine whether the results of

the study mentioned above translate across to the computer game domain. Therefore, a second study will be conducted, with some minor modifications, and be evaluated solely by computer gamers.

Chapter 2 explores the current state of the art of NLG systems, personality traits and how they are measured, the rule-based PERSONAGE system, and statistical analysis techniques as they relate to the studies documented in this report. The results of the original study performed by Mariesse and Walker are outlined towards the end of the chapter.

All sections related to the design of the study are documented in Chapter 3. These sections outline the aim and hypothesis, how the study is conducted, how the personalities are measured, the population sampling technique, how people are made aware of the study, the control used and any ethical implications or study problems that may arise.

Chapter 4 explains the implementation process. The flow of the study application is described, as well as internal structure of the program. The database workings are explained, and how the personality ratings, and other relevant information, are stored. The efforts of implementing the PERSONAGE system within the study are also detailed in this chapter.

The results of the study are analyzed and evaluated in Chapter 5. Results are also examined and discussed in light of the results of the original study.

Chapter 6 describes the second study, which is targeted at the computer game domain. Any modifications made to the original study are highlighted, before the new set of results are evaluated and presented.

Finally, Chapter 7 discusses future work that could be undertaken in the wake of this study and presents the conclusion reached as a result of the conducted research.

Chapter 2

State of the Art

This chapter reviews the current state of the art for all topics that are of interest to this report. These areas include NLG systems, the PERSONAGE system, how personality is measured and statistical analysis techniques, which will be used in Chapters 5 & 6.

2.1 Introduction

There are numerous state of the art systems in the area of NLG. Many of these are used to automatically produce documents that make large amounts of raw data more presentable, present arguments in an effort to change the reader's point of view on a particular subject, or complete work that would otherwise prove to be tedious and time consuming [6]. Two example systems that make raw data easier to understand are EasyText [7] and pCRU [8].

A class of systems which uses NLG to create output utterances, and that seems applicable to the game domain, is dialogue systems. A dialogue system is a computer system that can converse with a user. This involves receiving input from a user, such as speech, understanding that input, using the input in some meaningful way and finally producing an output to display, or verbalize, back to the user [9, p. 35]. A well-known example of a dialogue system is SUNDIAL [10]. Other examples, as described by Jurafsky and Martin, include telephone dialogue systems that guide travellers through the reservation process, or, more famously, HAL 9000 from the film "2001: A Space Odyssey" [9, p. 42].

While a dialogue system may translate easily to the computer game domain, as a non-player character dialogue generator, the method of receiving input through speech would not be applicable to most modern games, as they probably won't include microphone capabilities. Therefore, for the sake of this research, the assumption is made that the input to the dialogue systems will be in a multiple choice menu format.

Any one dialogue system, or conversational agent, is made up of various components, such as an automatic speech recognition unit, a natural language understanding unit, a natural language generator and possibly even a text-to-speech engine [9]. Rambow and Walker also state that although "advancements in automatic speech recognition technology has put the goal of naturally sounding dialog systems within reach", the generated responses may not yet be sufficiently sophisticated [6]. Therefore, in the hopes of finding an dialogue system that is sufficiently sophisticated for the game domain, the NLG component of a dialogue system will be the focus of this study.

A standard NLG system, however, may not be adequate to be implemented in a non-player character dialogue generator in an interactive environment such as a computer game. The utterances produced by each non-player character would either be the same or very similar, and would also lack personality. This is a big problem as a game's story is not only told through plot and narrative but also through the dialogue of its characters [11].

Mairesse and Walker mention that recent research in dialogue systems and affective NLG, such as the studies presented in [12, 13], has discovered that it is also beneficial to adapt the utterance to the recipient at the personality level [4]. This research inspired the development of NLG systems that can alter the produced utterance to suit different personality profiles, such as the CRAG-2 and PERSONAGE systems [4, 14].

The PERSONAGE system is a natural language generator that, unlike the CRAG-2 system, is highly parameterizable. It uses tweakable parameters to adapt the produced utterance to a personality profile's linguistic style [4]. The system appears to be very applicable to the game domain, and as such will be used as the NLG system for the research in this paper.

In this chapter the "Big Five" personality traits, which PERSONAGE is based on, are examined, and PERSONAGE is explained in more detail. The Ten-Item Personality Inventory, which is used to measure personalities with respect to the Five Factor Model, is also reviewed. Finally, statistical analysis techniques, which will be used to evaluate

the study results in Chapters 5 and 6, are explored.

2.2 Personality Traits

Over the past 50 years researchers in psychology have begun to define the core personality traits in humans. Starting with the work of Fiske [15] and expanded by research such as [16, 17], a personality model evolved that was accepted by the psychology community. This model is known as the "Big Five", or the "Five Factor Model" [18].

The Big Five personality model represents personality at the broadest level of abstraction, each dimension summarizing a larger number of more specific characteristics [19]. The five dimensions of the Big Five are *extraversion*, *emotional stability*, *agreeableness*, *conscientiousness* and *openness to experience*.

Extraversion refers, in a way, to the level of sociability of individuals. Extroverts are characterized as being predominantly concerned with obtaining gratification from others. They enjoy human interaction and are likely to be talkative, enthusiastic and assertive. A low level of *extraversion*, however, makes an individual an introvert. Introverts tend to be the opposite of extroverts in characteristics, i.e. solitary and reserved [20, p. 437].

Emotional stability refers to how emotionally balanced an individual is. A high *emotional stability* means a person is inclined to be more calm and relaxed than an individual with a low *emotional stability*, and less likely to experience negative emotions [20, p. 437].

Agreeableness refers to a person's tendency to be cooperative, compassionate and friendly towards others. A low *agreeableness*, in contrast, means the person is suspicious, unfriendly and more concerned with matters that revolve around themselves [20, p. 437].

Conscientiousness refers to how well organized an individual is, and how likely they are to have high aims or goals. A person with low *conscientiousness* prefers spontaniety and is more likely to live a "live in the moment" lifestyle [20, p. 437].

Openness to experience refers to how curious and open to new experiences a person is. A person with a high *openness to experience* tends to be more creative and free thinking than those with a low score. A person with a low score is also likely to be resistant to change [20, p. 437].

2.2.1 Trait Correlations with Linguistic Style

Studies in psycholinguistics also suggest that certain aspects of the Big Five personality traits are portrayed in speech [21, 22, 23].

Nowson analyzed blogs in an effort to determine the individual differences in linguistic styles, and whether personality and gender were projected linguistically in blogs [23]. A brief review of his findings, with respect to personality correlations with linguistic style, includes the following.

Highly neurotic people tend to write in the future tense, use a lot of anxiety words, talk more about themselves and their physical states and use more discrepancy words, such as 'needs', 'wants' or 'wishes' [23].

People with a high level of extraversion tend to talk in the present tense, use fewer discrepancy words, talk more and discuss their feelings more frequently [23].

People with a high level of openness to experience use long words, more prepositions and more communicative words related to positive feelings and inclusion [23].

Highly agreeable people use fewer discrepancy words, fewer tentative and certainty terms, talk less about negative emotions and use longer words, more articles and fewer motion words, like 'walk', or 'go' [23].

Highly conscientious people tend to use fewer words and avoid talking about friends or death [23].

In a separate study, Pennebaker also found that highly extrovert or conscientious people tend to make fewer distinctions, and people that are highly open to new experiences tend to use infrequent words, talk in a more immediate sense, and use a complex structure [21].

2.3 Ten-Item Personality Inventory

While there are various methods to measure personality with respect to the Five Factor model, they are usually time-consuming. Gosling highlights some of the methods in question, and the time they can take to complete: the 240-item NEO-PI-R, which can take around 45 minutes to complete [16], the 44-item Big Five Inventory, which can take 5 minutes to complete [24, 19], the 60-item NEO Five-Factor Inventory [16] and Goldberg's TDA [25], which both take around 15 minutes to complete [3]. Gosling

reasons that "we do not encourage the use of very brief measures, but we acknowledge that when brevity is a high priority, researchers may be driven to create their own very short measures of the Big Five, or even worse, to use no measure at all" [3]. Therefore Gosling et al presented the Ten-Item Personality Inventory (TIPI), for which he states "the psychometrics are known and are reasonable", to be used in such situations where time is short and of the essence [3].

The TIPI contains ten items, one to represent each pole of the Big Five traits. Every item has two descriptive words attached to it, which are separated by a comma. A 7-point scale is also attached, to be used for rating, which ranges from strongly disagree to strongly agree. The descriptive words for each item are as follows: High Extraversion (enthusiastic, extroverted), Low Extraversion (reserved, quiet), High Agreeableness (sympathetic, warm), Low Agreeableness (critical, quarrelsome), High Conscientiousness (dependable, self-disciplined), Low Conscientiousness (disorganized, careless), High Emotional Stability (calm, emotionally stable), Low Emotional Stability (anxious, easily upset), High Openness to Experience (open to new experiences, complex) and Low Openness to Experience (conventional, uncreative). Using the TIPI, the complete rating of a personality takes only a minute to complete [3].

2.4 PERSONAGE

PERSONAGE (PERSONAality GEnerator) was first presented by François Mairesse and Marilyn Walker in 2007 [26]. The generator was originally designed to produce utterances that portray extrovert personalities [26], however the system was later extended to work for all of the Big Five personality traits, making it the first system that successfully demonstrated that an automatic generator could manifest the Big Five traits in an utterance for a dialogue application [27].

Mariesse and Walker hypothesise that "personality can be made manifest in evaluative speech acts in any dialogue domain, i.e. utterances responding to requests to RECOMMEND or COMPARE domain entities, such as restaurants or movies" [26]. Thus, the PERSONAGE system was built to present utterances that recommend or compare restaurants around New York City. Each restaurant has eight attributes: the *name* and *address*, categorical attributes for *neighbourhood* and *type of cuisine*, and scalar attributes for *price*, *food quality*, *atmosphere* and *service* [26, 27, 4].

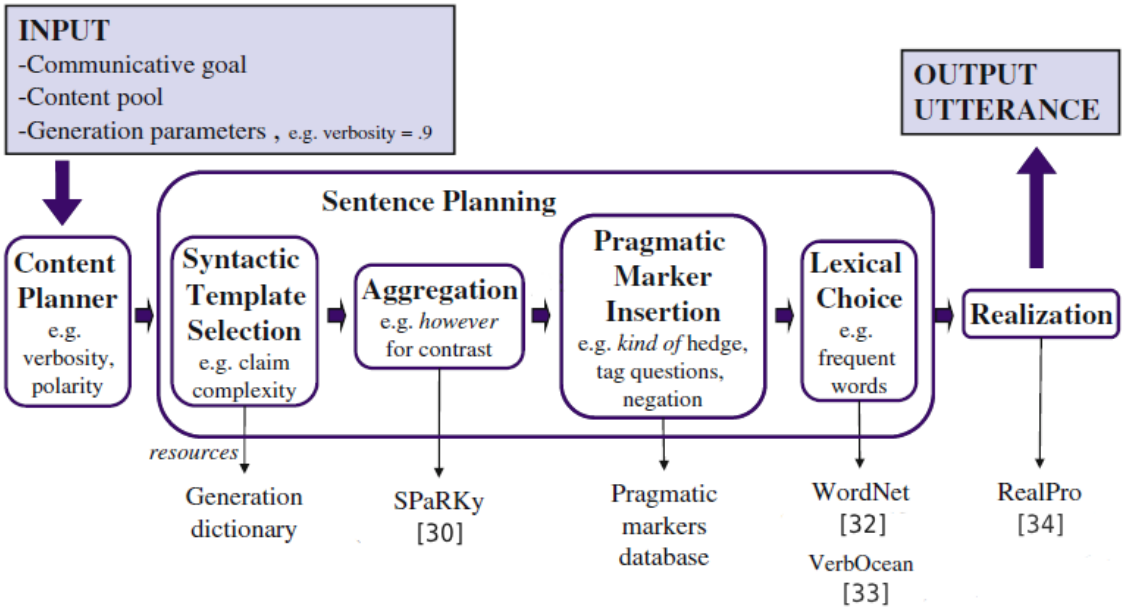

Figure 2.1: The PERSONAGE NLG system architecture. Source: [4].

PERSONAGE's architecture, as seen in Figure 2.1, is based on the standard NLG pipeline architecture as outlined in [1]. To generate an utterance three inputs are required: a content plan that represents a speech act, a content pool, which in this case is the database of restaurants and their attributes, and a set of generation parameters, which define how the produced utterance appears [4]. It is these parameters that are tweaked to portray the personality models in an utterance. However, Mairesse and Walker explain that thousands of different utterances can be produced for any content plan and simply randomly varying the parameters can lead to inconsistent personality cues. Therefore, in order to determine the parameter values that would produce utterances to match the personality profiles of the Big Five traits, Mariesse and Walker analyzed existing psycholinguistic studies, including [21, 22, 23], that document the linguistic reflexes of personality [4]. The parameters in question will be described in detail in the following subsections. The respective values of all parameters required to meet the personality profiles of extraversion, emotional stability, agreeableness, conscientiousness and openness to experience can be found in [4]. Further research shows that these parameters, and ultimately the personality models, can in fact be learned from example utterances, such as film dialogue [11].

Table 2.1: PERSONAGE Content Planning Parameters. Source: [4]

Parameters	Description
VERBOSITY	Control the number of propositions in the utterance
RESTATEMENTS	Paraphrase an existing proposition, e.g. *"X has great Y, it has fantastic Z"*
REPETITIONS	Repeat an existing proposition
CONTENT POLARITY	Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
REPITITION POLARITY	Control the polarity of the restated propositions
CONCESSIONS	Emphasize one attribute over another, e.g. *"even if X has great Z, it has bad Y"*
CONCESSION POLARITY	Determine whether positive or negative attributes are emphasized
POLARIZATION	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	Determine whether positive propositions are uttered first
REQUEST CONFIRMATION	Begin the utterance with a confirmation of the request, e.g. *"did you say X?"*
INITIAL REJECTION	Begin the utterance with a rejection, e.g. *"Im not sure"*
COMPETENCE MITIGATION	Express the speakers negative appraisal of the hearers request, e.g. *"everybody knows that "*

2.4.1 Content Planning

Content planning is concerned with establishing what to say in an utterance, and how it is structured [1, p. 47-49]. The content planning process takes a communicative goal, such as the recommendation or comparison of restaurants, and converts it into a high level rhetorical structure tree [28], which expresses information about the individual attributes of the communicative goal [4].

At this stage in the process, there are 12 parameters that affect the tree structure's size, content ordering, rhetorical relations used and the polarity of propositions. These are *verbosity*, *restatements*, *repetitions*, *content polarity*, *repetition polarity*, *concessions*, *concession polarity*, *polarization*, *positive content first*, *request confirmation*, *initial rejection* and *competence mitigation* [4]. Each parameter is explained, some with examples, in Table 2.1.

2.4.2 Sentence Planning

The content planning stage defined what was to be said in the utterance. The sentence planning stage takes that as input and controls how that information is to be conveyed [4]. The sentence planning process is split into four smaller processes: syntactic template selection, aggregation, pragmatic marker selection and lexical choice.

Syntactic template selection searches a handcrafted generation dictionary for the set of syntactic elementary structures that correspond to each proposition in the content plan [4]. These structures are referred to as Deep Syntactic Structures (DSyntS), which, Mairesse and Walker state, are based on the work of Mel'čuk [29]. The parameters that

Table 2.2: PERSONAGE Syntactic Template Selection Parameters. Source: [4]

Parameters	Description
SYNTACTIC COMPLEXITY	Control the syntactic complexity (e.g. syntactic embedding)
SELF-REFERENCES	Control the number of first person pronouns
TEMPLATE POLARITY	Control the syntactic structures connotation (positive or negative)

come into effect in the syntactic template selection stage, outlined in Table 2.2, are *syntactic complexity*, *self-references* and *template polarity*.

In order to produce a large number of sentences from the limited number of DSyntS, the aggregation stage combines the base DSyntS into larger syntactic structures [4]. A clause-combining operation is randomly chosen for each rhetorical relation in the tree, which Mariesse and Walker state is similar to the work in [30]. PERSONAGE employs a slightly modified version of SPaRKy (Sentence Planning with Rhetorical Knowledge) to perform the aggregation [31]. The aggregation parameters that can be tweaked are presented in Table 2.3.

Table 2.3: PERSONAGE Aggregation Parameters. Source: [4]

Parameters	Description
PERIOD	Leave two propositions in their own sentences, e.g. "X has great Y. It has nice Z."
RELATIVE CLAUSE	Join propositions with a relative clause, e.g. "X, which has great Y, has nice Z"
WITH CUE WORD	Aggregate propositions using with, e.g. "X has great Y, with nice Z")
CONJUNCTION	Join propositions using a conjunction, or a comma if more than two propositions
MERGE	Merge the subject and verb of two propositions, e.g. "X has great Y and nice Z"
ALSO CUE WORD	Join two propositions using also, e.g. "X has great Y, also it has nice Z"
CONTRAST-CUE WORD	Contrast two propositions using while, but, however, on the other hand, e.g. "While X has great Y, it has bad Z", "X has great Y, but it has bad Z"
WHILE CUE WORD	Contrast two propositions using while, e.g. "While X has great Y, it has bad Z"
HOWEVER CUE WORD	Contrast two propositions using however, e.g. "X has great Y. However, it has bad Z"
ON THE OTHER HAND CW	Contrast two propositions using on the other hand, e.g. "X has great Y. On the other hand, it has bad Z"
JUSTIFY-CUE WORD	Justify a proposition using because, since, so, e.g. "X is the best, since it has great Y"
BECAUSE CUE WORD	Justify a proposition using because, e.g. "X is the best, because it has great Y"
SINCE CUE WORD	Justify a proposition using since, e.g. "X is the best, since it has great Y"
SO CUE WORD	Justify a proposition using so, e.g. "X has great Y, so its the best"
CONCEDE-CUE WORD	Concede a proposition using although, even if, but/though, e.g. "Although X has great Y, it has bad Z", "X has great Y, but it has bad Z though"
ALTHOUGH CUE WORD	Concede a proposition using although, e.g. "Although X has great Y, it has bad Z
EVEN IF CUE WORDBUT	Concede a proposition using even if, e.g. "Even if X has great Y, it has bad Z"
THOUGH CUE WORD	Concede a proposition using butthough, e.g. "X has great Y, but it has bad Z though"
MERGE WITH COMMA	Restate a proposition by repeating only the object, e.g. "X has great Y, nice Z"
OBJECT ELLIPSIS	Replace part of a repeated proposition by an ellipsis, e.g. "X has it has great Y"

Table 2.4: PERSONAGE Pragmatic Marker Insertion Parameters. Source: [4]

Parameters	Description
SUBJECT IMPLCITNESS*	Make the presented object implicit by moving its attribute to the subject, e.g. "the Y is great"
NEGATION*	Negate a verb by replacing its modifier by its antonym, e.g. "X doesnt have bad Y"
SOFTENER HEDGES	Insert syntactic elements (sort of, kind of, somewhat, quite, around, rather, I think that, it seems that, it seems to me that) to mitigate the strength of a proposition, e.g. "X has kind of great Y" or "It seems to me that X has rather great Y"
EMPHASIZER HEDGES	Insert syntactic elements (really, basically, actually, just) to strengthen a proposition, e.g. "X has really great Y" or "Basically, X just has great Y"
ACKOWLEDGEMENTS	Insert an initial back-channel (yeah, right, ok, I see, oh, well), e.g. "Ok, X has great Y"
FILLED PAUSES	Insert syntactic elements expressing hesitancy (I mean, err, mmhm, like, you know), e.g. "ErrX has, like, great Y"
EXCLAMATION	Insert an exclamation mark, e.g. "X has great Y!"
EXPLETIVES	Insert a swear word, e.g. "the Y is damn great"
NEAR EXPLETIVES	Insert a near-swear word, e.g. "the Y is darn great"
TAG QUESTION*	Insert a tag question, e.g. "the Y is great, isnt it?"
STUTTERING*	Duplicate parts of a content word, e.g. "X has gr-gr-great Y"
IN-GROUP MARKER	Refer to the hearer as a member of the same social group, e.g. pal, mate and buddy
PRONOMINALIZATION	Replace references to the object by pronouns, as opposed to proper names or the reference this restaurant
REQUEST CONFIRMATION	Begin the utterance with a confirmation of the request, e.g. "did you say X?"
INITIAL REJECTION*	Begin the utterance with a rejection, e.g. "Im not sure"
COMPETENCE MITIGATION	Express the speakers negative appraisal of the hearers request, e.g. "everybody knows that ""

Pragmatic markers, or discourse markers, are elements that appear in an utterance, such as "maybe", "well" or "you know". As Mairesse and Walker state, the majority of these markers are context-independent and, in the PERSONAGE system, are inserted into an utterance by the use of syntactic pattern matching. They also mention that there are, however, some more complex markers that are embedded in the syntactic structure of an utterance, and thus require additional processing. To account for these oddities PERSONAGE adds extra entries in the generation dictionary that represent the embedded marker's syntactic structure. For each of these markers used in an utterance, the accumulated DSyntS is traversed to find the insertion point that satisfies the syntactic constraints specified in the database [4]. Table 2.4 lists, and provides examples of, the pragmatic marker insertion parameters. The parameters marked with an asterisk are non-trivial and were required to be implemented individually [4].

Table 2.5: PERSONAGE Lexical Choice Parameters. Source: [4]

Parameters	Description
LEXICON FREQUENCY	Control the average frequency of use of each content word (e.g. according to frequency counts from a corpus)
LEXICON WORD LENGTH	Control the average number of letters of each content word
VERB STRENGTH	Control the strength of the verbs, e.g. "I would suggest" versus "I would recommend"

Word length, word frequency and verb strength are controlled by the lexical choice process. Mariesse and Walker assert that "lexical choice is crucial to successful individual adaption in dialogue systems" [4, 32]. The process is implemented by analyzing the DSyntS and sequentially mapping each lexeme's corresponding WordNet [33] synonyms into a multidimensional space. The space's dimensions are the lexeme's strength, frequency of use and length [4]. The synonym's dimensional values are normalized and the synonym with the closest Euclidean distance is selected. Mairesse and Walker highlight that verb synonym strength is handled slightly differently. Here, the WordNet integration is replaced with the greater than semantic of VERBOCEAN [34]. Table 2.5 displays the lexical choice generation parameters.

2.4.3 Realization

The surface realization stage produces the final utterance string from the abstract sentence plan structure. In other words, the realizer converts the abstract sentence into a particular language, given that it has knowledge of the syntax, morphology and idiosyncratic properties of lexical items of the language [35]. PERSONAGE uses RealPro [35], an "off the shelf" realizer, to perform its realization [4]. RealPro takes the DSyntS as input, assuming lexical choice and syntactic selection has already been performed, and returns the comprehensible English utterance [35].

2.4.4 Evaluation

Mariesse and Walker evaluated PERSONAGE by asking a group of judges to rate a set of system generated utterances by completing the Ten-Item Personality Inventory (TIPI) questionnaire [3]. An extra criterion was added to the questionnaire, asking the judges to rate the naturalness of the utterance, i.e. how closely it resembled what a

Table 2.6: Example adjectives associated with the Big Five traits. Source: [4]

Trait	High	Low
Extraversion	Warm, gregarious, assertive, sociable, excitement seeking, active, spontaneous, optimistic, talkative	Shy, quiet, reserved, passive, solitary, moody, joyless
Emotional Stability	Calm, even-tempered, reliable, peaceful, confident	Neurotic, anxious, depressed, self-conscious, oversensitive, vulnerable
Agreeableness	Trustworthy, friendly, considerate, generous, helpful, altruistic	Unfriendly, selfish, suspicious, uncooperative, malicious
Conscientiousness	Competent, disciplined, dutiful, achievement striving, deliberate, careful, orderly	Disorganized, impulsive, unreliable, careless, forgetful
Openness to experience	Creative, intellectual, imaginative, curious, cultured, complex	Narrow-minded, conservative, ignorant, simple

real person might say. Each utterance presented to the judges was perceived as a new speaker, thus eliminating personality perceptions that could emerge over the course of a dialogue [4]. The judges were researchers of various areas and although they had no knowledge of NLG systems, they were familiarized with the Big Five personality models by being provided with lists of trait adjectives from [5]. The adjectives which the judges were presented with are displayed in Table 2.6. The results received from the judges were averaged for each utterance [4].

Extraversion was initially evaluated separately by three judges [27], as it was deemed to be the most important trait [5]. After receiving positive results from the study the other four traits were evaluated, by a group of two judges. The total set of utterances evaluated consisted of 240 utterances based on personality models and 320 random utterances generated with uniformly distributed parameter values. The randomly generated utterances were involved in the evaluation as a means to determine which parameters in each model were responsible for the judge's perceptions. 40 utterances were rated for each trait, with the exception of extraversion, which instead had 80. The sets of utterances were split in two, each half focusing on opposite dimensions of the associated trait. Each parameter was allowed to vary by a 15% standard deviation so that multiple outputs could be generated for every content plan, trait and personality model [4].

The results of the study, depicted in Tables 2.7 and 2.8, shows that 4 out of the 5 traits, when generated towards the positive ends of the trait, are modelled with a high accuracy. An utterance is defined as being correctly recognized if the average rating

Table 2.7: Average personality ratings for the PERSONAGE utterance sets. Source: [4]

Profiles	Low	High
Extraversion	2.96	5.98
Emotional Stability	3.29	5.96
Agreeableness	3.41	5.66
Conscientiousness	3.71	5.53
Openness to experience	2.89	4.21

Table 2.8: Generation accuracy (in %) for the PERSONAGE utterance sets. Source: [4]

Profiles	Low	High	Overall
Extraversion	82.5	100.0	91.3
Emotional Stability	80.0	100.0	90.0
Agreeableness	70.0	100.0	85.0
Conscientiousness	60.0	100.0	80.0
Openness to experience	90.0	55.0	72.5
All Utterances	85.0		

lies in the half of the scale predicted by its parameter setting. However, Mairesse and Walker found that the utterances with low parameter settings tend to be misrecognized. They reason that this could be due to either a bias of the judges towards the positive end or a bias in PERSONAGE's parameter values. Mairesse and Walker also report that it is clear that *openness to experience* is the most difficult trait to portray in an utterance [4]. The utterances on average, presented in Table 2.9, were judged to be moderately natural, with a rating of 4.59. It can also be seen that the utterances judged to be most natural were associated with extrovert personalities [4]. Mariesse and Walker state that it is likely that "some aspects of personality cannot be conveyed through language, or that more than a single utterance is required" [4].

Table 2.9: Average naturalness ratings for PERSONAGE utterance sets. Source: [4]

Personality Trait	Low	High	Random
Extraversion	4.93	5.78	4.75
Emotional Stability	3.43	4.63	4.72
Agreeableness	3.63	5.56	4.76
Conscientiousness	3.33	5.33	3.86
Openness to experience	3.98	3.85	4.75

2.5 Statistical Analysis

This section reports upon the statistical analysis techniques which will be used, in Chapters 5 and 6, to evaluate the results of the study.

2.5.1 *t*-test

The *t*-test is a statistical analysis technique used to determine the difference between two means in relation to the variance in data. It is used for experimental study designs that have two conditions that test one independent variable. It can be used in both related and unrelated studies, however, the equations differ depending on the type [36, p. 85, 89].

The need to account for the variance in data is important. Greene explains that there would be no problem claiming that Condition 1 scored more highly than Condition 2, if everyone in condition 1 scored X, giving a mean of X, and everyone in condition 2 scored Y, giving a mean of Y, where X is greater than Y. However, realistically everyone is different and will provide different scores. While the mean may still be X and Y, it is likely that there will be variation in the ratings where Condition 1 could be scored more extremely, while the other is scored more tightly around the mean. If this is the case, while Condition 1's mean would support the claim, the lower extreme scores seem to go against it [36, p. 27-30].

This report is interested in the related *t*-test, as the study will use the same participants for both conditions. As the same subjects take part in both conditions, it becomes possible to compare pairs of scores, for each individual, instead of using the means of each group [36, p. 89].

The equation presented in Figure 2.2 is used to calculate t, in a related design, where d is the difference in an individual's scores, calculated by subtracting Condition 2 scores from Condition 1 scores, and N is the number of subjects [36, p. 90]. A short program was written to calculate the t values for this study. These t values were later tested, using an online statistical program [37], and found to be correct.

$$t = \frac{\sum d}{\sqrt{\frac{N\sum d^2 - (\sum d)^2}{N-1}}}$$

Figure 2.2: Equation to calculate t. Source: [36, p. 90]

The t value represents the size of the difference between the subject's scores for the two conditions, which is then used, along with the degrees of freedom (N - 1), to look up the level of significance from a pre-defined, standardized, table. The degrees of freedom are used to look down the left hand column. If the t value is equal to or greater than a critical value in that row, t can be said to be true for that level of significance [36, p. 91].

2.5.2 Inter-Rater Agreement

It is often advantageous to measure the reliability of agreement between participants in a study, relating to the ratings or scores they provide. Measures such as Cohen's kappa [38] can be used for studies that have only two raters. However, the studies to be evaluated in this report will include many raters. Fleiss' kappa, a kappa which allows multiple raters, seemed to be the better choice for this report [39]. However, as Randolph states, Fleiss' kappa "assumes that raters are restricted in how they can distribute cases across categories" [40]. The raters in this report's studies are able to place a case into any category, i.e. they are not restricted. Therefore, Randolph's free-marginal multirater kappa [40] was chosen instead.

Randolph's kappa measures inter-rater agreement on a scale between, and including, -1 and 1. Where close to 1 is almost perfect agreement above chance, close to -1 is almost perfect disagreement below chance and around 0 is agreement equal to chance [40]. An online statistical program, hosted on Randolph's own website, was used to calculate the kappa values for the studies [41].

Chapter 3

Designing the Study

This chapter reports upon all the areas that are related to the study design process. These sections include the aim and hypothesis, how the study is conducted, how the personalities are measured, the population sampling technique, how people are made aware of the study, the control used and any ethical implications or study problems that may arise.

3.1 Aim

The aim of this study is to provide substantial, empirical evidence pertaining to the performance of the rule based PERSONAGE system, with regard to the portrayal of the Big Five personality profiles in, natural sounding, generated utterances.

The PERSONAGE system was previously evaluated, by Mairesse and Walker, which resulted in largely positive results [4]. However, the population was limited to a maximum of 3 judges, who were professional researchers and were familiarized with the nature of the Big Five personality traits, by being provided with a list of trait adjectives from [5], prior to taking part in the study [4]. This population may be too small a sample to derive any definitive conclusions and may not be a representative sample of the actual target population. The fact that the researchers provided the judges with information on the Big Five personality traits raises concerns of experimenter effects and ecological validity, which may confound the results of the study. Ecological validity, as Banister states, is concerned "with trying to make the research

fitting to the real world" [42, p. 4-7]. In the real world people are not primed with information relating to psychological personality profiles when judging an utterance. A brief overview of the original study design is described in Chapter 2. Examples of the list of adjectives the judges were provided with can be seen in Table 2.6.

The study documented in this report will perform research similar to the original study to determine: the recognition percentage of personality profiles in utterances and the perceived naturalness of the generated utterances. However, it will be targeted at a larger population, consisting of individuals, selected using volunteer sampling, whose knowledge of personality traits will not be influenced by the researchers.

A set of handcrafted utterances, which will be written by a real person, will also be rated alongside the generated utterances. This will be useful as it provides a means to compare PERSONAGE's performance, with respect to creating natural sounding utterances that manifest recognizable personalities, with that of a human. It may also help provide insight into whether more than a single, stand-alone utterance is required to consistently portray a personality.

3.2 Study Design

This design of this study is two-tailed, or bi-directional. In other words, although a difference in the perception of personality profiles between the PERSONAGE generated utterances and the handcrafted utterances is expected, the prediction does not specify a direction. Therefore, the hypothesis for the study specifies that, for each of the Big Five personality traits, on comparing the ratings of PERSONAGE generated utterances against the ratings of handcrafted utterances, there will be a significant difference in the percentage of correctly recognized personality profiles. This hypothesis will be tested against the null hypothesis, which states that, for each of the Big Five personality traits, on comparing the ratings of PERSONAGE generated utterances against the ratings of handcrafted utterances, there will be no significant difference in the percentage of correctly recognized personality profiles.

The research will be undertaken in the form of a quantitative, related, double-blind, experimental study. The independent variable for the experimental design is the presentation of utterances that portray personality profiles to the participants, while the dependant variable is the percentage of utterances whose related personality

profiles were correctly recognized.

There are two conditions for the study: participants are asked to rate the personalities of the computer generated utterances through a questionnaire, and participants are asked to rate the personalities of the human crafted utterances through a questionnaire. Both of these conditions can, and ideally will, be completed by the same subjects. The utterances of each questionnaire will be split up and combined into one set, which will be presented to the participants in an indeterminate order. A participant will only rate an utterance with each personality profile once.

Participants will also be asked to fill in an optional survey, before taking part in the study. The survey collects the following information: gender, age range, experience with NLG, experience with the Big Five personality traits and whether the participant has dyslexia. The collected information will not be used in the evaluation of this study, however, it may prove useful for further research.

3.3 Control

A control condition is used by researchers in an effort to minimize unpredictable, irrelevant variables that could affect their study. It is set up in such a way that the participants are not exposed to the independent variable, providing a set of results against which the set of independent variable results can be compared [36, p. 13].

For example, consider an experiment is being conducted to determine whether a specially designed toy encourages children to play together. The study will be split in two. In the control condition, two measures will be taken of the children's sociability, on two separate occasions. In the experimental condition, two measures will also be taken on two separate occasions, where a toy has been introduced after the first measurement. In this way, any significant increase in sociability in the children who were exposed to the toy, where there was little or no increase in sociability in the children in the control can give a definitive result: the toy was responsible for the increase in sociable behaviour in the children.

Originally, this study was set up as a uni-directional study in order to prove that PERSONAGE could generate utterances that closely resemble those spoken by a person with the matching personality profile. In this case, a control condition was designed as a set of handwritten utterances that match the personality profiles of the set of

generated utterances. The control's utterances were then to be rated with respect to the personality profile they represented.

However, after further research, it was found that sometimes comparing two levels of an independent variable is more appropriate than using a control [36, p. 14]. After much deliberation, a decision was reached to change the study to a bi-directional format and, instead, use the control utterances as a second condition. For the sake of this research it is important to measure the difference in the percentage of correct personality recognition between both independent variables as it will not only give a measure as to how well PERSONAGE performs, but it will also give an indication as to any further research that is required, concerning the ability to perceive personality when presented with a single utterance.

3.4 Medium

One of the criteria for this experiment is that the pool of participants involved should be relatively larger than in the original study. In order to reach as many people as possible, in a short time frame, it seems appropriate to perform the study in an online format.

Coincidentally, performing the experiment online will also attract a good representative sample for the study's population. That is, the people who will take part in the study will be computer literate and more likely to come across a PERSONAGE implementation in the future, than those who are computer illiterate.

When deciding upon which medium to use to bring the study online, the first idea to come to mind was to use an online survey service such as SurveyMonkey [43]. The service allows researchers to publish their surveys on the internet in the form of webpages. Although it matches most of the study requirements, in that it provides a means to ask participants to rate utterances in the form of a questionnaire and it provides a function to randomly order the questions, it does not provide the desired control over the prioritization of questions, depending on the number of ratings per trait already gathered. Ordinarily, in a controlled environment this would not be an obstacle, however, due to the online nature of the study and given that each section of the questionnaires takes a considerable amount of time to complete, it is likely that participants will become bored and fail to complete the full set of ratings. This

may result in either withdrawn ratings or an unequivalent number of ratings for each personality profile. Neither of these outcomes is desirable, and played a part in the decision to find a more suitable medium.

It is also of importance to note that the survey services would not provide a real-time generation of PERSONAGE utterances. It would require a predetermined number of utterances to be generated before setting up the study. While this would be acceptable, the most desirable case would be for the utterances to be generated in real-time, so that each participant would be presented with slightly different utterances, as would occur in a real world scenario.

These concerns prompted the search for a more controllable medium toward a more customizable approach: to code the online experiment, exactly as required. Java was the obvious choice for the coding language, as it is the language in which the PERSONAGE system itself was coded, and would allow for a quick integration. With Java as the core language, one decision remained that would determine how the online study would be presented: whether to use Java Server Pages, and present the questionnaires through multiple web-pages, or to use a Java Applet in a single web page.

Implementing the study using Java Server Pages would provide a quick and easy way to set up the questionnaire interfaces, as HTML forms could be used. However, considering that only one questionnaire section is to be displayed at a time, the web page would require reloading every time a new question is asked, in order to access the server side data. On the other hand while a Java applet would take longer to implement, and requires the Java runtime environment to be installed on a participants machine, it does not require the reloading of the page to access new server side data. It also keeps all of the coding practices to Java, and does not require any additional HTML or JavaScript to assist in the layout and animation of menus.

As most people who, nowadays, use the internet for social and recreational purposes have more than likely come across a Java applet, requiring them to install a version of the Java runtime environment on their machine, using a Java Applet as the medium for the study seemed to be the most beneficial option.

3.5 Gathering Participants

The target population for this study includes anyone who can competently use a computer. Given the online nature of the study, this means that anyone who decides to take part in the study is actually a part of the target population. This is fortunate as it allows the researchers to gather a representative sample of the population through a technique called volunteer sampling.

Volunteer sampling is a non-random population sampling technique, where the sample is made up of participants who have volunteered to take part in the study. Using this technique a particularly large sample can be acquired rather quickly, however, as the researcher has little control over the sample, it rarely matches the representativeness requirement [42, p. 6]. As mentioned above, this is not a problematic issue in this study.

One problem with volunteer sampling, and other non-probabilistic methods, is that the results of the study cannot make statistical inferences about a population as a whole [44, p. 322]. The reason for this is that certain people will be more likely to take part in a study than others, for example if the research is a particular area of interest for them. The individuals who have no interest in the research are unlikely to volunteer to become a part of the sample, and as a result a bias is introduced to the research results. Rosenthal also points out that people who are more likely to volunteer to take part in a study tend to be friendlier, brighter, younger and less conventional, but with a strong need for approval [45]. With regard to this study, no statistical inferences will be made regarding the population as a whole, but rather regarding the difference in performance of the utterance generation.

For people to be able to volunteer to take part in the study, they first have to be made aware that the study is taking place. The primary method, by which this study will be advertised, will be through social media sites such as Facebook and Twitter. Using the researchers' personal accounts, a call for participants will be made to friends, family and co-workers. They will also be asked to share the message with their contacts in the hopes of acquiring more participants through word of mouth.

While Facebook and Twitter are good for reaching people who may not have heard of NLG, LinkedIn offers a better means to reach industry professionals, whom the researchers don't know, and who may have an interest in NLG. This is advantageous as ideally the study should be as well rounded as possible, and include as many people

from different types of backgrounds as possible. Posts, detailing the specifics of the study, will be made in LinkedIn's professional forums that are focused on computer science and interactive entertainment technologies.

Emails will also be sent to academics in Trinity College Dublin, informing them of the research being conducted and providing instructions on how to take part.

3.6 Utterances

The utterances which will be rated by the participants, both PERSONAGE generated and handcrafted, will be single, stand-alone utterances and not in a conversational format. In other words, the previous utterance will be not be related to the utterance that is currently under examination. This is to avoid any manifestations of unintended personality perceptions that could occur as the result of a dialogue.

As in the original experiment, all of the utterances will be recommendations for restaurants in the New York City area. They will be presented to the participant with the understanding that they are to be rated as if they were spoken by a friend.

3.6.1 PERSONAGE Utterances

PERSONAGE will generate utterances in real time, as they are required by the applet. No modifications have been made to the PERSONAGE system by any of the researchers involved in this study. However, it does stand to reason that the system has evolved and changed since the original study, thus minor discrepancies are expected between the generation of the original and current utterances.

Each time an utterance is required, a restaurant will be randomly chosen from the database of New York City restaurants to be used as the subject for the statement. The opinion of the speaker toward said restaurant, which is portrayed through the utterance, is determined by four variables which are also stored in the database: *price*, *food quality*, *atmosphere* and *service*. Remarks are also sometimes made that refer to the restaurant's *address*, *neighbourhood* and *type of cuisine*.

Every utterance is generated using a different content plan. This is a slight variation from the original study in that sets of 20 utterances were generated using the same content plan. This variation is in the interest of keeping the study as close to what

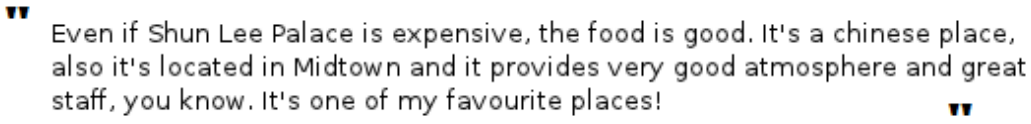

Figure 3.1: An example of a generated utterance, depicting a high level of extraversion.

may happen in a real world scenario as possible.

The way the utterance is presented is determined through a set of generation parameters, which are described in Tables 2.1, 2.2, 2.3, 2.4 and 2.5. Parameter XML files, that specify values for each of these variables, were provided along with the PERSONAGE system. Ten of these files were set up to account for the Big Five personality profiles, one file for each extreme of every trait. It is these parameter files that will be used to determine the presentation of utterances for this study.

In an effort to remain consistent with the original study, each parameter will be allowed to vary by a standard deviation of 15%. This will allow for the generation of multiple outputs from every content plan, trait and personality model.

An example of a generated utterance with an extrovert personality is shown in Figure 3.1

3.6.2 Handcrafted Utterances

The ideal set of hancrafted utterances would consist solely of utterances that were spoken by people with the related personality profiles. However, it would be very difficult and time-consuming to gather such data, and is, unfortunately, outside the scope of this research. Instead, contact was made with a linguistics expert to obtain expertly handcrafted utterances. Unfortunately, a set of expertly handcrafted utterances could not be acquired. Finally, in order to create the hand-written utterances, research into the Big Five personality traits and trait correlations with linguistic styles was conducted. The results of this are explained in Chapter 2. Ten utterances were crafted by the researchers; one to portray each extreme end of every personality trait.

The crafted utterances were also based on the database of New York City restaurants. Five restaurants were chosen from the database using a random selection program, to be used as the topics for the utterances. The details to be used as talking

Table 3.1: The handcrafted utterances for each extreme of the Big Five traits

Profiles	Study 0
High Extraversion	Provence is the best Italian place in Manhattan. It's expensive, but the food is amazing! The staff are extremely friendly and it has a fantastic atmosphere. I love going here with my friends.
Low Extraversion	Menchanko-Tei should be okay. It is inexpensive and the food, the service and the ambiance are good.
High Emotional Stability	If you enjoy Italian, you'll like San Pietro. I know I did. The place has good food, with a peaceful atmosphere.
Low Emotional Stability	Em.. Danal, I.. I think it's in Manhattan. I heard the food is mediocre, b-but I wouldn't know.
High Agreeableness	Cafe Centro? Sure, I can recommend it. It's a French place. The food is lovely. The price is expensive but it's worth it.
Low Agreeableness	Obviously, Menchanko-Tei isn't the best place to go. God, the service is damn bad and the food isn't anything special.
High Conscientiousness	Cafe Centro is in Midtown. I know it's got acceptable food, nice staff and it's cheap enough.
Low Conscientiousness	My friends recommended Provence to me. They said it's in Manhattan and it's got nice food ... and it's dear ... oh and it's better than most.
High Openness to experience	Why not try San Pietro? It's an Italian restaurant down in Midtown. I guess you will find the cuisine enjoyable and the staff are outstanding throughout the meal.
Low Openness to experience	I don't know... but I always eat at Danal's. It isn't pricy, but it's not the best.

points in the utterances were the same as were used for the generated setup. In this way, it is hoped that the utterances will be similar to each other and provide similar remarks about the restaurants in question.

One issue arises: while the generated utterances have the capability of presenting the information in more than a thousand different ways, the handcrafted utterances will always be the same. Implications may arise if a participant decides to take part in the study again, as they will already be familiar with the handcrafted utterances. Ideally, a participant will only see the study once. However, as the study is online there is no way of enforcing this rule, short of introducing a registration system, which would only discourage participation and work against the study.

The scripted utterances are shown in Table 3.1.

3.7 Questionnaire

The questionnaire is designed to collect the participant's rating of utterances. There are two questionnaires, one for the generated utterances and one for the handcrafted utterances. Each of these questionnaires has ten parts, one for each extreme of the Five Factor model, where the participants are asked to rate a single utterance.

Each part presents an utterance at the top of the page followed by eleven Likert

items, the first ten of which are part of the Ten-Item Personality Inventory, which is explained in Chapter 2, and are used to assess the personality type of the utterance. The last Likert item is used separately to rate how natural the utterance seems. The design of this modified Ten-Item Personality Inventory was introduced by Mariesse and Walker for their original study [26].

A Likert item asks a participant to rate their level of agreement with a provided statement, which, in this case, is targeted at the utterance. The statement is normally rated using options similar to the following: *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree* [46, p. 264].

The two questionnaires are presented as one set. As the complete questionnaire is made up of 20 parts, each of which has eleven ratings, the process may become time consuming and tedious. Therefore, it seems appropriate to ask the judges to save their ratings after intervals of four sections, in the hopes of minimizing the number of ratings lost if they decide to leave the survey. Intervals of four were chosen as it splits the entire survey into five even rounds, giving the participant closer targets to reach. A smaller interval was not chosen as it would become laborious for the respondents being asked to save so often. The saving process is made explicit to the participants, so that it is obvious when their information is about to be saved.

3.8 Avoiding Study Problems

3.8.1 Order Effects

This is a related study using the same participants for both experimental conditions. One advantage is that more data can be generated from a smaller participant pool. A second advantage is that any individual peculiarities which can be problematic in a small study, should be balanced out over both conditions. However, in a same subjects study, consistently presenting one experimental condition before the other presents difficulties. For instance, familiarity with the task may favour results in the second condition, or conversely in a performed task fatigue may have a negative effect in the second condition. Therefore it is standard practice to counterbalance the order of presenting the conditions [36, p. 21-22].

In this study, participants will not be presented with each condition separately.

Items from both conditions will be presented to each participant in an indeterminable order. As neither participants nor researchers will be aware which condition any given item belongs to, the study is referred to as being double-blind.

The indeterminate order for questionnaire section presentation is achieved through the following process. When a new participant begins the study, the applet contacts the server to find out the number of times each personality profile has already been rated. It then orders the profiles to be presented to the participant in order of priority, from the least rated to most. As the applet presents and saves the utterances in sets of four, which are populated from the priority list and then randomly ordered, it takes some time before a save is processed. If within this time-frame another participant begins the study they will receive the same priority ordering from the server, meaning that both participants will be rating the same four sets of utterances for the duration of the study. As it is unlikely that both participants will stay for the entire process, and are likely to quit at different stages, the priority ordering will be modified. For example, if person A rates 6 utterances before quitting, while person B only rates 5, the first 5 personality profiles will have been rated twice, while the sixth will only have been rated once. This means that the next time a participant enters the study they will receive a priority list with the profiles unrated by persons A and B first, followed by the profile that was rated once, followed by the profiles that were rated twice.

3.8.2 Experimenter Effects

Experimenter effects refer to any influence the experimenters exert on the subjects, whether intentional or unintentional, that may affect the results of the study. For example, Rosenthal mentions that as experimenters are always anxious to get a good result, the anxiety they exhibit can transfer across to the subject and affect the way they feel [47].

Banister et al state that the only way to ensure experimenter bias cannot influence a study is to prevent the experimenter from either knowing how the data was collected or meeting or communicating with the subjects [42]. They do, however, mention that the bias can be minimized through double-blind procedures.

The study reported in this paper, is double-blind. Also, the experimenters will not meet the subjects as it will be performed online. However, it is also expected that

friends and co-workers of the experimenters will also take part in the study, which may introduce some degree of experimenter bias.

3.8.3 Demand Characteristics

Often, a problem arises in studies where the subjects try to make sense of the research in their own way. They tend to make assumptions about the purpose of the study, and respond in accordance with these assumptions. If their suspicions are incorrect this introduces unintentional confusion to the results, while if their suspicions are correct they become too compliant. It is very difficult to prevent demand characteristics, and attempts to do so tend to introduce even less desirable effects, such as the destruction of the ecological validity [42].

Demand characteristics should not prove to be problematic in the study conducted for this paper. As participants are required to rate utterances in terms of personality profiles, the purpose of the study is made explicit to them from the outset. This precludes the possibility of participants forming their own theories. The unknown presentation order of utterance type, generated or crafted, maintains the integrity of the task.

3.9 Ethically Sound

Before initiating the study, ethical approval had to be received from the SCSS Research Ethics committee in Trinity College Dublin. To begin the approval process four documents were required: A proposal for the research project, a signed application form, a participant informed consent form and a participant information sheet.

The research project proposal details the study aims, methods, debriefing arrangements, participant related information and any ethical issues that may arise. It also mentions relevant legislation, such as the Data Protection Act 1998, and outlines how the study adheres to the required laws.

Both the lead researcher and supervisor were required to sign the application form, which was mainly concerned with the photographing, recording and misleading of participants and whether they were to be exposed to any physical or psychological distress or discomfort. Mention also had to be made if the intended participants were under

18, were patients or had intellectual or communication difficulties.

The participant informed consent form is to be presented to participants before taking part in the study. It outlines the study aim, what the participant will be asked to do, the amount of time it will take to complete, confidentiality information and withdrawal procedure. It also presents the participant with a set of bullet points which they must accept and agree with before taking part in the study. Normally, the consent form is signed by the participant, but in the case of this online study, the participants will instead be asked to enter their name in a textbox and click an "I have read the consent form and agree to take part in this study" button.

The participant information sheet simply describes what will be required of participants in more detail, and will be accessible through a web page before consenting to take part in the study.

The informed consent form and participant information sheet can be found in the Appendix.

Chapter 4

Implementing the Study

This chapter explains everything that is related to the implementation process. It reports upon the original implementation of the study before problems were identified, explains the modifications made to account for these problems and adresses the details concerning menu flow, internal structure of the program and database workings.

4.1 Original Implementation

As the study design required that the study be conducted in a questionnaire format, the applet was implemented as a 2-dimensional program that resembles a questionnaire. The main requirements of the implementation design were for the applet to have a simple layout and to be easy to navigate.

As all of the information and study components to be presented to the participants could be easily separated into distinct parts, the entire applet was implemented as a menu system, where each distinct part is presented through a single menu screen. The menu's flow is controlled using a simple screen management system, where only one screen is actively receiving input at a time, and, depending on the input, can transition to a different screen. The menu flow, or study flow, is depicted in Figure 4.1.

When a new participant navigates to the study web page, they are met with the first screen in the menu system, the Main Menu Screen, and a consent form. Before they are able to proceed with the study, they are first asked to fill in the compulsory fields and consent to the details outlined in the informed consent form, which can be

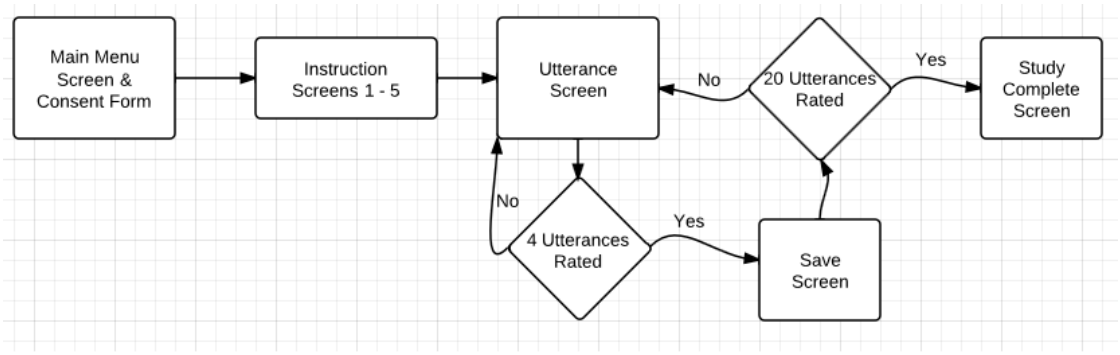

Figure 4.1: The applet's menu flow

found in the Appendix. After accepting the terms of the study, the new participant's ID and their related details are saved to the database on the server. They can then proceed to click the start button on the Main Menu Screen to begin the study.

Directly after the study begins, the participant is shown a series of instruction popup screens, informing them as to how to complete the questionnaires, how many utterances there are in total and how to save their ratings.

The questionnaire utterance rating screens follow the instruction set. At each of these screens, the participants are presented with an utterance and asked to rate it using the modified Ten-Item Personality Inventory. The Likert items are implemented as rating bars that have 7 ranges, from strongly disagree to strongly agree. The UtteranceManager determines the type of utterance to be retrieved, either hand-written or generated, and returns the required text. If it is a generated utterance the UtteranceManager uses the PERSONAGE system to generate the utterance in real-time. Conversely, the hand-written utterances are hardcoded into the system.

After each set of four utterance rating screens, the participant comes to a save screen. The save screen gives the participant the option to save the ratings of the previous four personality profiles. If they choose to save, the DatabaseManager contacts the MySQL database on the server, stores the new ratings and increments the related number of ratings for each of the processed profiles.

There are five rounds of ratings in total. If the participant has yet to complete all five, they are brought through the next set of four utterance rating screens. Otherwise they are brought to the study complete screen.

4.2 Modifications

After the complete applet had been implemented, and was found to be working as desired on the test machine, the time came to test the program online. Unexpected problems arose that had not occurred in the offline version.

One such problem was the method by which the applet was loading files. Files, such as images, were originally loaded by determining the directory using the *getCanonicalPath()* function call and appending the image name. This function, however, finds the absolute directory to the files, which includes the user's home directory. While this is not a problem running on the testing machine where the user, and the applet through the IDE, have administrator privileges, when the applet is downloaded to the clients machine from the server, a Security Exception is thrown.

One way to get around this Security Exception is to get the applet signed. Signing an applet, in a sense, is a way of saying "I have made this applet and it won't harm your computer". When an applet is signed, the user is shown a certificate before it loads. If they accept the certificate, the applet gains more permissions on the client's machine. A major problem with this approach, however, is that it is very costly to get an applet signed legitimately.

Fortunately, a better method of file loading was available. Instead of using the *getCanonicalPath()* to determine the directory, *ClassLoader.getResource()* was used. This new function can load files directly from the jar file, which is the compressed form the applet will take when it is put online, and bypasses the security exception in this way.

Having found the fix to this problem, it was discovered that the PERSONAGE system itself, and its accompanied libraries, were also loading files in a similar way that was unsuitable for applets. Instead of rewriting all of these systems, a decision was made to pre-generate the utterances using the original implementation, and hardcode them into the UtteranceManager, similar to the way the handwritten utterances are handled. Three utterances were created for each of the ten generated personality profiles, from which one is randomly chosen each time that profile is to be rated.

4.3 Textures

All of the textures required by the applet are loaded at the initialization stage. They are stored in a HashMap, using the filepath as the key and the image as the value.

A HashMap is a collection that is used to store data using a key-value pair. The key is used to look up whatever value is stored as its pair.

Whenever a texture is required, the path of the desired texture is passed to the HashMap and a reference to the image is returned. Although this method of resource handling takes more time at the loading stage, and uses more memory as the process is running, it makes the transitions between screens smooth, as no additional loading is needed. It also prevents a texture from being loaded more than once as each key in the HashMap must be unique.

The images are stored in the HashMap as BufferedImage types. The main difference between the Image and BufferedImage types, is that the BufferedImage provides a way to access the pixel data. BufferedImages were used as the storage type in anticipation of later implementing per-pixel collision detection for clickable regions in the applet. This type of accuracy was later deemed excessive for an applet of this calibre.

Any images that are used in the applet were either created from scratch or sourced, and slightly modified, from royalty free websites.

4.4 Double Buffering

Each time a frame of the 2D scene is drawn, the window is cleared and then all of the required images are drawn to it. This drawing process can take time, and if it isn't fast enough the eye picks up on the cleared screen colour and perceives it as a flicker. This artifact is undesirable, distracting, disorientating, and breaks the illusion of animation.

A technique commonly used to fix this artifact is called double buffering. Double buffering involves drawing all of the images in the scene to an off-screen image, or buffer, and then drawing the buffer directly to the computer screen. In this way the computer screen isn't cleared until the full scene is ready to be drawn to it in one step. While double buffering is usually implemented at the hardware level in graphics cards, it was implemented at the application level for this applet.

4.5 ScreenManager

The ScreenManager keeps track of the drawable screens in the scene, and passes any mouse input to the screen that is currently active. The screens are stored in an ArrayList, and the ScreenManager provides functionality to add and remove screens to or from this list. Although all of the screens in the list are drawn to the computer screen, only the most recently added screen is active at any given time.

If a new screen is added to the list, the previously active screen transitions off, before being removed from the list, while the new screen transitions on. How the screen will transition is defined in the Screen object. All of the screen transitions in this menu system are simple pan from left to right transitions.

The Screen class is an abstract class with function placeholders for updating, drawing and input handling. This means that any sub-class that has the Screen class as its parent, must provide its own functionality for updating, drawing and input handling. In this way many different types of screens can be created, however, the only types of screens needed for this implementation are menu screens.

4.5.1 MenuScreen

The MenuScreen type is a sub-class of Screen. When presented with a menu screen, users expect to see options from which they can choose how to progress through the menu flow. As such, the MenuScreen keeps a list of MenuEntry objects which can be clicked on from the menu screen. The MenuEntry objects have display text, a texture to use as a backdrop, a position on screen and a screen to add to the ScreenManager if it is clicked.

When creating a new type of MenuScreen, the new class becomes a sub-class of MenuScreen and populates the MenuEntry list in the loading phase. An example of this, the MainMenuScreen, which has only one MenuEntry, is depicted in Figure 4.2.

4.5.2 PopupScreen

The instruction and study complete screens are implemented as PopupScreens, which means that when they transition on, the previously active screen does not transition off, but still loses its active status to the newly added screen. PopupScreen is a subclass

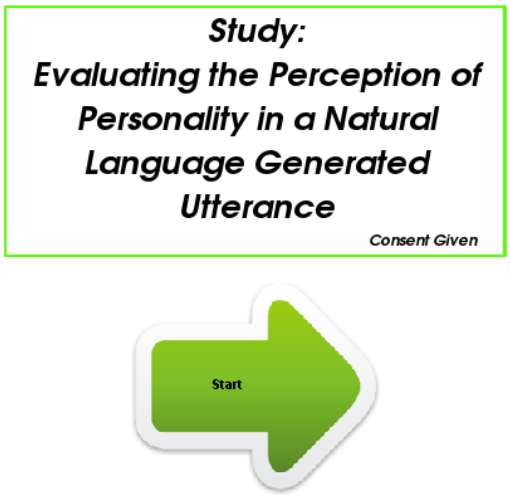

Figure 4.2: The Main Menu Screen

of MenuScreen, which defines menu options to skip to the final instruction or proceed to the next instruction or screen. It also presents the participant with one line of text and an image which can be used to illustrate a point.

4.5.3 UtteranceScreen

UtteranceScreen is a MenuScreen with one menu option, to progress to the next screen. It presents the user with an utterance, which is received from the UtteranceManager, and uses a set of RatingBar objects to represent the Likert items of the questionnaire. Each RatingBar presents a statement to the user and provides them with the means of judging their level of agreement with said statement through seven clickable boxes. Only one of the seven boxes can be selected for each RatingBar.

Each time a new UtteranceScreen is created a new set of personality profile ratings is added to the DatabaseManager. These ratings are modified any time a RatingBar's value is changed. After a set of four UtteranceScreens, and upon receiving permission from the participant, the ratings are committed to the database on the server.

An example of the UtteranceScreen can be seen in Figure 4.3.

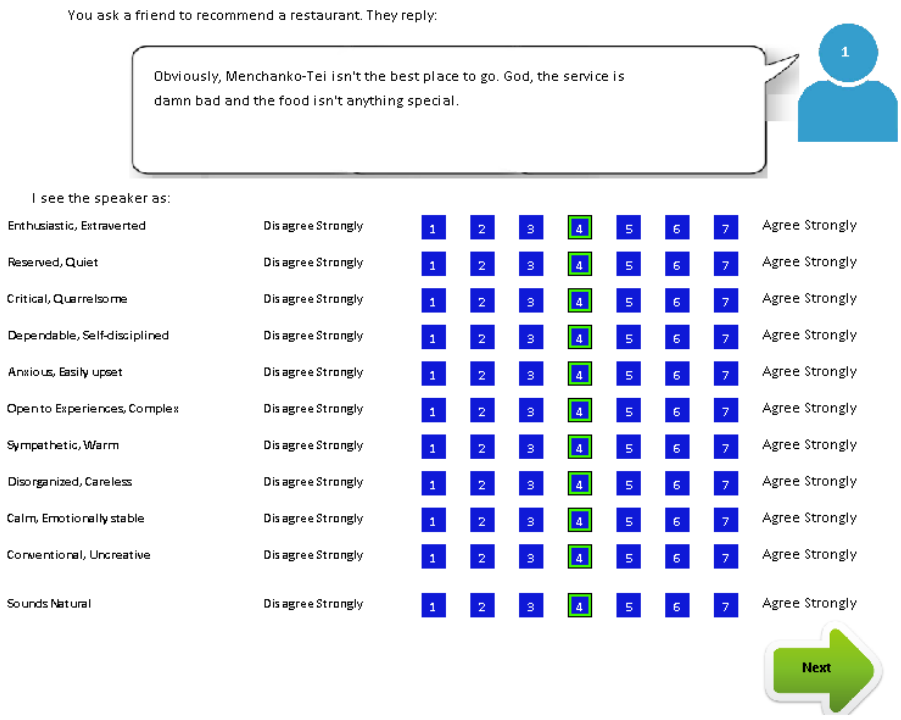

Figure 4.3: An Utterance Screen

4.6 DatabaseManager

The DatabaseManager takes care of all communication between the applet and the database server. The database is a MySQL database and therefore the mysql-connector-java library is used to create the connection.

The main function provided by the DatabaseManager includes: saving a new participant, and their related survey information, to the database and returning their unique ID, retrieving the number of ratings per personality profile and determining the priority order of utterances for the UtteranceManager, keeping track of the personality profile ratings that are to be saved and, most importantly, saving the personality profile ratings to the database. It also ensures that for each set of four utterances each personality profile is unique, i.e. generated and handwritten utterances of the same profile do not appear in the same set.

The database server uses three tables to keep track of the required data. One table called 'Participants' is used to store every new participant's age-range, gender,

experience with the Five Factor model, experience with NLG systems and whether they have dyslexia. A participant's name is not stored in the database in an effort to keep the data confidential. Instead, a unique ID is generated for each participant and used to match the survey results with the participant's ratings.

Another table is used to keep track of how many of each type of personality profile has already been rated. There are 20 columns in this table, one for each extreme of each trait, but only one row. When a personality profile is rated the value in this row with the corresponding column name is incremented by one.

A third table is used to keep track of the profile ratings which the participants rate using the modified Ten-Item Personality Inventory. There are columns to store the numeric ratings of all 10 traits and how natural the utterance seemed. The full utterance that was shown to the participants is also stored, along with the intended personality profile. Once again, the participant's ID is used to denote which ratings are theirs.

4.7 UtteranceManager

The UtteranceManager, upon receiving the profile priority list from the DatabaseManager, determines the utterances that are presented at each UtteranceScreen. When a new round starts, the UtteranceManager removes the first four profiles from the priority list and creates a smaller round list to insert them into. This new round list is then shuffled. When the UtteranceScreen asks for a new utterance, the UtteranceManager uses the utterance screen count, the place of the screen in the set of four, as the position at which the next personality profile is stored in the list. Using this personality profile, which specifies whether it is handwritten or generated, an utterance with the corresponding profile is sent back to the UtteranceScreen.

For each personality profile there is only one matching handwritten utterance. However, there are three generated utterances, hardcoded into the system, for each profile. When a generated utterance is required, an utterance is randomly chosen from the corresponding profile group of three.

4.7.1 PERSONAGE Implementation

Originally, in the offline version of the applet, the UtteranceManager generated the non-handwritten utterances in real-time. This was achieved by downloading the PERSONAGE source code from the website of the Natural Language and Dialogue Systems Lab, at the University of California, Santa Cruz, and compiling and packaging it into a library file so that it could be used within the applet [48].

Maven [49] was used, as outlined in the instructions which accompanied the PERSONAGE download, to compile and build PERSONAGE. However, before compilation could be completed, contact had to be made with CoGenTex, the owners of the RealPro realizer [35], in order to acquire a RealPro serial number for research purposes.

After the library file was added to the applet implementation, the samples included in the PERSONAGE download were studied in order to determine how to correctly use the rule-based PERSONAGE system.

To generate utterances in the restaurant domain, the PersonageMATCH property file had to be loaded. This specifies the directories that the system requires, including where the generation dictionary is stored. The generation dictionary defines the Deep Syntactic Structures that represent dialogue acts, as explained in Section 2.3.2.

XML files that specify the parameter values that control the personality profile of the utterance were also included in the download. Contact was made with Francois Mairesse to determine the correct files to use, which represent the high and low extremes of the Five Factor model. He confirmed that the XML files following the naming convention of "params-trait-extreme.xml" were the files to use.

With the natural language generator producing the desired utterances, 30 utterances, 3 for each trait extreme, were output to a text file so that they could be later hardcoded into the final applet implementation.

Chapter 5

Analyzing the Study

This chapter is concerned with the analysis and evaluation of the Study described in Chapters 3 and 4. The raw data will be presented first, followed by the t-test and inter-rater agreement findings. Finally, the results will be discussed and compared to the original study conducted by Mairesse and Walker. To make the presentation easier to understand the original study, presented in [4], will be referred to as Study 0 and the study documented in this report will be referred to as Study 1.

5.1 Raw Data

The study was conducted online for a period of 6 days. 16 participants took part in the experiment, and between them 236 utterances were rated. Every personality profile, high, low, handcrafted and generated, was rated 13 times, with the exception of handcrafted agreeableness. It was instead rated 12 times. This was unexpected as the ratings were saved in sets of four. This unexpected occurrence may be due to server downtime, which happened in the middle of processing the save request.

The percentages of recognized personality profiles for each utterance are presented in Table 5.1. As in Study 0, a profile was deemed correctly recognized if the average of both related rating items ended up on the same end of the scale as the intended profile [4]. For example, consider a participant rated an utterance, that was intended as highly extrovert, with 7 “enthusiastic, extroverted” and 3 “quiet, reserved”. The second rating item is scored in reverse, and as such is counted as being 5 (8 - 3). The

average of these two items is 6, which is on the high extrovert side of the scale, i.e. the average is greater than 4. If the average instead turned out to be less than 4, it is considered to be on the lower end of the trait. Otherwise, if the average is equal to 4, it is counted as being misrecognized.

High Extraversion seems to be the trait that is most easily recognized overall. The handcrafted High Agreeableness was also highly recognized (100%), which is much higher than the result of its generated counterpart (31%). The results suggest that the most difficult traits to identify from the generated utterances, in descending order starting with the most difficult, are High Openness to Experience (31%), High Agreeableness (31%) and High Emotional Stability (46%). All other generated traits were recognized more than 61% of the time. On the other hand, the most difficult traits to identify from the handcrafted utterances, again in descending order starting with the most difficult, are Low Emotional Stability (31%) and Low Conscientiousness (46%). All other handcrafted traits scored above 68%. On average, personality seems to be easier to recognize in the handcrafted utterances (74.6%) than in the generated utterances (59.5%), however there is only a 15% difference. Simply analyzing the percentages does not fully account for the variance in data, therefore a t-test will be performed to test for any significant differences between the generated and handcrafted ratings.

The average perceived naturalness for each personality profile is given in Table 5.2. An utterance's naturalness is scored on a scale between 1 and 7, where 1 implies the utterance does not sound natural at all, and 7 implies it is a completely natural sounding utterance.

Table 5.1: Correctly Recognized Personality Profiles (Rounded %)

Profiles	Generated	Handcrafted
High Extraversion	100	92
Low Extraversion	62	69
High Conscientiousness	62	85
Low Conscientiousness	62	46
High Openness to Experience	31	69
Low Openness to Experience	77	77
High Emotional Stability	46	85
Low Emotional Stability	62	31
High Agreeableness	31	100
Low Agreeableness	62	92
Mean	59.5	74.6

Table 5.2: Average Naturalness of Utterances (Scale 1 - 7)

Profiles	Generated	Handcrafted
High Extraversion	4.6	6.2
Low Extraversion	4.7	4.8
High Conscientiousness	2.8	5.3
Low Conscientiousness	3.3	4.5
High Openness to Experience	3.1	5.3
Low Openness to Experience	3.2	5.1
High Emotional Stability	3.4	6.3
Low Emotional Stability	2.7	5.2
High Agreeableness	3.0	5.4
Low Agreeableness	5.0	5.1
Mean	3.58	5.32

Once again, overall, it seems that High Extraversion tested the best. On average, it was rated as being moderately natural for generated utterances (4.6), and highly natural for handcrafted utterances (6.2). The most natural sounding of the generated utterances was Low Agreeableness (5.0), which is surprising as it was scored as the third least natural utterance in Study 0 [4]. The utterances that sound most natural of the handcrafted condition portray High Emotional Stability (6.3). While the least natural sounding of the handcrafted utterances is Low Conscientiousness (4.5), the generated utterances' naturalness varies between 2.7 and 5.0.

A mean of the naturalness averages suggests that the handcrafted utterances (5.32) are perceived as being considerably more natural than the generated utterances (3.58). This is supported by the fact that, for each profile, the handcrafted utterances rated more natural than the generated. A t-test will be conducted, to account for variance in ratings, and to determine whether the difference in naturalness is significant.

5.2 *t*-Test

A related t-test was performed for each personality profile to determine whether there is a significant difference between the ratings of generated and handcrafted utterances. The t value was calculated for each extreme of every profile, using the formula described in Section 2.5.1. The differences were determined by subtracting the generated profile rating from the handcrafted profile rating. The profile ratings were calculated by averaging the profile's related rating item scores. The resulting t values are presented in Table 5.3.

Significant differences were found between the handcrafted and generated profile ratings for 6 of the 10 profiles. They are, in order of increasing t, Low Emotional Stability ($p<.20$), High Conscientiousness ($p<.05$), High Openness to Experience ($p<.05$),

Table 5.3: Profile Ratings t-test (t values)

Profiles	Low	High
Extraversion	0.807	0.968
Conscientiousness	0.928	2.491
Openness to Experience	-0.456	2.127
Emotional Stability	1.606	3.464
Agreeableness	2.196	6.253

Table 5.4: Naturalness t-test (t values)

Profiles	Low	High
Extraversion	0.185	2.719
Conscientiousness	1.402	4.382
Openness to Experience	3.184	3.713
Emotional Stability	4.822	4.926
Agreeableness	0.121	2.504

Low Agreeableness (p<.05), High Emotional Stability (p<.01), and High Agreeableness (p<0.001). All of the significant differences' related profiles tended to be better recognized in the handcrafted condition, with the exception of Low Emotional Stability, which was, in fact, the least significant of the six.

A related t-test was also carried out for each personality profile to determine whether there is a significant difference between the naturalness of the generated and handcrafted utterances. The resulting t values are displayed in Table 5.4.

Out of the 10 personality profiles 8 significant differences were found, all of which favoured the handcrafted condition. The profiles, in order of increasing significance, are Low Conscientiousness (p<.20), High Agreeableness (p<.05), High Extraversion (p<.02), Low Openness to Experience (p<.01), High Openness to Experience (p<.01), High Conscientiousness (p<.001), Low Emotional Stability (p<.001) and High Emotional Stability (p<.001).

5.3 Inter-Rater Agreement

The inter-rater agreement was calculated, for each extreme of every profile, using Randolph's free-marginal multirater kappa [40]. This inter-rater statistic was chosen as it can deal with more than two raters. Brennan and Prediger also suggested using a free-marginal kappa if the study does not force a rater to assign a specific number of cases to each category [50], which is the case in this study.

For every kappa calculation there are two cases, which are the rating items that are

Table 5.5: Inter-Rater Agreement for Profile Ratings

Profiles	Generated	Handcrafted
High Extraversion	0.364	0.319
Low Extraversion	0.035	0.095
High Conscientiousness	-0.025	0.200
Low Conscientiousness	0.005	0.013
High Openness to Experience	0.088	0.125
Low Openness to Experience	0.185	0.08
High Emotional Stability	0.095	0.162
Low Emotional Stability	0.080	0.110
High Agreeableness	0.005	0.213
Low Agreeableness	0.028	0.205

related to the current personality profile, e.g. 'enthusiastic, extroverted' and 'reserved, quiet', and seven categories, ranging from strongly disagree to strongly agree. In the study the 13 raters, 12 for handcrafted Low Agreeableness, rated each of these cases into one of the seven categories. The number of cases in each category are added up and used by the kappa statistic to determine the extent to which the raters agree. A kappa value close to -1 implies complete disagreement below chance, a value close to 1 implies complete agreement above chance and a value around zero implies that agreement is equal to chance. The kappa values for each of the profiles are presented in Table 5.5.

By examining these results, it becomes apparent that, for most of the profiles, the agreement is almost equal to chance, with kappa values that are close to zero. This means that there is not a general agreement among the raters concerning these profiles. Some traits are slightly more agreed upon, such as generated and handcrafted High Extraversion (0.364, 0.319), handcrafted High Conscientiousness (0.200), and handcrafted High and Low Agreeableness (0.213, 0.205). However, none of the kappa results are significant enough, for any of the profiles, to state that there is an adequate agreement among the raters.

5.4 Discussion

In Study 0, Mairesse and Walker found that "Extraversion is the easiest trait to project in our domain", with an accuracy of 91.3% over both extremes of the trait [4]. The recognition percentages for the generated condition in Study 1 support this statement, as Extraversion is also found to be the most recognizable trait, with an 81% accuracy over both extremes.

Study 0 also found that High Openness to Experience was the least recognized trait, with an accuracy of only 55%, where all other traits scored 80% or above. Study 1 also found reason to believe that High Openness to Experience is one of the least recognizable traits, as it was only recognized 31% of the time. However, High Agreeableness was also found to be equally poorly recognized, which is in contrast to Study 0, where it scored 100% recognition. The recognition of other traits were also found to vary largely from their counterparts in Study 0, such as High Emotional Stability, High Conscientiousness, High Openness to Experience and Low Extraversion, which

had differences of 54%, 38%, 24% and 20.5% respectively. The overall average difference between the accuracy of Study 1 and Study 2 is 24.25%. The comparison of recognition percentages, between Study 0 and Study 1, are displayed in Table 5.6.

The poorer recognition percentages in Study 1, as well as the greater range of results, are likely to be either a result of a larger participant pool in Study 1, or the priming of participants with adjectives that are associated with each trait in Study 0. In either case, the new set of results can give an idea of how well PERSONAGE portrays the intended personalities in generated utterances, with a more representative sample.

The overall average perceived naturalness of Study 1 only deviates from the findings in Study 0 by 0.87, on the scale between 1 and 7. Although there was a notable difference in the number of raters between the studies, the naturalness was rated similarly for most traits, however, the percentage of recognized traits differs largely. This could suggest that the cause for the poorer recognition percentage in Study 1, as mentioned above, is more likely to be due to the priming rather than the number of participants, as there was no priming involved in the naturalness rating process. It could also, however, simply be due to the participants being familiar with judging the naturalness of an utterance, as it is a part of everyday human interaction.

While most of the traits were rated only slightly less natural than Study 0, some traits did have a notable difference. High Agreeableness was rated 2.56 less natural, High Conscientiousness was rated 2.53 less natural and, interestingly, Low Agreeableness was rated 1.43 more natural.

Table 5.6: Comparing Recognition Percentages (Study 0 [4] and Study 1)

Profiles	Study 0	Study 1	Difference
High Extraversion	100	100	0
Low Extraversion	82.5	62	20.5
High Conscientiousness	100	62	38
Low Conscientiousness	60	62	2
High Openness to Experience	55	31	24
Low Openness to Experience	90	77	13
High Emotional Stability	100	46	54
Low Emotional Stability	80	62	18
High Agreeableness	100	31	69
Low Agreeableness	70	62	8
Mean	83.75	59.5	24.25

Table 5.7: Comparing Average Naturalness (Study 0 [4] and Study 1)

Profiles	Study 0	Study 1	Difference
High Extraversion	5.78	4.6	1.18
Low Extraversion	4.93	4.7	0.23
High Conscientiousness	5.33	2.8	2.53
Low Conscientiousness	3.33	3.3	0.03
High Openness to Experience	3.85	3.1	0.75
Low Openness to Experience	3.98	3.2	0.78
High Emotional Stability	4.63	3.4	1.23
Low Emotional Stability	3.43	2.7	0.73
High Agreeableness	5.56	3.0	2.56
Low Agreeableness	3.63	5.0	1.43
Mean	4.45	3.58	0.87

In comparing the Big Five generated utterance results from Study 0 and Study 1, it can be seen that, on average, the participants in Study 1 recognized considerably fewer of the personality profiles, and judged them to be slightly less natural, than in Study 0.

Study 1 also found, through the use of a t-test, that there were significant differences in ratings between the generated and handcrafted utterances for Low Emotional Stability, High Conscientiousness, High Openness to Experience, Low Agreeableness, High Emotional Stability and High Agreeableness. For each of these trait extremes the null hypothesis can be rejected, and by finding the corresponding recognition percentages in Table 5.1, we can conclude that, for these traits, the handcrafted profiles were recognized better than the generated, with the exception of Low Emotional Stability, which was instead recognized better in the generated condition. For the other four traits, the null hypothesis, that there were no significant differences between ratings of the generated and handcrafted conditions, is accepted. These results support the PERSONAGE system, in that it can generate utterances that manifest personality for half of the Big Five trait extremes, that are rated as well as, or better than, the related handcrafted utterances. The average recognition percentage of these traits is 72.6% for the generated condition.

The fact that there are low profile recognition percentages in both the generated and handcrafted conditions suggests that raters may need more than a single stand-alone utterance to be able to correctly perceive the intended personalities. This suggestion is supported by the low level of inter-rater agreement among each of the profile extremes. Mairesse and Walker made a point of presenting the utterances in a stand-alone fashion, in an attempt to avoid invalid personality perceptions from emerging over the course of a dialogue [4]. However, in reality, people don't gain an understanding of other's personalities through a single utterance, but instead develop their understanding by experiencing things such as the person's dialogue, body language and facial expressions over time. It would be useful for future research to examine whether PERSONAGE proves more successful in the ecologically valid situation of repeated exposure to utterances rather than the emphasis on stand-alone utterances.

Study 1 also evaluated naturalness using a t-test. Differences of a significant level were found for all of the trait extremes except Low Extraversion and Low Agreeableness, and also favoured the handcrafted condition. While all utterances were expected

to sound more natural in the handcrafted condition, as they were written as if spoken by a real person, some were rated as low as 4.5, on average. This may suggest that what may sound natural in a spoken utterance may not seem natural in a written utterance.

Although, for five of the trait extremes, the PERSONAGE system can produce utterances with personalities that are recognized as well as, or better than, the handcrafted profiles, the only profiles that tested above 4 on the naturalness scale were High and Low Extraversion. Of these two extremes, only the High Extraversion had a moderate inter-rater agreement (0.364). While this is not an adequate level of inter-rater agreement, it is the best produced in this study. As such, of the generated condition, High Extraversion appears to be the trait that tested best overall, and is the trait which PERSONAGE can produce with the best mix of consistency, naturalness, and personality recognition.

Chapter 6

A Second Study

As this paper is primarily interested in the application of NLG to interactive entertainment technology, or, more precisely, computer games, it is useful to see if the results from Study 1 translate to the computer game domain. Therefore, the study will be repeated, with some minor adjustments to more closely simulate a gaming situation. This chapter will detail the modifications to the original study design for Study 2 and evaluate and present the new set of results.

6.1 Study Modifications

While in Study 1 PERSONAGE produced utterances that recommended restaurants, it does not seem to be very ecologically valid that the generator would be used for such a purpose in a computer game. Therefore, as suggested in Chapter 1, the generator will instead create utterances to recommend particular "in-game" quests to players. In order to change the domain in which PERSONAGE generates, a new generation dictionary and textplan had to be created. This is explained in the following subsection. The handcrafted utterances were also slightly modified to match the new domain. They can be seen in Table 6.1.

Given that the study was retargeted at the computer game domain, a participant pool was constructed consisting exclusively of computer game players, for a more representative sample.

The change in the sample also introduces a problem to the sampling process. Using

Table 6.1: The handcrafted utterances for each extreme of the Big Five traits

Profiles	Study 0
High Extraversion	King's Quest is the best quest. It's low level, but it's extremely easy! The enemies aren't hard, and it's really short. I love doing this quest.
Low Extraversion	King's Quest should be okay. It isn't long and the loot, the rewards and the experience are good.
High Emotional Stability	If you like grinding, you'll like King's Quest. I know I did. The quest has good rewards, and some attractive loot.
Low Emotional Stability	Em.. King's Quest, I.. I think it starts in Ravi's Inn. I heard the difficulty is mediocre, b-but I wouldn't know.
High Agreeableness	King's Quest? Sure, I can recommend it. It's a good quest. There's some lovely loot. It can be difficult but it's worth it.
Low Agreeableness	Obviously, King's Quest isn't the best quest to do. God, the difficulty is damn hard and the loot isn't anything special.
High Conscientiousness	King's Quest starts in Ravi's Inn. I know it's got acceptable loot, nice rewards and it's short enough.
Low Conscientiousness	My friends recommended King's Quest to me. They said it starts in Ravi's Inn and it's high level ... and it's difficult ... oh and it's longer than most.
High Openness to experience	Why not try King's Quest? It's a grinding quest down in Xavi's Inn. I guess you will find the difficulty enjoyable and the loot is outstanding throughout the quest.
Low Openness to experience	I don't know... but I always do King's Quest. It isn't difficult, but it's not the best.

a volunteer sampling technique, it becomes more difficult to gather participants who match the sample requirements. Therefore, purposive sampling was used instead. Purposive sampling is a technique generally used to gather participants for a study where there are a limited number of people with a required expertise. The researcher chooses the sample based on who they deem to be appropriate [44, p. 244]. While the new sampling technique means the researcher knows the subjects, which could introduce some experimenter effect, the participants are still required to complete the study in their own environment.

All other aspects of the study design are the same as in Study 1.

6.1.1 Modifying PERSONAGE's Domain

PERSONAGE's domain was modified by following instructions on Natural Language and Dialogue Systems Lab's website [51]. There were three major steps required.

First of all a new generation dictionary had to be created. A generation dictionary, as explained in Chapter 2, is made up of a set of Deep Syntactic Structures that represent individual dialogue acts, which are usually single sentences. Each dialogue act is stored in a single XML file. However, multiple Deep Syntactic Structures, each

representing the same dialogue act, but expressed in different ways, can be stored in the same XML file. Each of these DSS has an associated polarity and complexity value. In this way, PERSONAGE can determine which DSS to use to represent the dialogue act at runtime, depending on the speakers personality and by examining these variables. The RealPro manual provided an understanding of the command terms used to specify sentence structure in the DSS [52].

As the second study is interested in how well the system directly translates to the game domain, the DSS were only slightly modified. The overall structure remained unchanged. Only the words relating to the recommendation of a restaurant were changed to words relating to the recommendation of an in-game quest.

Secondly, a textplan XML file had to be created for the new generation dictionary. While the generation dictionary specifies how single sentences should appear, the textplan specifies relations between these single sentences, in order to express multiple dialogue acts in the same utterance. For example if a DSS expressed "Bistro is the best restaurant", and the textplan combined it with another DSS that specified "Bistro has great food", using the justify relation it could get combined into something similar to "Bistro is the best restaurant because it has great food".

A sample textplan, which was used to test PERSONAGE, was modified slightly in order to create a textplan for the desired domain. PERSONAGE actually creates the textplan at run-time in order to make the resulting utterances more restaurant specific. It is likely that a computer game would use a similar technique, however, for the purpose of this study a hardcoded textplan is sufficient.

Finally, a new property file had to be created that could then be loaded into PERSONAGE. The property file simply specifies the directories in which the generation dictionary, textplans and other required components can be found.

Shortly before initiating the study, it was discovered that synonyms for some of the newly added words could not be determined by WordNet. As time was limited, this problem could not be remedied before performing the study. Therefore, some discrepancies are expected in the results, especially regarding the utterances generated with a high level of openness or agreeableness, as they tend to use longer, less frequently used words.

6.2 Analysis

This study was evaluated in a similar fashion to Study 1. The raw data will be presented first, followed by the t-test and inter-rater agreement findings. All findings and results will then be discussed and compared with Study 1 towards the end of the section.

6.2.1 Raw Data

The complete set of 20 utterances was rated by a group of 4 computer gamers. In other words, 80 utterances were rated in all, and each personality profile was rated 4 times.

The percentages of correctly recognized personality profiles are shown in Table 6.2. As in Study 1, it appears that, overall, High Extraversion is one of the most recognizable traits. However in Study 2, Low Agreeableness matches it in both generated and handcrafted recognition percentages.

The results suggest that the gamers found the traits that are most difficult to identify, from the generated utterances, to be High Agreeableness (25%) and High Openness to Experience (25%). This may be due to the aforementioned synonym problem. However, as the two percentages are very close to those related in Study 1, within 6%, and are equivalent to each other, which is also the case in Study 1, they will still be used in the evaluation of this study. Of the handcrafted utterances, all of the gamers had difficulty recognizing Low Conscientiousness (0%), which could suggest a flaw in the modified handcrafted utterance. The average rating for Low Conscientiousness was

Table 6.2: Correctly Recognized Personality Profiles (Rounded %)

Profiles	Generated	Handcrafted
High Extraversion	75	100
Low Extraversion	50	50
High Conscientiousness	50	100
Low Conscientiousness	50	0
High Openness to Experience	25	75
Low Openness to Experience	75	75
High Emotional Stability	75	50
Low Emotional Stability	50	50
High Agreeableness	25	50
Low Agreeableness	75	100
Mean	55.5	65

Table 6.3: Average Naturalness of Utterances (Scale 1 - 7)

Profiles	Generated	Handcrafted
High Extraversion	4.25	5.75
Low Extraversion	5.5	5.75
High Conscientiousness	4.0	5.5
Low Conscientiousness	4.5	5.0
High Openness to Experience	2.75	5.5
Low Openness to Experience	4.75	4.5
High Emotional Stability	3.0	5.25
Low Emotional Stability	3.25	3.5
High Agreeableness	3.5	6.5
Low Agreeableness	3.75	5.25
Mean	3.925	5.25

4.375. All other utterances, both generated and handcrafted, scored above 49%. The mean recognition percentage for the handcrafted profiles (65%) scored almost 10% less than in Study1. The mean recognition percentage of the generated profiles (55.5%), however, only scored 4% less than in Study 1, leaving the mean recognition difference between the generated and handcrafted profiles at 10%.

The average perceived naturalness for each of the personality profiles is presented in Table 6.3. Extraversion, again, appears to produce the most natural of the utterances, overall. However, in this study the low end of the trait (5.0) seems to be more natural than the high end (5.625), when averaged between the handcrafted and generated versions. The profiles rated as least natural for the generated and handcrafted conditions are, respectively, High Emotional Stability (3.0) and Low Emotional Stability (3.5). Emotional Stability seems to be the profile rated as the least natural, overall. The profiles rated as most natural for the generated and handcrafted conditions, are, respectively, Low Extraversion (5.5) and High Agreeableness (6.5).

The mean of the naturalness averages suggests that the handcrafted utterances (5.25) sound more natural than the generated utterances (3.925). Naturalness was also rated higher for the handcrafted condition for all profiles, with the exception of Low Openness to Experience.

6.2.2 *t*-Test

For Study 2, a related *t*-test was carried out to determine whether there are any significant levels of difference between the gamer's ratings of each personality profile for the generated and handcrafted utterances. The calculated t values are presented in Table 6.4.

Only 4 significant differences were found out of the set of 10. In order of increasing significance level, they are High Extraversion ($p<.20$), High Conscientiousness ($p<.20$), Low Extraversion ($p<.02$) and High Openness to Experience ($p<.01$). All of these personality profiles have a higher recognition percentage in the handcrafted condition, with the exception of Low Extraversion, which, interestingly, has a rating of 50% in both conditions.

A *t*-test was also performed, for each profile, to find any significant levels of difference between the naturalness ratings of the generated and handcrafted utterances, the

Table 6.4: Profile Ratings t-test (t values)

Profiles	Low	High
Extraversion	4.899	1.647
Conscientiousness	1.608	1.698
Openness to Experience	0.775	8.660
Emotional Stability	0.522	0.0
Agreeableness	0.728	0.835

Table 6.5: Naturalness t-test (t values)

Profiles	Low	High
Extraversion	1.0	1.260
Conscientiousness	0.420	1.567
Openness to Experience	0.151	2.2
Emotional Stability	0.397	2.029
Agreeableness	1.732	4.243

resulting t values of which can be seen in Table 6.5. Again, only 4 significant levels of differences were found. They are, in order of increasing significance, Low Agreeableness ($p<.20$), High Emotional Stability ($p<.20$), High Openness to Experience ($p<.20$) and High Agreeableness ($p<.05$). Each of these profiles, also, seemed more natural in the handcrafted condition.

6.2.3 Inter-Rater Agreement

The level of inter-rater agreement, concerning the personality profile ratings, was evaluated as in Study 1, using Randolph's free-marginal multirater kappa [40]. The resulting kappa values can be seen in Table 6.6.

The inter-rater agreement analysis returned similar results to Study 1, in that most of the kappa values were close to zero. Therefore, the agreement that occurred wasn't significant and could have happened by chance. There are also a few traits that were slightly more agreed upon, handcrafted Low Emotional Stability (0.222) and Low Conscientiousness (0.222), and generated High Conscientiousness (0.222) and Low

Table 6.6: Inter-Rater Agreement for Profile Ratings

Profiles	Generated	Handcrafted
High Extraversion	-0.167	0.417
Low Extraversion	-0.069	-0.125
High Conscientiousness	0.222	0.125
Low Conscientiousness	0.125	0.222
High Openness to Experience	-0.167	0.167
Low Openness to Experience	0.222	0.125
High Emotional Stability	0.028	0.167
Low Emotional Stability	-0.069	0.222
High Agreeableness	-0.069	0.028
Low Agreeableness	-0.069	0.028

Openness to Experience (0.222). The most agreed upon trait was handcrafted High Extraversion (0.417), which has the highest value of kappa out of both the studies.

6.2.4 Discussion

In comparing recognition percentages of Study 2 and Study 1, it can be seen that the majority of traits, both handcrafted and generated, were recognized a similar percentage of times in Study 2 as they were in Study 1. A little difference is expected, as there were only four participants in Study 2, which means that each participant's correct recognition is worth 25%. Most of the trait recognition differences between the generated and handcrafted conditions also maintained their direction, i.e. the condition that was better recognized in Study 1, was also the better recognized in Study 2. The exception to this being High and Low Emotional Stability and High Extraversion. This could suggest that, for most traits, the intended personality is recognized a similar percentage of times in the new game domain as it was in the original restaurant domain.

As in Study 1 and Study 0, Extraversion was found to be, overall, the most recognizable trait. The least recognizable generated profiles were High Openness to Experience and High Agreeableness, which were also the least recognized in Study 1. Surprisingly, in both studies these two traits shared the same ratings, 31% in Study 1 and 25% in Study 2. Also of interest is that Low Openness to Experience had identical scores for the generated and handcrafted conditions in both studies, 77% in Study 1, and 75% in Study 2.

It is noteworthy that while Handcrafted Low Conscientiousness had a low recognition percentage in Study 1 (46%), it wasn't recognized by any of the participants in Study 2. This could be due to either a flaw in the modified handcrafted utterance, or it may just be an anomaly made possible by the small participant pool. In both studies the generated condition fared better.

There was a slight decrease in the mean recognition percentages between Study 1 and Study 2. However, they were fairly small differences, a 4% and 9.6% difference for the generated and handcrafted conditions, respectively.

The differences in the average perceived naturalness between Study 1 and Study 2, which are based on a scale between 1 and 7, are also minimal. In Study 2, there

was a 0.345 increase in perceived naturalness for the generated utterances and a 0.07 decrease in perceived naturalness for the handcrafted utterances. As with the recognition percentages, if a trait was perceived to be more natural in one condition in Study 1, it was also perceived to be more natural in that condition in Study 2, the exception being Low Openness to Experience.

In Study 2, a t-test revealed 4 differences of a significant level between the ratings of the generated and handcrafted utterance. The differences were for High Extraversion, High Conscientiousness, Low Extraversion and High Openness to Experience. The null hypothesis can, therefore, be rejected for each of these traits. Interestingly, while all but one of these differences again favoured the handcrafted condition, Low Extraversion was actually recognized 50% of the time in both conditions. This may be a result of the small participant pool. For the other 6 traits, the null hypothesis is accepted. This could suggest that, in the game domain, PERSONAGE generates utterances that portray recognizable personality profiles equally as well as the handcrafted utterances, for 6 of the Big Five trait extremes. The average recognition percentage of these utterances, however, is only 58.3%. Ideally, this study should be repeated, with a larger participant pool to confirm these findings.

A t-test was also used to determine any significant differences between the perceived naturalness of the generated and handcrafted utterances. The only significant differences found were for Low Agreeableness, High Emotional Stability, High Openness to Experience and High Agreeableness, all of which fared better in the handcrafted condition. In comparing the number of significant differences between Study 1 and Study 2, it was found that there were 4 fewer significant differences in Study 2. This might suggest that computer gamers rate the naturalness of the utterances in the game domain more tolerantly than participants in the restaurant domain. This could be due to the gamer's familiarity with, and acceptance of, a certain level of unrealistic behaviour in a gaming environment .

The overall level of inter-rater agreement was also found to be fairly low for Study 2. Handcrafted High Extraversion was, however, found to be the most agreed upon trait of the two studies.

As with Study 1, the trait that tested best overall, and the only trait that can be consistently and naturally conveyed through PERSONAGE, was found to be High Extraversion.

Chapter 7

Conclusions and Future Work

This chapter discusses future work that could be undertaken in light of this study and presents the conclusions reached from the research in this report.

7.1 Summary

Theoretically, the addition of a sophisticated dialogue system to a computer game, to control all non-critical, non-player character dialogue interactions with the player, could save a game company a lot of time and money, as vast amounts of dialogue would no longer require scripting. More impressively, it could allow the players to actually converse with the non-player characters, creating a more immersive experience. Role-playing computer games, or, more specificly, massively multiplayer online role playing games, would benefit greatly from such a system, as they tend to have large amounts of non-player characters that are very limited in their interactions. However, for a dialogue system to be sufficiently sophisticated for the game domain it must be able to produce responses that can manifest personality profiles while sounding natural. Otherwise, all the non-player characters in the game would present their responses in the same manner and would not be very believable.

Therefore, the aim of this research was to determine if a natural language generation system, that can portray personality profiles, is sufficiently sophisticated to be used as the natural language generator component in a dialogue system that controls non-player characters' speech, in a computer game.

The natural language generation system that currently seems the most applicable to the computer game domain was found to be PERSONAGE, which can generate utterances that portray the extremes of the Big Five personality traits.

Although PERSONAGE was previously evaluated, the study was only conducted with a maximum population of 3 judges, who were researchers in various fields and who were primed with adjectives that were related to each of the Big Five traits. The study did not have a high level of ecologically validity or a representative population. Therefore, a new study was designed, performed and evaluated to determine PERSONAGE's performance, with respect to its ability to correctly portray personality profiles in generated utterances, when compared against handcrafted utterances, and how natural the generated utterances seemed. The study was conducted online and had a participant pool of 16.

The results of this study, hereafter referred to as Study 1, found that, on average, the personality profiles were recognized 24.25% less often in Study 1 than in the original study, which will be referred to as Study 0. There was also a lot more variation in recognition percentages for each of the profiles in Study 1. This suggests that the priming of participants or the low participant pool in Study 0 may have introduced a bias to their study.

While similarities were found between the studies, such as Extraversion being the most recognizable trait and High Openness to Experience being the least recognizable trait extreme, it can be seen that, on average, the participants in Study 1 recognized considerably fewer of the personality profiles, and judged them to be slightly less natural, than in Study 0. The average of the correctly identified profiles in the generated utterances in Study 1 is 59.5%, which implies that the PERSONAGE system cannot produce utterances for which the intended personalities are consistently recognized, for all extremes of the Big Five traits. The only profiles which can be consistently recognized are High Extraversion, which was recognized by all participants, and, to a lesser extent, Low Openness to Experience, which was recognized by 77% of participants.

Study 1 also evaluated a set of handwritten utterances, with respect to the percentage of correctly recognized personality profiles, in order to determine whether there is a difference in the recognition percentages between PERSONAGE generated utterances and human crafted utterances. The results suggest that the handcrafted condition fared better than the generated condition for half of the trait extremes. However,

for the other half of the trait extremes, the generated utterances were found to be rated similar to, or better than, the handcrafted utterances. This suggests that, as the handcrafted utterances did not always fare better than the generated, and most of the traits could not be consistently recognized, more than a single stand-alone utterance is required to consistently portray personality.

Also, for the traits extremes mentioned above, the only traits to test above 4 on the naturalness scale, which is between 1 and 7, was High and Low Extraversion. From these results, it can be concluded that the only trait the PERSONAGE system can consistently produce that is both highly recognized and natural sounding is High Extraversion.

As this research is primarily interested in the application of a natural language generation system to a computer game, Study 1 was repeated, with some minor modifications to better simulate a computer game domain. This new study will be referred to as Study 2. The participant pool of Study 2 consisted of 4 computer gamers. The results of Study 2 were similar to Study 1, with respect to the recognition percentages and naturalness of utterances. Some discrepancies did occur, however, they were expected as a result of the small participant pool.

As with Study 1, the trait that tested best overall in Study 2, and the only trait that was found to be consistently and naturally conveyed through PERSONAGE, was High Extraversion.

7.2 Conclusion

In conclusion, as it stands, the number of personality traits that PERSONAGE can consistently convey through generated utterances, and that sound natural, are too few for the system to be a sufficiently sophisticated NLG component in a non-player character dialogue system. However, if the PERSONAGE system's ability to generate natural sounding utterances that manifest recognizable personality traits is improved, the system could in fact be directly applied to the computer game domain.

7.3 Future Work

In light of the research documented in this report, it would be useful to determine whether there would be any significant differences in PERSONAGE's performance given some of the following modifications.

Although Study 1 and Study 2 have been evaluated in a more ecologically valid scenario than Study 0, there is still room for improvement in the level of ecological validity. The mentioned studies have measured the level to which personality profiles, with respect to the Five Factor model, are correctly recognized when presented with a single, stand-alone utterance. While using single utterances minimizes unintended personality profiles from emerging over the course of a dialogue, it would be more ecologically valid to measure how well PERSONAGE can portray personality profiles through a dialogue of generated utterances, as this is what happens in reality, and is more than likely what will happened should PERSONGE be used in a dialogue system.

It would also be useful to repeat Study 1 with a set of handcrafted utterances that are sourced directly from people with the related personality profiles. This would increase the level of ecological validity in the study.

Concerning the naturalness of the utterances evaluated in Study 1 and Study 2, the generated condition was seen on average as being between not very and moderately natural. Surprisingly, the handcrafted utterances on average, although scoring higher than the generated condition, were not rated as highly as the researchers expected, considering that they were written by a real person. This could suggest that what may sound natural in a spoken utterance, may not necessarily seem natural in a written utterance. Therefore, it would be useful to conduct a study to determine if there are any significant differences between the naturalness ratings of PERSONAGE generated utterances, where there are three conditions: to rate the generated text utterances, to rate the spoken generated utterances and to rate the generated utterances which are presented using a Text-To-Speech system.

Also, participants may have expected computer generated utterances to be less natural and as they were unaware which utterance type they were rating, they may have been predisposed to perceive unnaturalness more strongly in all utterances. It would be of interest to determine, in a new study, whether utterances are rated as being more natural, where the participants are unaware of the presence of computer

generated utterances.

In this report, Study 2 was conducted to determine if the results from Study 1 can translate to the computer game domain. However, as it was late in the research process and time was limited, only 4 participants were found to take part. To ensure the findings are robust, the study should be repeated but targeted at a larger participant pool.

Following this, it would be useful to evaluate PERSONAGE under more ecologically valid circumstances in a computer game environment. As mentioned above, the system could be evaluated in a dialogue fashion, instead of using stand-alone utterances, and by using spoken instead of written utterances. However, in a computer game environment, it is likely that the player will also be able to see the non-player character that is the source of the PERSONAGE utterances. Therefore, to be ecologically valid when performing a study to evaluate PERSONAGE's ability to portray personality profiles in generated utterances in the computer game domain, it is essential to take the character's facial expressions and body language into account.

Appendix

List of Acronyms

NLG Natural Language Generation 01

TIPI Ten Item Personality Inventory 09

XML Extensible Markup Language 26

IDE Integrated Development Environment 34

DSS Deep Syntactic Structures 51

Page 1 of the informed consent form participants must accept in order to take part in the study.

Informed Consent Form (Web Page)

Evaluating the Perception of Personality in an Interactive Natural Language Generation System

Please read this document carefully before you decide to participate in this study

Aim: The aim of this study is to evaluate how well the personality intended by a natural language generator is portrayed in a produced expression, and perceived by its audience.

What you will be asked to do in the study:

In this interactive study you will be presented with a series of expressions, produced by a natural language generation system. After reading each expression, you will be asked to rate the personality of the "person" who produced it, with respect to six traits: Extraversion, Emotional Stability, Agreeableness, Conscientiousness, Openness to Experience and Naturalness. The traits will be explained when you are presented with the expressions.

Time required: There is no specified time for this study. You can quit at any time, and upon quitting you will be asked if you wish your ratings to be included in or withdrawn from the study.

Compensation: You will not receive compensation for participating in the experiment.

Confidentiality: Your identity will be kept confidential. This consent form, if accepted by you, will be kept locked separately from the data we collect during the experiment. It will not be possible to link the data we collect to your name. The data we collect will be analysed together with the data collected from other participants, and generalised results and conclusions will be drawn from these experiments will be submitted for publication at conferences and/or scientific journals. Your name will not be used in any report or article.

Voluntary participation: Your participation in this study is voluntary. There is no penalty for not participating.

Right to withdraw from the study: You have the right to withdraw from the study at any time without consequence.

Please check the following before signing:

- I understand that my participation in an experiment is entirely voluntary.
- I am informed about the purpose of the study.
- I understand that I can withdraw from the study at any time without having to give a reason.
- I understand that I may take a break/rest from the study at any time if I feel any discomfort or in any way tired.

Page 2 of the informed consent form participants must accept in order to take part in the study.

- I have been given the option of omitting questions that I do not wish to answer if a questionnaire is used.
- I understand that I should not participate in this study if I suffer from epilepsy
- All the information collected will remain confidential.
- I agree that my data is used for scientific purposes and therefore have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I acknowledge that the College requires that all research data be held for a period of ten years after the completion of a research project.
- Neither my participation in the experiment, nor the results obtained, have any bearing on my academic record. If I withdraw, this has no bearing on my academic record.
- I will be given answers to any question I have about the study.
- I am over 18 years of age

Statement of investigator's responsibility: I have explained the nature and purpose of this research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

Researchers Contact Details: Christopher McCormick, Computer Science, Lloyd Institute, Email: mccormc4@tcd.ie

Signature: _____

Date: _____

The participant information sheet which the participants are presented with upon navigating to the study.

Participant Information Web Page

Evaluating the Perception of Personality in an Interactive Natural Language Generation System

Researchers:
Christopher McCormick
Email: mccormc4@tcd.ie
Location: School of Computer Science and Statistics, Trinity College Dublin

In this interactive study you will be presented with a series of expressions, produced by a natural language generation system. After reading each expression, you will be asked to rate the personality of the "person" who produced it, with respect to eleven traits: Enthusiastic, Critical, Self-disciplined, Anxious, Open to new Experiences, Reserved, Sympathetic, Disorganized, Calm, Uncreative and Naturalness.

You can rate as many expressions as you like, and can quit at any time without penalty. Upon quitting you will be asked if you wish your ratings to be included in or withdrawn from the study.

Your results will be kept strictly confidential, stored using a unique ID number and the experimenter will not be able to identify your data or link them with your personal details.

The aim of this study is to evaluate how well the personality intended by a natural language generator is portrayed in a produced expression, and perceived by its audience.

This study is being conducted as part of my dissertation, in the pursuit of a M.Sc. in Computer Science (Interactive Entertainment Technology).

If you have any further questions, please do not hesitate to contact the experimenter at mccormc4@tcd.ie.

Bibliography

[1] E. Reiter and R. Dale, *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press, 2000.

[2] F. D. Rosis and F. Grasso, "Affective natural language generation," 1999.

[3] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the big-five personality domains," *Journal of Research in Personality*, vol. 37, pp. 504–528, 2003.

[4] F. Mairesse and M. A. Walker, "Towards personality-based user adaptation: psychologically informed stylistic language generation," *User Modeling and User-Adapted Interaction*, vol. 20, pp. 227–278, Aug. 2010.

[5] L. R. Goldberg, "An alternative description of personality: the big-five factor structure.," *Journal of Personality and Social Psychology*, vol. 59, no. 6, pp. 1216–1229, 1990.

[6] O. Rambow, S. Bangalore, and M. Walker, "Natural language generation in dialog systems," in *Proceedings of the first international conference on Human language technology research*, HLT '01, (Stroudsburg, PA, USA), pp. 1–4, Association for Computational Linguistics, 2001.

[7] L. Danlos, F. Meunier, and V. Combet, *EasyText: an Operational NLG System*. 2011.

[8] A. Belz, "Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models," *Nat. Lang. Eng.*, vol. 14, pp. 431–455, Oct. 2008.

[9] D. Jurafsky and J. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence, Prentice Hall, 2000.

[10] N. Youd and S. Mcglashan, "Generating utterances in dialogue systems," in *the Sixth International Workshop on Natural Language Generation*, pp. 135–150, Springer-Verlag, 1992.

[11] M. A. Walker, R. Grant, J. Sawyer, G. I. Lin, N. Wardrip-Fruin, and M. Buell, "Perceived or not perceived: film character models for expressive nlg," in *Proceedings of the 4th international conference on Interactive Digital Storytelling*, ICIDS'11, (Berlin, Heidelberg), pp. 109–121, Springer-Verlag, 2011.

[12] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? an experimental test of similarity-attraction," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '00, (New York, NY, USA), pp. 329–336, ACM, 2000.

[13] A. Tapus and M. J. Mataric, "Socially assistive robots: The link between personality, empathy, physiological signals, and task performance.," in *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*, pp. 133–140, AAAI, 2008.

[14] A. Isard, C. Brockmann, and J. Oberlander, "Individuality and alignment in generated dialogues," in *Proceedings of the Fourth International Natural Language Generation Conference*, INLG '06, (Stroudsburg, PA, USA), pp. 25–32, Association for Computational Linguistics, 2006.

[15] D. Fiske, "Consistency of the factorial structures of personality ratings from different sources.," *Journal of Abnormal Psychology*, vol. 44, no. 3, pp. 329–344, 1949.

[16] P. T. Costa and R. R. McCrae, *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five- Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources, 1992.

[17] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual Review of Psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[18] L. R. Goldberg, "Language and individual differences: The search for universals in personality lexicons.," vol. 2, (Beverly Hills, CA), pp. 141–165, Sage, 1981.

[19] O. P. John and S. Srivastava, "The big five trait taxonomy: history, measurement and theoretical perspectives.," *Handbook of Personality: Theory and Research 2nd edn*, pp. 102–138, 1999.

[20] R. Atkinson, *Hilgard's Introduction to Psychology*. Harcourt College Publishers, 2000.

[21] J. Pennebaker and L. King, "Linguistic styles: language use as an individual difference.," vol. 77, (Austin, Texas), pp. 1296–1312, 1999.

[22] M. Mehl, S. Gosling, and J. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," vol. 90, (Austin, Texas), pp. 862–877, 2006.

[23] S. Nowson, "The language of weblogs: A study of genre and individual differences," tech. rep., School of Informatics, University of Edinburgh, 2006.

[24] V. Benet-Martinez and O. P. John, "'los cinco grandes' across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english," *Journal of Personality and Social Psychology*, vol. 75, pp. 729–750, 1998.

[25] L. R. Goldberg, "The development of markers for the big-five factor structure," *Psychological Assessment*, vol. 4, pp. 26–42, 1992.

[26] F. Mairesse and M. Walker, "PERSONAGE: Personality generation for dialogue," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, (Prague), pp. 496–503, 2007.

[27] F. Mairesse and M. Walker, "A personality-based framework for utterance generation in dialogue applications," in *Proceedings of the AAAI Spring Symposium on Emotion, Personality, and Social Behavior*, (Palo Alto), 2008.

[28] D. Marcu, "Building up rhetorical structure trees," in *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2*, AAAI'96, pp. 1069–1074, AAAI Press, 1996.

[29] I. Mel'čuk, *Dependency Syntax: Theory and Practice*. 1988.

[30] O. Rambow, M. Rogati, M. A. Walker, and A. S. Labs, "Evaluating a trainable sentence planner for a spoken dialogue travel system," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 426–433, 2001.

[31] A. Stent, R. Prasad, and M. Walker, "Trainable sentence planning for complex information presentation in spoken dialog systems," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, (Stroudsburg, PA, USA), pp. 79–86, Association for Computational Linguistics, 2004.

[32] J. Lin, "Using distributional similarity to identify individual verb choice," in *Proceedings of the Fourth International Natural Language Generation Conference*, INLG '06, (Stroudsburg, PA, USA), pp. 33–40, Association for Computational Linguistics, 2006.

[33] C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[34] T. Chklovski and P. Pantel, "Verbocean: Mining the web for fine-grained semantic verb relations," in *Proceedings of EMNLP 2004* (D. Lin and D. Wu, eds.), (Barcelona, Spain), pp. 33–40, Association for Computational Linguistics, July 2004.

[35] B. Lavoie and O. Rambow, "A fast and portable realizer for text generation systems," in *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97, (Stroudsburg, PA, USA), pp. 265–268, Association for Computational Linguistics, 1997.

[36] J. Greene and M. D'Oliveira, *Learning to Use Statistical Tests in Psychology*. McGraw-Hill International, 2005.

[37] GraphPad, "*t*-test calculator," http://graphpad.com/quickcalcs/ttest1.cfm/, Last Accessed on August 21st 2012.

[38] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[39] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[40] J. J. Randolf, "Free-marginal multirater kappa (multirater kfree):: An alternative to fleiss' fixed-marginal multirater kappa," in *Joensuu University Learning and Instruction Symposium 2005*, 2005.

[41] J. J. Randolph, "Kappa calculator," http://justusrandolph.net/kappa/, Last Accessed on August 21st 2012.

[42] P. Banister, E. Burman, I. Parker, M. Taylor, and C. Tindall, *Qualitative methods in psychology: a research guide*. Open University Press, 1994.

[43] "Survey monkey," http://www.surveymonkey.com/, Last Accessed on August 20th 2012.

[44] V. Jupp, *The Sage Dictionary of Social Research Methods*. Sage, 2006.

[45] R. Rosenthal, "The volunteer subject," *Human relations*, vol. 18, pp. 389–406, 1965.

[46] N. Malhotra and M. Peterson, *Basic marketing research: a decision-making approach*. Pearson/Prentice Hall, 2006.

[47] R. Rosenthal, *Experimenter effects in behavioral research*. Century psychology series, Appleton-Century-Crofts, 1966.

[48] Natural Language and Dialogue Systems Group at UCSC, "Personage source code," https://nlds.soe.ucsc.edu/software/personage/setup/, Last Accessed on August 21st 2012.

[49] Apache, "Maven," http://maven.apache.org/, Last Accessed on August 21st 2012.

[50] R. L. Brennan and D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and Psychological Measurement*, vol. 41, pp. 687–699, 1981.

[51] Natural Language and Dialogue Systems Group at UCSC, "Personage domain modification," https://nlds.soe.ucsc.edu/software/personage/newdomain/, Last Accessed on August 21st 2012.

[52] CoGenTex Inc., "Realpro general english grammar user manual," 2000.