UNIVERSITY OF CALIFORNIA SANTA CRUZ

PLAYING WITH WORDS: FROM INTUITION TO EVALUATION OF GAME DIALOGUE INTERFACES

A dissertation submitted in partial satisfaction

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

Serdar Sali

December 2012

The Dissertation of Serdar Sali

is approved:

_____

Professor Michael Mateas, Chair

_____

Associate Professor Noah Wardrip-Fruin

_____

Associate Professor Sri Kurniawan

_____

Professor Marilyn Walker

_____

Tyrus Miller

Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Dialogue is central to many gameplay experiences, yet it remains widely unexplored and understudied. In this dissertation, we present our findings from a series of studies we conducted on dialogue systems in games. Our results point to deeper insights not only on dialogue system design, but also on deeper issues regarding how players experience games. Furthermore, we also design a survey instrument that can be used to evaluate user satisfaction in dialogue systems in games, and propose a set of quantitative metrics that can be used to evaluate player behavior. The result is a more formal and complete approach to evaluating this complicated design space.

ACKNOWLEDGMENTS

The road to getting a Ph.D. degree is long and arduous, and without the help of many loved ones it wouldn't be possible. I was so lucky to arrive at UC Santa Cruz when the Center for Games and Playable Media was being formed. Being a part of the Ezpressive Intelligence Studio broadened my horizons on what can be done with software, let me have fun while getting a PhD, and fall in love with my occupation again. I'm proud to have been a member, and will miss my days there dearly.

I'd like to thank my brilliant committee for their guidance and supervision. Thanks to Noah Wardrip-Fruin, who sowed the seeds of this project. Sri Kurniawan has been of great help with protocol, instrument and user study design. Marilyn Walker's work on the Paradise system has been one of the guiding references for this work. I'd also like to thank Steven Dow for his guidance and support during the earlier stages of this project – his ARFacade project was one of the inspirations for this dissertation. Thanks to Aaron Reed for taking the lead on writing the script for our menu-based interfaces. Thanks to Alexander McCaleb, Brian Kopleck and Robb Steel for all their hard work.

Thanks to many friends whose calls have not been returned or emails have not been replied to for months. I'll make it up to you all, I promise. A special thanks goes to Turhan for accompanying me during the countless hours spent working on my dissertation in various coffee shops in Santa Cruz, and for being a great friend during my last two years there.

Finally, huge thanks to my advisor Michael Mateas – he truly made this possible by extending a helping hand when I needed it most. I'll miss his uncanny ability to impart wisdom at the speed of light. I don't think I could ask for a better advisor.

One of the greatest luxuries in life is having people who you know love you unconditionally and will support you no matter what. I dedicate this to my family, who has been through this all with me.

1

INTRODUCTION

Dialogue is central to many gameplay experiences, yet it remains widely unexplored and understudied. While there has been extensive work on evaluation of task-based systems, current insights into this space in the games domain remains limited to conventional wisdom by game designers and scholars.

One important reason for the lack of studies in this area is that it presents a very complicated design space. For task-based systems, efficiency and task success are the main goals – consequently, design and evaluation has focused on optimization of those parameters. The main goal when designing virtual agents has been realism and believability – these goals are also ambitious and multi-faceted, but the goal is single, and evaluation methods have managed to detach themselves from those dimensions by relying entirely on human judgment of believability.

Game designers, on the other hand, design for a variety of goals. Some designers attempt to create experiences where players have a high degree of control over the experience. Some designers want their games to feel as realistic as possible, whereas others want players to feel completely immersed in the experience. Those goals are quite ambitious on their own: there are multiple ways to give players a high sense of control, make them feel more present or engaged in the game world, or impart upon them a high sense of realism and believability. Making things even more complicated is the fact that these goals can sometimes conflict with or contradict each other. Last but not the least, these qualities are highly subjective and players' perceptions of them vary greatly.

Despite all these challenges, however, if we want games to evolve into a more mature medium, we need to take the challenges presented by this complicated design space head-on. Our most potent form of communication, dialog, has barely made its way into games. It is my belief – and hope - that further inquiry into understanding this complicated space will pave the way for new gameplay experiences that we have yet to realize.

This thesis presents the results of my personal attempts to do so. Over the course of my PhD, I've conducted several user studies using the interactive drama Façade (Michael Mateas and Andrew Stern 2005) aimed at understanding and exploring this complicated design space. Starting with initial exploratory studies, my research has evolved to include quantitative instruments, information visualization techniques, and finally resulted in the development of a framework for evaluating dialogue

systems in games. In this dissertation I document this path, hopefully also contributing to a better understanding of how to do evaluation work in the complicated domain of games.

OBJECTIVES

My main goal when I started working on this area was, in very general terms, to understand this complex, intriguing design space. To that end, I set out to answer the following questions:

1. How do we evaluate different dialogue systems with regards to important properties of interactive experiences such as agency, flow, enjoyment or presence? Putting aside assumptions and conventional wisdom, what design issues and considerations actually exist in this space when we run controlled studies? What do they tell us about not only dialogue systems but also game design and player behavior?
2. What are the quantifiable aspects of player behavior? How do the changes in player perception and behavior reveal themselves in those metrics? How do they tie into important properties of interactive experiences mentioned above?
3. Finally, using these metrics, how do we come up with a formal evaluation framework for dialogue systems in games?

CONTRIBUTIONS

Our quest to answer the research questions outlined above led to the following contributions:

1. To my knowledge, this thesis presents the results from the first controlled studies on dialogue systems commonly employed in games. Using Façade as our test bed, we were able to compare different dialogue systems while allowing participants to play a complete story arc instead of a subsection of a game or a subplot that might not get resolved. In our first study, we compared menu-based systems with Façade's natural language understanding (NLU) system, whereas in our second study we implemented variations on Façade's NLU interface. These studies revealed interesting design guidelines and considerations that should be of interest to game designers and scholars.
2. Our results also give us deeper insight into game design. Specifically, we found support towards a more mature concept of agency that also takes into account system understanding and player perception. Also, our results seem to support a more goal-oriented design process where designers need to carefully consider what mediation they want to introduce in their interfaces in accordance with their goals. Rather than realism being the ultimate goal, explicit mediation introduces important design parameters.
3. Despite frequent dismissal from the game design community, our results point to support from players in favor of NLU interfaces. Players enjoy

interacting with NLU interfaces despite reporting many problems and difficulties.

4. We show how to make use of information visualization techniques to observe interaction patterns. We demonstrate the usability of our toolset by presenting two case studies of Façade and Prom Week.
5. With the help of knowledge we collected as a result of our aforementioned efforts, we propose a set of metrics and a formal evaluation framework based on the task-based dialogue system evaluation framework PARADISE (M. A. Walker, Litman, Kamm, Kamm, et al. 1997) for evaluating dialogue systems in games.

I believe the result of this work is a better understanding of not only dialogue system design but also player behavior in games and different aspects of game design in general. I also hope our methods provide some guidance on how to do evaluation work in the complicated interactive drama domain.

ORGANIZATION

This thesis is structured as follows: In Chapter 2, we present a brief overview of the various dialogue systems employed in digital games. Chapter 3 gives the unfamiliar reader some basic concepts regarding Façade's dialogue system and a summary of previous user studies on dialogue systems conducted using Façade. In Chapter 4, we detail the results of our very first study comparing various interface modalities on Façade in a controlled study. In Chapter 5, we present the results from another study we conducted on the effects of artificial pacing options on player perception and

behavior. In Chapter 6, the implementation details of a visualization tool we developed to study quantitative aspects of player behavior are explained, along with two case studies. Chapter 7 presents the results of another study in which we focused on collecting quantitative metrics, and use statistical regression methods on those metrics to come up with a framework for evaluating dialogue systems in games. Finally, Chapter 8 summarizes our conclusions and future plans for our work.

2

RELATED WORK

Dialogue systems are used extensively in a huge variety of application domains ranging from task-based systems to digital games. In this section, I will go over task-based systems, dialogue systems for virtual and believable agents, and dialogue systems in games and interactive stories. I will also briefly summarize existing evaluation frameworks and existing critique for these systems.

DIALOGUE IN TASK-BASED SYSTEMS

Dialogue systems have been widely employed in task-based systems that are designed for carrying out a specific task, such as making a flight reservation, paying bills or getting technical help services. These systems are usually focused on efficiency and successful completion as end-goals, and can significantly differ in the modalities they use for user input: They can be operated either by a textual interface or by speech (in this case they are called spoken dialogue systems) - there are also

multi-modal systems which use a combination of text and speech as input. In addition to differing modalities, these systems can also differ on how they allow user input, which might be through menus or a natural language understanding (NLU) module, which parses the user input, analyzes it to determine user intent, and decides on an appropriate reaction to the perceived meaning of user input. Finally, a task-based dialogue system might take the initiative itself, with the user responding to a series of questions that it asks, in which case it's called a system-initiative system; or in user-initiative system the user might be required to initiate the conversation; or it might be a mixed-initiative system.

Since these systems are strictly focused on task-completion, both from the users' and the system's perspective, users' experiences and satisfaction with the system are assumed to be directly correlated with how likely they are to complete their intended goal successfully. Their design, therefore, is focused on optimizing quantitative metrics such as task completion rate, completion time, number of irrelevant utterances, and so on. However, there is another class of task-based systems that try to maintain interest and usability by providing a positive experience for the user – virtual agents used in education and simulation, personal assistants found in mobile platforms, and virtual characters found in many games are prime examples. In the next sections, we further elaborate on these different systems and related design considerations.

DIALOGUE SYSTEMS FOR VIRTUAL AND BELIEVABLE AGENTS

Dialogue systems are also central to the goal of creating intelligent virtual agents. These systems can also employ the various types of dialogue systems mentioned above, but of particular interest to us are chatbots that employ natural language understanding. One earlier significant example is ELIZA (Weizenbaum 1966) – a brief excerpt from an interaction session with the DOCTOR script of ELIZA, which tried to create a satirical take on a non-directive psychotherapy session, is presented in Table 1 (Wardrip-Fruin 2009).

ELIZA manages to fulfill player expectations by both exploiting players' initial expectations for a non-directive psychotherapeutic dialog, and from "*the susceptibility of people to read far more understanding than is warranted into strings of symbols — especially words — strung together by computers*" (Hofstadter 1995) - the latter is called the ELIZA effect among computer scientists. Unless – or until - the conversation breaks down, ELIZA does a good job of acting like a therapist – or, more accurately, in line with its author's goals, a satire of a psychotherapy session. However it only employs very simple, stateless, pattern matching, and mostly responds to player input with a question or phrase obtained by applying syntactic transformations to the player utterance. This inevitably becomes apparent as the player starts typing more complex sentences.

Table 1. Sample interactions with ELIZA. Player statements are in italic, statements in uppercase are ELIZA's responses. While ELIZA manages to be convincing initially (a), it inevitably breaks down as players type utterances that aren't covered by the transformations and patterns built into the system (b).

Men are all alike.

IN WHAT WAY?

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE?

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

It's true. I am unhappy

DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY

I need some help, that much seems certain.

(a)

Can I ask you for help

DO YOU WANT TO BE ABLE TO ASK I FOR HELP

(b)

Of course, ELIZA is primitive by today's natural language processing standards. More advanced chatbots that employ more complex NLP algorithms and learning methods have been implemented since (Richard Wallace) (Jabberwacky, Rollo Carpenter) (Do-Much-More, David Levy et. al). Task-based NLP systems went on to gain greater acceptance in tutorial applications and task-oriented dialogue systems, where the domain for player input is usually small, making errors less likely.

DIALOGUE SYSTEMS IN GAMES AND INTERACTIVE STORIES

Dialogue has been central to many game genres as a gameplay activity. Interactive dramas, role-playing games and adventure games have extensively employed dialogue as part of gameplay as well as a storytelling tool. As a result, various dialogue systems, each with a different set of advantages and disadvantages, have evolved.

Probably the most commonly employed dialogue systems in digital games are menu-based systems, where conversation proceeds by the player picking up pre-scripted options from a menu, which might either be displayed to the players when their input is required, or might be initiated by the player with the push of a button. These systems usually differ in what is displayed to the player in the menus. Games like *Star Wars: Knights of the Old Republic* (Bioware 2003), the *Monkey Island* series (Lucas Arts 1990; Telltale Games 2009) and the more recent *Dragon Age: Origins* (Bioware 2010) display actual lines of dialogue in their menus. In our work, we classify these as sentence-selection interfaces.

Figure 1. Dialogue interface in Dragon Age: Origins.

A concern with conversational flow is one of the common objections to this type of dialogue interface. As Brent Ellison argues, "reading all the possible responses takes time and brings conversation flow to a halt" (Ellison 2008). Similarly, Lee Sheldon observes that "While [a sentence selection interface] gives the writer even more opportunity for character revelation, especially of the player character, it adds more text to read—one reason it only occasionally shows up in console games, and why designers are forever trying to find ways to shorthand it" (Sheldon 2004).

Sheldon's observation that sentence selection interfaces provide an opportunity for revealing information about the player character touches on two further pieces of design wisdom about such interfaces. First, as Sheldon points out, game writers can make player characters "far more witty, articulate, and wise (or boring, tongue-tied,

and stupid!) than the player himself.... We give him the chance to stand up in a conversation with Albert Einstein or Dorothy Parker or Dennis Miller and hold his own." In addition, Sheldon argues that sentence selection interfaces can, through well-written sentences, provide multiple dimensions of information about the choices players are making. This can include the topic, the approach to the topic, and how forcefully the character presents her case.

In pursuit of addressing these concerns, further evolution of menu-based systems have even made explicit use of additional indicators and symbols that will give the players clues as to what the outcome of speaking a certain line of dialogue will be. Perhaps the most common alternative to sentence selection interfaces is a different kind of menu that displays shorter, less fully realized conversational options. The display may be of discourse acts, topics, tones of response, partial responses, more diverse conversational actions, and so on. One of the best-known recent examples of such work is the dialogue interface for *Mass Effect* (Bioware 2007). The dialogue interface in *Mass Effect* reportedly went through 10-12 iterations, aimed primarily at speeding up the player's ability to choose responses (to preserve conversation flow) and secondarily at allowing the player character to perform lines without repeating aloud something the player had just read silently (Nutt 2007). The resulting interface is what we call an abstract-response interface, where instead of full sentences, the menus contain shorter, more abstract versions of what the player character will say, such as the tone or a brief summary of the response, expressing the gist of each line

of dialogue. With games like *Mass Effect* and *Indigo Prophecy* (Quantic Dream 2005) adopting menus that present a combination of these to the player, we can recognize this form as a prominent alternative to sentence selection.

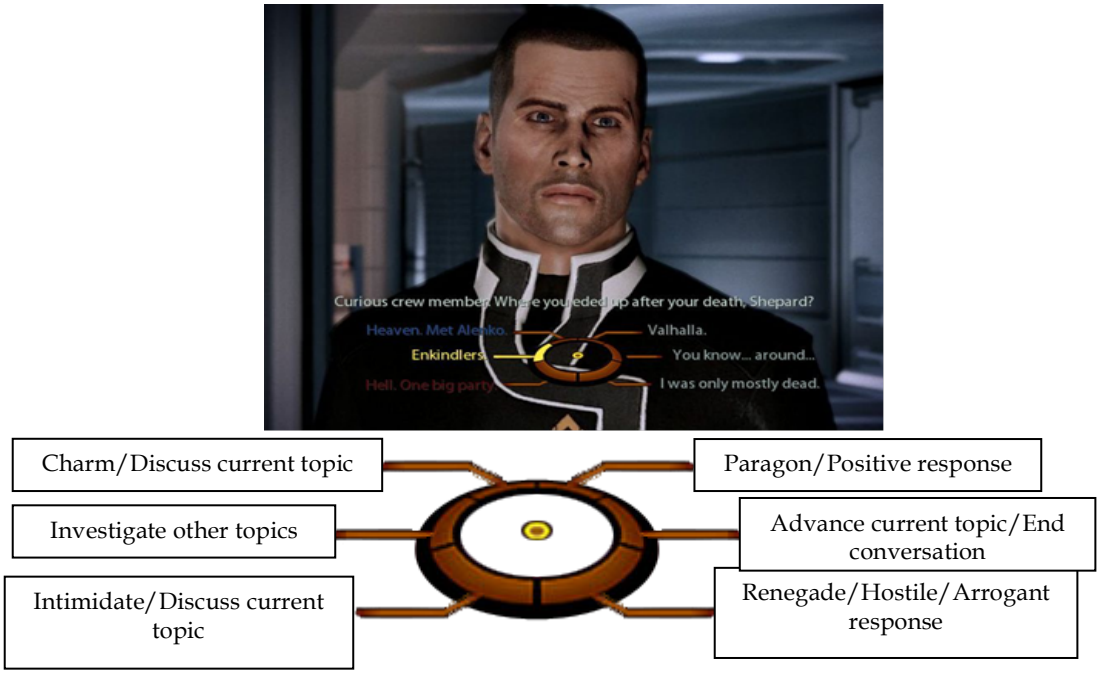


Figure 2. Dialogue interface in Mass Effect. The dialogue wheel makes use of spatial placement to give players clues on what the outcome of choosing an option will be.

In *Mass Effect*, responses picked on the dialogue wheel are short three-to-five-word sentences that give an idea of what the character will say. This provides greater conversation fluidity by providing quicker interactions and removing repetition caused by characters repeating what the player just read. In addition to presenting less text for the player to read, and making the thematic gist of each line of dialogue clearer, *Mass Effect*'s dialogue system also makes use of spatial placement as additional clues for the player by using a radial interface commonly referred to as

the "dialogue wheel," as pictured in Figure 2. Choices the player can make are placed on the left and right of the radial interface, and represent dialogue actions the player can take. Choices that tend to involve progressing through the conversation towards an end goal are placed on the right side of the interface. On the left side of the dialogue wheel, the player has options that can drastically sway conversations via hostile or friendly tones. In addition, the left side of the wheel allows players to further investigate situations, people, and the world. Following this spatial structure in most of the conversational exchanges allows the player to learn what the probable tone of their response will be even without knowing what the character will say.

Dragon Age 2 uses a radial wheel and abstract-response menu system similar to *Mass Effect*, but adds indicators to the middle of the dialogue wheel that attempt to demonstrate the nature and tone of the player's response. This, as depicted in Figure 3, allows players to better judge how the player character will respond when certain options are chosen on the dialogue wheel.

Icon	Description	Notes
	Diplomatic	Core dialogue choice for the Diplomatic personality type. Very rarely causes offense, likely to engender friendship increases.
	Helpful	Part of the Diplomatic personality type. Generally appears when Hawke is poised to volunteer his services to help someone or if sympathy or understanding is called for.
	Humorous	Core dialogue choice for the Humorous personality type. Rarely causes offense, tends to be a fairly "neutral" option where companions are concerned, though Isabela and Varric sometimes respond favorably.
	Charming	Secondary dialogue choice for the Humorous personality type.
	Aggressive	Core dialogue choice for the Aggressive personality type. Most likely to cause rivalry increases with companions.
	Direct	Secondary dialogue choice for the Aggressive personality type.

Figure 3. Dialogue interface of Dragon Age 2. The icon in the center gives the player hints on the tone of response.

Heavy Rain's dialogue system is another example of an abstract-response menu system (Quantic Dream 2010). In contrast to *Mass Effect* and *Dragon Age*, it uses a dialogue interface that consists of short 1-4 word options bound to the buttons on the

gamepad. These options are given to the player when the player is near specific objectives or characters, and are more strictly focused on conveying the pragmatic outcome of performing the dialogue action associated with that option.

Figure 4. Dialogue interface in Heavy Rain.

When the player chooses a dialogue option, the player character gives a full response in the context of the situation. The player usually has limited time to make a selection, which adds to the fluidity as the system allows for even quicker interactions and decisions due to the shortness and simplicity of the options. As depicted in Figure 4, the number of options ranges from two to four. The player is also occasionally confronted with pivotal moments where dialogue choices can clearly define the rest of the narrative. These options often reflect the tense nature of the situation by making the player's dialogue choices faded, blurry, or shaking. As depicted in Figure 5, during intense moments for the player character, dialogue options are rendered distorted and blurry to illustrate stress, fear, or other intense emotions. This adds to the intensity of situations and gives the player insight into how the player character is feeling in response to these dramatic situations.

Heavy Rain makes use of this system in some further interesting ways. At one point in the game, the player must deal with a religious fanatic who has been pushed over the top by the player character's partner. As the fanatic threatens to shoot your partner the player must make a decisive decision or risk deaths. By choosing the R1 option the player kills the fanatic and loses further evidence that would be gained by keeping him alive. By making this option clear, and keeping the rest shaking and blurry, the game effectively conveys the dilemma of taking "the easy way out" of the situation by shooting the fanatic versus following the more difficult and stressful

path of trying to convince a person in distress to drop his gun and calm down, perhaps risking the death of his partner and himself in the process.

Figure 5. Heavy Rain makes use of visual cues to reflect player character's mood to the player.

In all the abstract-response systems mentioned above, despite players having a vague sense of what the tone of their response might be, it's still hard to gauge what the player character is going to say based on the player's choice. There are often instances where the mapping between the options presented to the player in the menu and the actual outcomes feels unexpected and disconnected. Ellison writes of sentence selection: "There is no ambiguity in the player's decision." Sheldon says that, on the other hand, abstract response interfaces "can interject an immersion-harming game played between designer and player—What is my player- character going to say next?" Within the gaming community, the early level *Mass Effect*

"renegade" response (on Eden Prime) that resulted in physically hitting another character, without this being indicated in the interface, may have helped cement the impression that in this approach the mapping between selection and performance can be unpredictable for players.

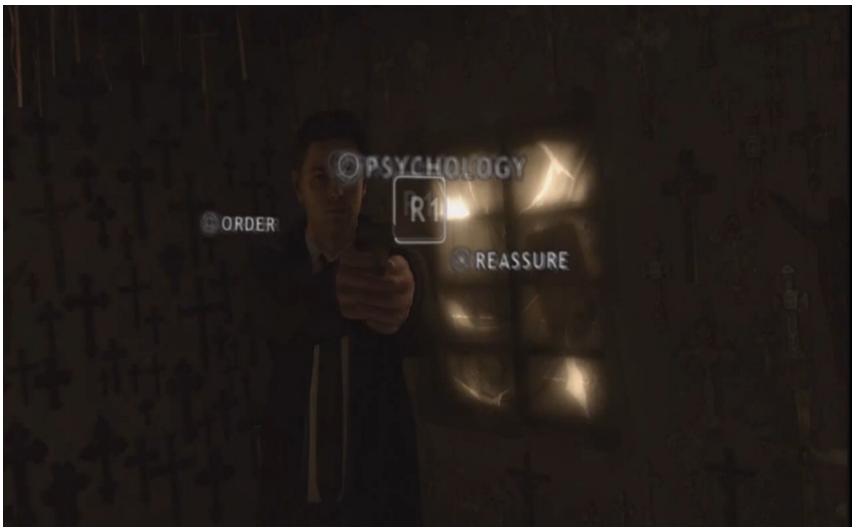


Figure 6. Heavy Rain also uses visual cues to manipulate the player in interesting ways. In this example, the "easy-way-out" of shooting the lunatic murder suspect is placed front and center.

Additionally, player enjoyment may heavily depend on making good use of the authorial control menu-based systems offer - if the options presented to the player and the possible outcomes for each dialogue exchange are always the same, dialogue may cease to be an enjoyable gameplay activity. One such recent example is *LA Noire* (Team Bondi 2011), which uses an abstract-response menu interface. In *LA Noire*, the player first selects a dialogue topic via the in-game journal that keeps track of information useful to the case, and after listening to what the NPC has to say about

the subject, may decide that the NPC is lying, telling the truth, or acting suspicious. Unlike games like *Heavy Rain* or *Mass Effect* where conversation options aren't strictly right or wrong, in *L.A. Noire* there's only one correct option at each step – the player can use the cues from the body languages and gestures of the NPCs to find the correct option, and is immediately notified on whether she made the right choice by a sound cue. The player can also spend intuition points to select the most commonly picked option among the online community of the game. The most frustrating moments occur when conversation turns into a guessing game trying to figure out how the designers dialog intent, including what option should be selected at each step and how evidence should be presented. Additionally, failing any dialogue segments make no difference as to the successful completion of the case. In solving the case, the players' decisions do not affect the ultimate outcome, only the speed at which the player arrives at the ending and the final grade for the case. As noted in a review of the game, "[Even if you get too many dialogue options wrong] you'll still end up completing the case successfully – thanks to branching outcomes, the game always finds a way, whether it's by introducing a last-minute witness or redirecting you back to some clue you missed – but you'll likely get yelled at by your captain for incompetence" (Mikel Reparaz 2011). The failures of *L.A. Noire*'s dialogue system point to the importance of making good use of the authorial affordances such interfaces offer, as well as the sense of nuance they can invoke through well-written and thoughtful options, which the underlying computational system can reasonably respond to as the outcomes of options are known at design-time.

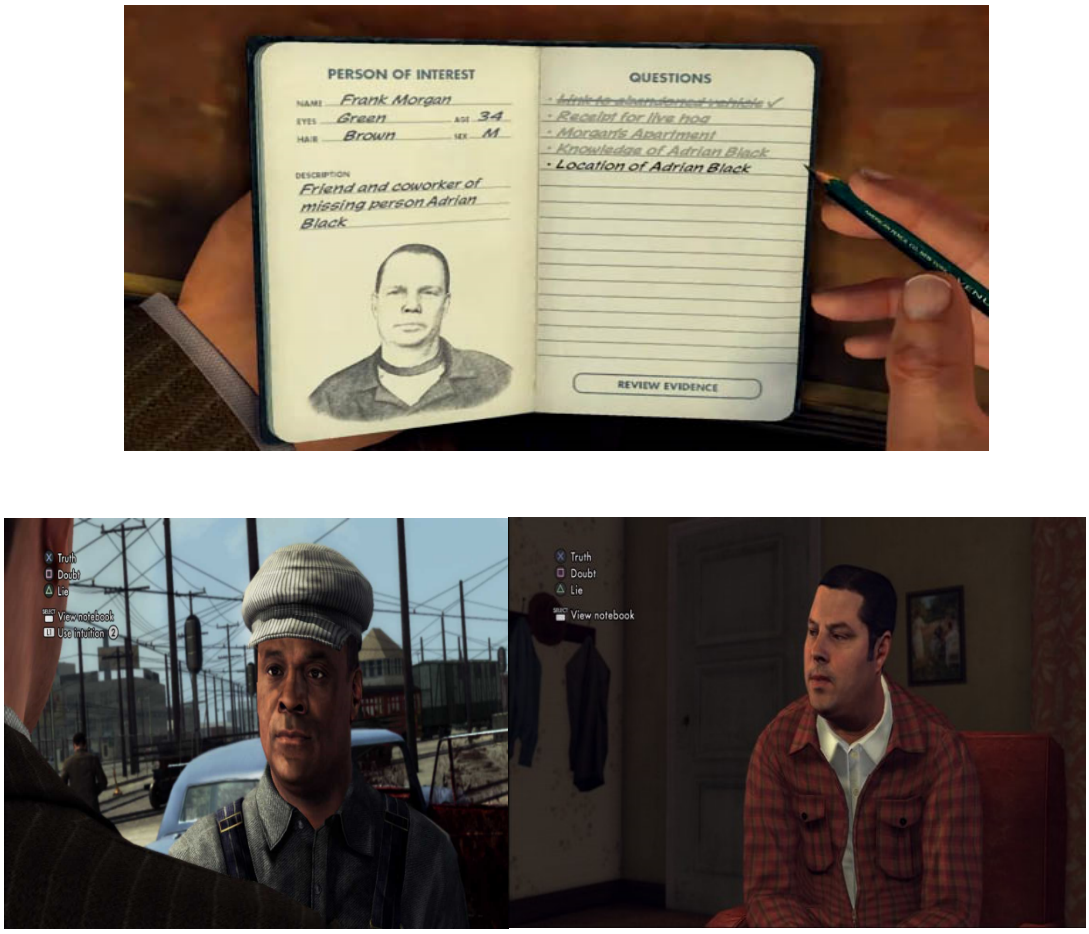


Figure 7. The dialogue system in *L.A. Noire*. In *L.A. Noire*, every dialogue action is reduced to a correct answer among the same possible three outcomes.

However, there are conflicting views as well: The fact that menu-based interfaces offer players enticing things to say (and enticing roles to play) points in the opposite direction of another piece of design wisdom about such interfaces. This is an outgrowth of the conventional "player character as transparent cipher" viewpoint that has led even story-rich games such as the *Half-Life* series to choose the odd conceit of a silent player character. As Richard Rouse puts it:

> *[W]hen players want to play games, often they want to play themselves. If the character they are controlling has a very strong personality, there is a distancing effect, reminding players that the game is largely predetermined and making them feel like they are not truly in control of what happens in the game. Particularly frustrating are adventure games that feature strongly characterized player characters who keep speaking irritating lines of dialogue. (Rouse and Ogden 2005)*

This brings us to a rarely employed type of interface in current commercial games: Natural language understanding (NLU) interfaces. By allowing (and attempting to interpret) free form textual input from players, NLU interfaces potentially enable a much greater range of player response than any single-depth menu could display. This type of interface is common in the independent game design/writing community of interactive fiction practitioners, who create games in which most actions are specified textually and interpreted by a parser. In most works of interactive fiction, this parser is also used for taking all physical action in the game. For the purposes of this discussion, we are only focused on the conversational interfaces. The community has also built up significant discussion of the issues such as misinterpretation problems or the problem of authoring responses for every possible player action (Short, *Conversation*) (Emily Short 2009). Earlier examples include the *Adventure* (Crowther 1975) and the *Zork* series (Infocom 1980) which rely

on very rudimentary parsers that can only recognize a very limited set of verbs such as *ASK [NPC] ABOUT [TOPIC]* or *TALK TO [NPC]*. This type of interface has been a staple of the interactive fiction genre, but there has been some experimentation on the complexity of the parsers, flexibility of user input, and the interconnectivity of gameplay and conversation. One such recent example is *'Mid the Sagebrush and the Cactus* (Victor Gijsbers 2010) in which the player can interleave combat and conversation and constantly make decisions on whether to shoot, try to reason with the character, or think and reflect on her own thoughts. *Glass*, by Emily Short (Short 2006), implements a system in which NPCs also have their own conversational goals and try to move the conversation along the best path to achieve those goals – the player character (a parrot) can address certain topics, and what is uttered next is determined by a combination of what the player character has said before, and what the conversational goals of NPCs are. *Alabaster* (Cater et al. 2009) has a complicated guidance system in the form of quips (actually realized snippets of conversation) for conversational topics that the player can explore. *Blue Lacuna* by Aaron Reed (Reed 2008), which takes place in an island with a resident hermit named Progue, eschews traditional verbs entirely in favor of shorter keywords that the player can type to further the conversation. In addition to specifically formatted keywords in the game text, the list of topics that the player can address is displayed to the user at the bottom of the screen, and topics can expire in a number of turns if they are no longer relevant or they are not vital to the progression of the plot.

The examples mentioned above use very rudimentary text parsers – in fact, it is hardly possible to classify them as NLU systems at all. The game community has rarely explored a true NLU system for conversation. Sheldon writes, "We're nowhere near ready to turn over conversations with major characters to AI" and characterizes this interface option as "outside the scope" of his book. Ellison argues:

> *[NLU interfaces] are rare in modern games for two reasons. The first is that the freedom they provide is extremely time-consuming to produce. The system needs hundreds of potential responses to accurately simulate a single, short conversation.*
>
> *The second reason is that even the most robust parsers frequently misinterpret the player's input. In Façade, an innocent inquiry can send the NPCs [non-player characters] into shock, horrified by what they thought the player just said. These misunderstandings ruin virtual relationships and frustrate the player, while at the same time exposing the program's failings and distracting the user from the interaction. (Ellison 2008)*

Within the game industry, there have been few attempts at implementing systems that attempt to give the player, at least on the surface level, the sense of freedom and flexibility that free-form input offers. *Deikto* by Chris Crawford is such an attempt (Crawford 2008). *Deikto* is a computer-legible mini-language that allows users to specify sentences from simpler subsets of phrases that the game understands. These

phrases can refer to relations, actions, moods or tones. While it doesn't offer complete freedom, provided that a reasonable range of language constructs is supported by the system, it can offer deep expression and interaction possibilities for conversation; in Chris Crawford's own words, *"Deikto may not be the most inspiring language to write sonnets in, but the singular beauty of interactive storytelling is not in its representation - it is in the richness, depth, variety and drama of the interactions it allows."*(Crawford). *Starship Titanic* by Douglas Adams (D. Adams 1998) was also an early attempt at an NLU interface, but the capabilities of the NLU module were severely limited - the system lacked any deep semantic processing and therefore mostly responded to keywords in player input.

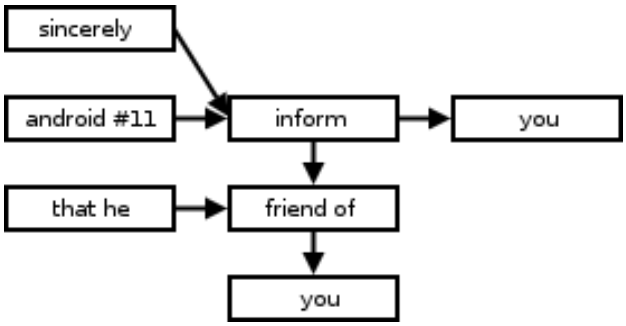


Figure 8. A sample Deikto sentence that says "Android #11 sincerely informs you that he's a friend of you."

Perhaps the best-known contemporary experiment with this type of interface is *Façade* (e.g., well-known industry commentator Ernest Adams writes, "*Façade* is one of the most important games ever created"(E. Adams 2005)). *Façade* is a one-act interactive drama where the player is supposed to help a couple work through their marital problems. *Façade* attempts to implement a conversation system that is as

natural and realistic as possible: the player is allowed to type in whatever they want to say - within a preset length limit - anytime, and the game parses player input and maps it to ~30 discourse acts that represent meaningful actions that a player is expected to take within the game. While *Façade* definitely has its shortcomings, it still remains one of the most notable examples of a practical application of an NLU system in a game, and its performance has been shown to be remarkably good despite the rudimentary parsing and the timing constraints that the system operates under (Mehta et al. 2007). This is part of the reason we used *Façade* for our work, along with its depth of conversational interaction and the fact that its underlying structure was amenable to supporting multiple dialogue interface options – we will discuss *Façade*, and previous studies performed using *Façade* as a test bed in more detail in the next chapter.

EVALUATING DIALOGUE SYSTEMS

There has been extensive research on formal evaluation methods for dialogue systems in the computational linguistics and HCI literature (McTear 2004). However, the focus is mostly on task-based NLU systems in this domain – therefore, high task completion rate and usability are considered to be the most significant factors contributing to user satisfaction. To that aim, various metrics have been established, such as word error rate and word accuracy[1], sentence understanding rate, contextual appropriateness, user and system turn correction ratio, response appropriateness,

[1] If the system includes a speech recognition component.

implicit and explicit recovery rates, transaction success rate and average number of turns for each transaction (Danieli and Gerbino 1995).

The main problem with using those metrics is that they don't generalize to different tasks and domains, and there's no way to tell, by using these metrics alone, how those metrics overlap or compensate for each other or how significantly each of them contributes to overall user satisfaction.

Evaluation methods for task-based systems: The PARADISE framework

The PARADISE framework proposed by Walker et al (M. A. Walker, Litman, Kamm, Kamm, et al. 1997) solves these problems by using a task representation that decouples what the task is from how the task is achieved, and performing linear regression on the efficiency and cost metrics based on external user satisfaction scores – the result is a general framework that allows evaluation of dialogue systems that employ different dialogue strategies and evaluation across different tasks, while at the same time calculating relative contributions from various cost and success metrics. The PARADISE framework been shown to be a very flexible framework, with applications to vastly different systems and extensions to even multi-modal architectures (M. Walker, Kamm, and Litman 2000; Beringer et al. 2002)

The PARADISE framework assumes that user satisfaction is maximized when task success is maximized, and the costs associated with the dialogue are minimized. The costs are determined by a combination of efficiency and quality measures such as

number of utterances, dialogue time, agent response delay, inappropriate utterance ratio, and so on.

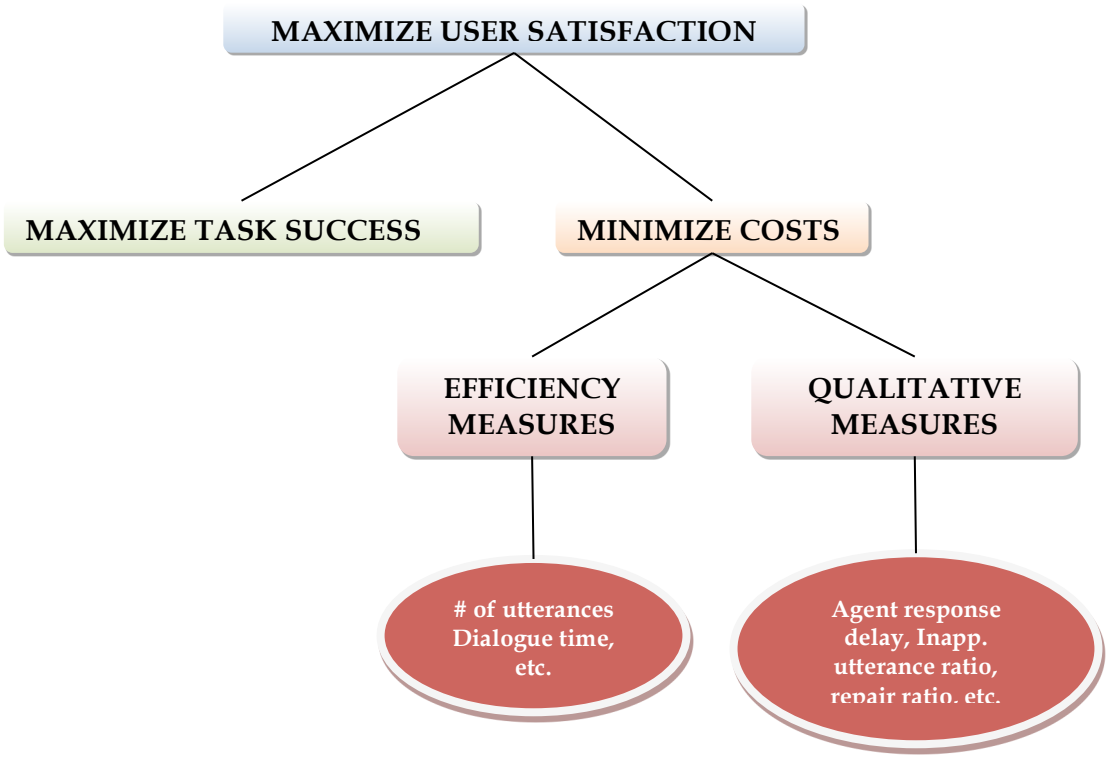


Figure 9. PARADISE's objectives and associated metrics. Figure reproduced from (Walker, Litman, Kamm, & Abella, 1997).

PARADISE assumes user satisfaction is correlated with system performance, which is a linear weighted combination of task success and cost:

$$P = \propto * N(k) - \sum_{i=1}^{n} w_i * N(c_i)$$

where N is a Z normalization, n is the number of cost metrics, c_i are the values of the cost metrics, and $\propto$ and w_i are the weights for success rate (represented here as k) and each cost metric c_i respectively. PARADISE then performs multiple linear regression to determine the weights of each metric, with the metrics as independent variables and the actual user evaluation of the system as the dependent variable. The resulting model is also predictive since it can predict user ratings from quantifiable metrics obtained from logs of users' interactions with the system. The framework we developed is inspired by PARADISE – we will discuss PARADISE in more detail in the related section.

Evaluating believable agents

The Turing test (French 2000) has long been considered as the standard evaluation method for evaluating chatterbots – although it's use has been a debated issue in the field (this discussion is beyond the scope of this book, but interested readers can look at (Pinar Saygin, Cicekli, and Akman 2000) (Scheiber, 2003)). The Loebner prize competition pits many AI entities against each other every year – while the gold and silver medals have never been won, *ALICE* and *Jabberwacky* have won the bronze award in recent years.

Evaluating dialogue systems in games

As we discussed above, previous work on evaluating dialogue systems in other domains have mostly focused on task-based metrics, since efficiency and productivity are key factors for those systems. However, games are experiences that

are meant to be enjoyed and thus are subject to different criteria. In this section we discuss important properties of immersive interactive experiences such as agency, flow, immersion and engagement in detail.

Agency

Agency was proposed by Murray as a fundamental aesthetic property of interactive experiences along with immersion and transformation (Murray 1998). According to Murray, agency is the "satisfying power to take meaningful action and see the results of our decisions and choices." Agency is now accepted as the most fundamental characteristic of interactive experiences, and it has been explored both in game studies and game design as an important concept and design tool.

While Murray's model of agency mostly addresses how an interactive experience should feel, another important question is how to design for maximizing agency. Church addresses this issue by proposing two abstract design goals aimed at achieving high agency when used in conjunction: intention and perceivable consequence (Church, 1999). In a game, the player should be able to make plans intentionally based on an understanding of game rules and current state of the game world, and these plans should result in visible, clear consequences: a game should allow and motivate plans which result in sensible and rationalizable outcomes in the game world.

Mateas, in Preliminary Poetics for Interactive Drama, further refined the concept of agency in the interactive drama domain by integrating it into the Aristotelian model of drama. As Mateas puts it,

> *"A player will experience agency when there is a balance between the material and formal constraints. When the actions motivated by the formal constraints (affordances) via dramatic probability in the plot are commensurate with the material constraints (affordances) made available from the levels of spectacle, pattern, language and thought, then the player will experience agency." (Mateas 2001)*

According to Mateas, the work itself should invoke desires for action that are actually supported by the game, and, when the player takes one of these actions, the game should respond in a way that makes sense to the player. This formulation also acts as a potential design guideline for interactive drama as it also hints at ways to design for maximizing player agency.

As a further addition to this formulation, Noah Wardrip-Fruin suggests *The SimCity Effect* as an alternative route to agency in his book Expresssive Processing (Wardrip-Fruin 2009):

> *"SimCity … begins with audience expectation – using it to evoke desires to take city planning actions using the tools represented on its surface. This initiates a process designed to transition players, through*

experimentation and feedback, from their initial assumptions to an understanding of its procedural city. This understanding is what enables agency in SimCity, and it accomplishes this at a level much more ambitious than simply moving through virtual space."

A high degree of agency is achieved when the transition from initial expectation evoked by surface representation to understanding the procedural model is smooth and coherent, and the mapping is consistent and in match with underlying system mechanics. The role of audience expectation in agency was also further revealed in recent work by Dow et al (Dow et al. 2007) in which the players were placed in an augmented reality version of the interactive drama *Façade*, with a control group playing the regular desktop version. Their findings suggest the increased presence offered by the AR version moved people out of the magic circle (out of a safe sense of mediated distance from the experience), and that the stress associated with this resulted in a decreased sense of agency.

We believe a synthesis of these models leads to a more mature concept of agency and a more useful model for future research (Wardrip-Fruin et al. 2009). In this work, we find further support for such a model.

Flow

Mihály Csíkszentmihályi proposed flow in his seminal book *Flow: The Psychology of Optimal Experience* (Csíkszentmihályi 2008). Csíkszentmihályi wanted to answer the question of why people pursue activities that result in no net gain to them in terms

of monetary value or other material gains – in other words, why people pursue certain activities just for their own sake. He interviewed various performers, ranging from chess players to surgeons, and as a result proposed the concept flow. Csíkszentmihályi defines flow as follows:

> *"We have seen how people describe the common characteristics of optimal experience: a sense that one's skills are adequate to cope with the challenges at hand, in a goal-directed, rule-bound action system that provides clear clues as to how well one is performing. Concentration is so intense that there is no attention left over to think about anything irrelevant, or to worry about problems. Self-consciousness disappears, and the sense of time becomes distorted. An activity that produces such experiences is so gratifying that people are willing to do it for its own sake, with little concern for what they will get out of it, even when it is difficult, or dangerous."(Csíkszentmihályi 2008)*

Csíkszentmihályi called such activities "optimal experiences" or "flow activities." Although Csíkszentmihályi didn't propose flow in the human-computer interaction domain, and it has been employed as a measure in a variety of fields from games to sports performance, looking at the definition it's easy to see how flow is an important part of *play* and *gaming* – and indeed, flow has been shown to be an important factor in player enjoyment of games (Weibel et al. 2008). Within the context of a game, flow addresses the most appropriate level of challenge, resulting

in the player playing the game just for the sake of it, with the feeling of a perfect match between her skill set and the level of challenge as the game progresses: A high level of challenge out of match with player capability results in anxiety, whereas a very easy game makes the player bored. A survey for measuring flow called the Flow Short Scale (FKS) was developed by Rheinberg et al. In this work, we used the English translation of this scale (Vollmeyer and Rheinberg 2006).

Immersion and Presence

The terms immersion and engagement are often used interchangeably – in fact, there's little consensus in the HCI community about what those terms mean and how they relate to each other, with many inconsistent and ambiguous definitions and results in different works.

Presence was first coined (as telepresence) by Marvin Minsky (Minsky 1980). Minsky formulated telepresence in the context of machinery operated by remote control. The term has since expanded to include similar experiences facilitated by all forms of media. It's now accepted that computer generated media is one of the main facilitators of presence, and presence has become an important focus of media studies. In that context, presence is defined simply as "being there in the mediated environment". Lombard and Ditton (Lombard and Ditton 1997) categorized six possible facilitators of presence as "*social richness (the 'warmth' or 'intimacy' possible via a medium), realism (perceptual and/or social), transportation (the sensations of 'you are there,' 'it is here,' and/or 'we are together'), immersion (in a mediated environment), social*

actor within medium (e.g., parasocial interaction), and medium as social actor (e.g., treating computers as social entities)." They then define presence as "*the perceptual illusion of non-mediation*".

According to Witmer and Singer (Witmer and Singer 1998), immersion and involvement are precursors to a sense of presence, but the definition of immersion remains unclear. Brown and Cairns (Brown and Cairns 2004) claim that immersion is achieved in three stages: engagement, engrossment and total immersion, where total immersion is presence. O'Brien recently proposed that there are six main components to engagement: *aesthetic appeal, novelty, focused attention, involvement, perceived usability,* and *endurability* (O'Brien 2008). Kim and Biocca (T. Kim and Biocca 1997) modeled presence as being influenced by two factors: "*arrival*," the sense of being present in the mediated environment, and "*departure*," not being present in the physical environment. They also found that being in the mediated environment is not equivalent to not being in the physical environment.

Following Lombard and Ditton's categorization, it is apparent that social interaction is an important source of presence in mediated environments. In games, this social interaction is usually achieved through some form of conversation between the mediator and the interactor. The specifics of the implementation of conversation, therefore, should be an important influencing factor in a user's sense of presence.

In this work, we are interested in how different dialogue system choices affect player experience with respect to agency, flow and presence. We have created semi-

structured interviews based on a review of literature, and employed standard surveys for measuring flow and presence. We will discuss our survey instruments and interview questions in depth in the relevant chapters.

3

A PRIMER ON FAÇADE

In this section, I'll briefly go over *Façade*, the interactive drama developed by Michael Mateas and Andrew Stern in 2005 (Michael Mateas and Andrew Stern 2005) which we used as our test bed for our experiments. I will discuss *Façade*'s story, the implementation details of the original NLU conversation interface, and go over the different interfaces we implemented in *Façade* for our experiments.

FAÇADE

Façade is an interactive drama developed by Michael Mateas and Andrew Stern as part of Michael Mateas' PhD work. *Façade* has been a groundbreaking experiment in interactive storytelling with its interesting attempts at merging storytelling and gameplay.

In *Façade*, the player is invited to a dinner party by old friends from college, Trip and Grace, who are now in the midst of an unhappy marriage. As the evening wears on,

small quibbles between Trip and Grace turn into a full-blown argument about their marriage and its future. It's up to the player how this night will end: Will Trip and Grace stay together and agree to talk over and try to resolve their differences, or will one of them leave? *Façade* doesn't force a single goal on the player, it's up to the player to decide whether she wants to keep Trip and Grace together by trying to calm them down and be the voice of common sense, blame either one of them for their marital woes, try to get them even angrier at each other, or just be a passive bystander.

One of the most interesting things about *Façade* is how it tells a non-linear story without making story branching points obvious to the player. It achieves this by employing a story-beat based system along with a drama manager that sequences the beats to achieve dramatic effect. Each story-beat is composed of many possible joint dialogue behaviors the selection and sequencing of which is determined by player input. The end result is a short story segment that changes depending on player interaction and behavior, while the drama manager sequences whole beats depending on the tension level associated with each beat and the story trace that the player has encountered. The characters Trip and Grace are driven by a complex behavior system implemented using A Behavior Language (ABL) (Mateas and Stern 2002). They can perform concurrent behaviors and have joint goals that they can perform synchronously, which may result in interesting emergent behaviors and situations.

Façade's story structure is composed of 5 main sections: the first part is the affinity game, where Trip and Grace put the player in a series of situations where they disagree on a certain topic, such as the decoration of the apartment or what they should have for drinks, and try to get the player to side with one of them. This is followed by a crisis beat in which the seemingly small arguments up to that point turn into a full-blown crisis about their marriage – which is then followed by the therapy game where Trip and Grace move on to discuss deeper and more fundamental problems about their marriage, and ask the player for help on those issues. Finally, after the therapy game ends, Trip and/or Grace can make a series of revelations about their real feelings and concerns that they have come to realize over the course of the evening with player's help, and then either one of them may decide to leave, or they may decide to stay together and agree to work more on their differences. In addition to this main story arc, the player can also address a variety of topics that are more loosely related to the game's domain: these might include referring to objects in the room, referring to satellite topics such as marriage, infidelity, career or divorce, and to hot-button topics, which evoke a reaction in the NPCs when the player brings them up a few times in a short amount of time, anytime during the story. The ability to intermix satellite and hot-button topics with conversations related to the main story line is made possible by the expressive freedom offered by the NLU dialogue system, which will be explained in more depth in the next section.

FAÇADE'S DIALOGUE SYSTEM

Façade uses a natural language understanding system for conversation. The player can type in any utterance any time within a fixed-length limit (Figure 10). The game will move on if the player takes too long to respond.

Figure 10. Façade's dialogue system. Façade implements and NLU system that allows players to type anytime.

Façade's NLU system is responsible for deciding what a player utterance means and how the characters should react to that statement. Its approach is rather pragmatic – it's more concerned in what the utterance implies in terms of the outcome on the game world than the syntax and semantics of the sentence. Dialogue is modeled as an exchange of discourse acts between NPCs Trip and Grace, and each player utterance is mapped to one or more discourse acts that are relevant within the current context, which is determined by the story beat that the player is currently experiencing.

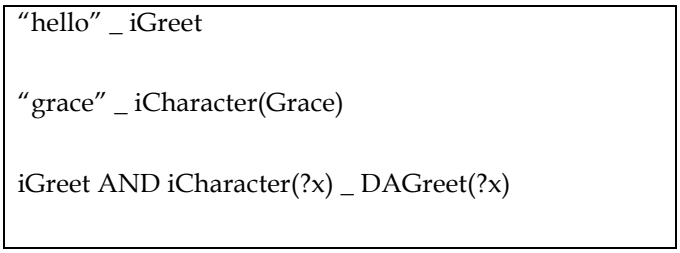


Figure 11. An example of Phase I processing in Façade's NLU system. The sentence "hello grace" is parsed word by word. The system knows hello is a greeting word, and Grace refers to an in-game NPC. It can then execute the rule in the third line to map this combination to a greeting discourse act from the player towards Grace.

Once an utterance has been mapped to discourse acts, phase II of NLU processing takes over and decides a reaction to the discourse acts chosen in the first phase. In addition to situations where player input is directly related to current beat context, player input can be mapped to global mixins, satellite topics, push-too-far reactions or discourse-act mixins. These actively increase the freedom of expression allowed to the player as they allow the player to pursue subplots that are not vital to the progression of the plot, or let the game maintain believable performance even when the player says something that is not expected or relevant within the current context.

Façade is perhaps the most well known example of NLU interfaces in games – and therefore has been the subject of extensive user studies, which we will review in the next section.

PAST USER STUDIES ON *FAÇADE*

NLU Accuracy

When discussing the performance of a dialogue system in a game, it's important to define what is meant by accuracy. In contrast to a task-based system where user

goals are much clearer and efficiency and task completion are the main important performance measures, in a game player strategies and goals can change constantly, and unexpected outcomes and surprising results can add to the experience. What defines a successful dialogue exchange is also related to player's interpretation of the outcome. In fact, a previous study by Mehta et al (Mehta et al. 2007) shows that when talking about the accuracy of an NLU system, it's useful to distinguish between actual accuracy vs. perceived accuracy. In this study, the authors explored the accuracy rate of both phases of the NLU system as explained above, and tried to uncover how player reaction changes in the face of communication failures, which might be a result of the failure of the NLU processing. For their study, they recruited 12 participants, and recorded videos of gameplay sessions and interviewed the participants using retrospective protocols. Data from AI logs and gameplay scripts were also employed in their analysis. Their results highlight important features of NLU systems and characteristics of player behavior when using NLU systems.

The authors' analysis of the phase I of the NLU system consisted of categorizing the system understanding for each utterance into one of the five categories: correct, which means the system correctly captured the semantics of the player utterance, wrong, which means the system completely missed the semantics of the utterance, doesn't understand, which means the system wasn't able to map the player utterance onto any of the discourse acts it supports, conflicting discourse act, which means the system mapped the player utterance to two different discourse acts that

are complete opposites, or typing issues, which occur when player makes a typing error, or tries to split a sentence into two to overcome length limitations. Their findings revealed that the phase I of the NLU architecture of *Façade* works relatively well, with an average 74% correct recognition rate (Table 2).

Table 2. Performance measures for Façade's NLU system. A significant portion (74%) of the utterances were mapped to correct discourse acts, showing that Façade's NLU system performs quite well despite its shallow processing.

Category	**Average**
Correct(%)	74
Conflicting discourse act(%)	3
Doesn't understand(%)	9
Wrong discourse act(%)	9
Typing problem(%)	5

The authors then analyze the perceived breakdowns of the system with respect to the technical shortcomings, and reach these main conclusions:

Narrative cues can help players rationalize conversation breakdowns. When a conversational breakdown occurs, players are able to use the narrative cues provided by the game to rationalize the characters' response to the player utterance by constructing elaborate backstories or justifications based on characters' personalities or current moods. This most successfully occurred when the characters deflected the player utterance, and when the NPCs performed a PushTooFar

reaction. In *Façade*, when the NLU system doesn't understand the player utterance, in order to maintain immersion and flow of conversation, the characters deflect instead of explicitly stating they don't understand the player. Since the authoring burden of providing varied responses to discuss each specific topic is too high, *Façade* also prevents the player from drilling down on a certain topic by using PushTooFar reactions, which are triggered when the player insists on talking about a certain topic. In this case, the NPCs usually say they don't want to talk about that topic – these reactions were usually rationalized as typical believable reactions to sensitive topics, and did not evoke a sense of failure in the player.

Shallow semantic understanding hurts engagement. This happens when the NLU system maps a player utterance to a discourse act that is not entirely disconnected from what the player meant, but still is not relevant to the player's intention, i.e. the NPCs started talking about the decoration or style of their room when the player referred to the view from the balcony, or when a phrase such as "not happy" gets mapped to "depressed". However, this kind of failure was usually observed to happen when the NPCs explicitly communicated to the player how they understood the player utterance, leaving little room for rationalization strategies. This observation points to an important design lesson for NLU systems: Players need to be subtly trained in what the underlying system supports and what level of understanding it has.

Believable NPC performance can still maintain engagement even in the face of perceived failure. Even when the players become explicitly aware of a failure in the underlying system, the NPCs' responses can still be entertaining and interesting, therefore maintaining an engagement in the experience. In fact, further studies have supported these findings and reveal that the novelty of an NLU interface and the high degree of "playability" resulting from the freedom the interface offers to players results in a software toy that is fun to play with. I will talk about these results in-depth when I discuss our studies.

Coping with failures

An important study conducted on the failures of the NLU system in *Façade* is by Knickmeyer et al (Knickmeyer and Mateas 2005). In this study, the authors looked at how players behave in the face of interaction failures of the NLU system. In their study, the participants were asked to play *Façade* twice. Players' gameplay sessions were videotaped, and then the videos were analyzed and coded according to a coding scheme developed by the authors. After the play session, the participants were also given a script of their gameplay session and interviewed while watching their gameplay video. Then, using the data, they plotted Gannt diagrams and discovered three main patterns of player response to conversational breakdowns (Figure 12).

One of these patterns is ***background interest***. This happens when Trip and Grace misunderstand the players, but as a result, reveal interesting info about the

background story, which results in players changing strategies to pursue this new interesting story piece. This shows how the freedom and flexibility offered by the NLU system allows participants to pursue new goals in an opportunistic manner even in the face of communication breakdowns.

Another interesting pattern is the ***player affective response***, where an inappropriate response by one of the NPCs was attributed not to the shortcomings of the system or the game, but to a flaw or trait of the NPCs character or background. This is also an interesting side effect of NLU systems: when players are truly immersed in the game world and the characters, the ambiguities of language itself might help them rationalize the failures in the context of the game world and the character traits of the NPCs.

The final pattern was called ***meta-play***, where the players explicitly recognized the failure of the NLU system, and shifted their strategy accordingly. These shifts happened two ways: they either abandoned their old strategy and moved on to a new one, or tried to "repair" their old strategy and get back on track. This also shows how the flexibility of an NLU system gives the players more freedom in reformulating and repairing their strategies, an obvious advantage other systems described above mostly lack.

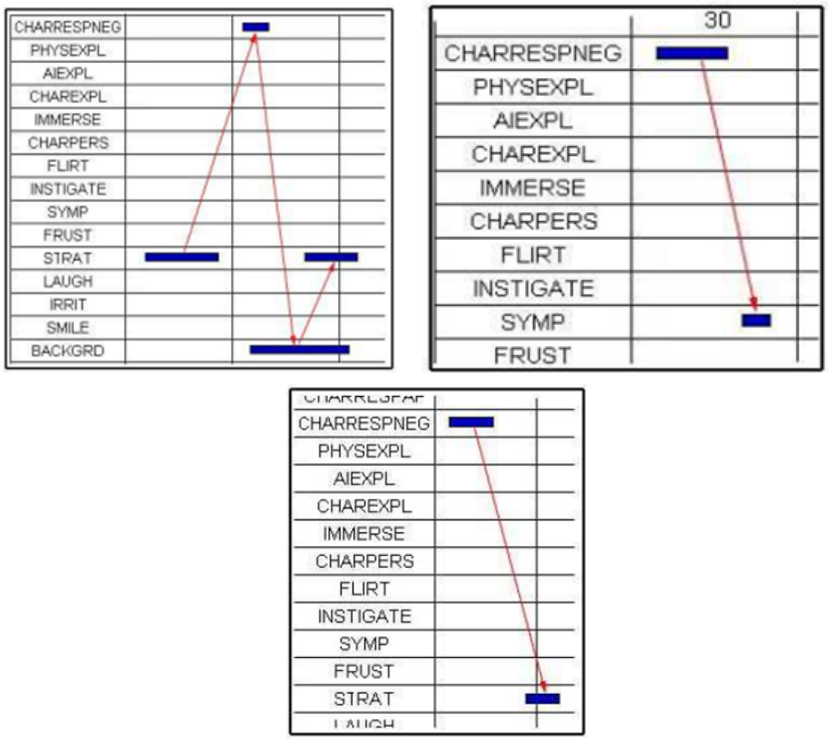


Figure 12. Coping strategies for Façade's NLU system's failures: Background interest (top left), player affective response (top right), and meta-play (bottom).

Effects of mediation on presence and engagement

Another interesting study on dialogue systems by Dow et al (Dow et al. 2007) looked at how various degrees of mediation affected presence and engagement in an interactive drama. To investigate this the authors built an augmented reality version of *Façade* called *AR Façade*, in which the player interacts with the couple Trip and Grace in a physical space decorated identically to the couple's apartment, and Trip and Grace are superimposed on this physical space through a head-mounted

display. Dow et al compared three different versions of *Façade*: *AR Façade* (the augmented reality version, where players use speech and physical gestures to interact), the original desktop version of *Façade* in which the player types to speak, and a speech input version in which the player speaks to a microphone. A wizard operator typed player statements into the system in the AR and speech input versions (Figure 13).

They recruited 12 participants, balanced across genders, ages, races and education levels. Each participant played all three versions in a randomized order, and then answered open-ended interview questions. The interviews were then transcribed and coded and categorized using Grounded Theory (Martin and Turner 1986).

Perhaps the most important finding of this study is that ***increased presence doesn't necessarily lead to increased immersion or better gameplay***. The increased sense of presence in the AR version led to some players feeling too close to the tense situation between Trip and Grace. The setup felt more like a real-life situation than a game, and some participants found it hard to embody a certain character fully in the physical space, with the gestures, behaviors and body language, whereas in the desktop version this required just embodying a certain conversational style. As a result, for some participants the experience stopped being enjoyable. Half of the participants ended up preferring the desktop versions to the AR version, because they desired some distance from the experience. This suggests that ***more natural interfaces do not necessarily result in more compelling entertainment experiences –***

in fact, we found a similar result when we compared the desktop NLU version with more traditional menu-based dialogue interfaces in another study, the results of which will be discussed Chapter 4.

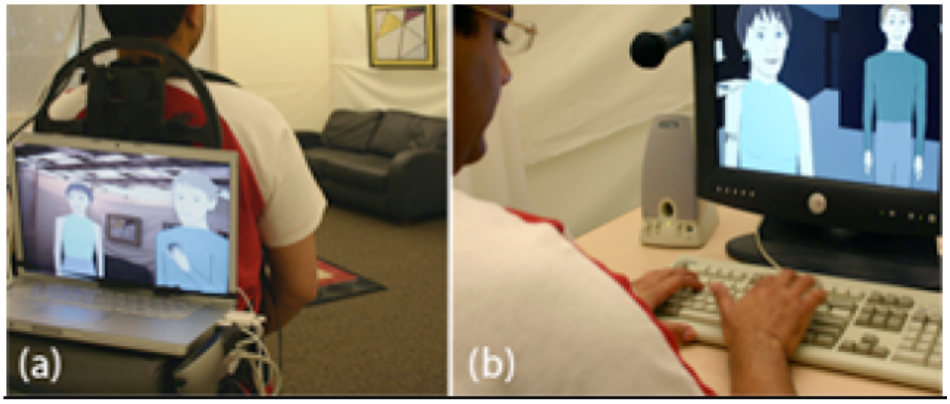


Figure 13. ARFacade. In this study, the authors compared the augmented reality version with a speech-input version.

IMPLEMENTING DIFFERENT DIALOGUE SYSTEMS ON *FAÇADE*

General Implementation Details

As part of our studies on dialogue systems, we implemented various different dialog interfaces for *Façade*. We kept all versions where the user was only allowed to interact at pre-determined points identical so that user experience across those different versions was as similar as possible, barring possible different story traces experienced by the user due to *Façade*'s underlying non-linear story structure. Recognized interactive fiction author Aaron A. Reed, author of *Blue Lacuna*, the longest work of IF to date (Reed 2008) was the lead author for the options in the

menus for the sentence-selection and abstract-response version. Those versions underwent a good number of iterations and extensive testing so that the experience was as smooth as possible. While we list all the different versions we implemented here, it should be noted that not all these versions were tested in the same experiment – the details of the experiments, along with our design goals, will be discussed in the related chapters.

Sentence-selection

The sentence-selection version we built presents an interface similar to the dialogue interfaces found in games such as *Knights of the Old Republic*, the *Monkey Island* series, or *Dragon Age: Origins*, which were discussed above. In this version, the user selects the actual line of dialogue spoken from a pre-scripted list (Figure 14). User interaction is only allowed at fixed points, which was identified by going through the source code for *Façade* and implementing interaction points in places where user interaction is expected by the game. The game waits indefinitely for player input when a sentence selection menu is presented. In our implementation, the number of options in the menus ranged from two to nine. In this version, we circumvented the NLU processing so that each option mapped to the correct discourse act. A response to the player statement was then chosen by the underlying AI system depending on the discourse act and the current context. We made sure that the player is able to take part in every major decision that made a significant impact on plot progression. In addition to those, we also made some of the global mixins available to the player

among the options in the menus whenever we felt it was relevant to the context and when it allowed us to use the authorial leverage this "more-authored" mode of interface provided, evoking in the player a sense of experiencing a hand-crafted narrative with predetermined interaction points. Obviously, due to limitations in the number of options that can be presented to the user at each interaction point, and the fixed-point nature of interaction, this version presents a more linear experience to the user, as expected by this type of interface.

Figure 14. Sentence-selection version of Façade.

Abstract-response

In the abstract-response version, the user is presented with a short abstract representation of the line of dialogue to be spoken by the player character (Figure 15). Whenever the user selects an option, a corresponding line of dialogue is presented to the user. The game again waits indefinitely for the user. While due to

the non-linear structure of *Façade* the story arc experienced by the player might still change, we paid special attention to make this version identical to the sentence selection version except for the options in the menus, so that our comparative study is not affected by the authorial differences among the menu-based versions.

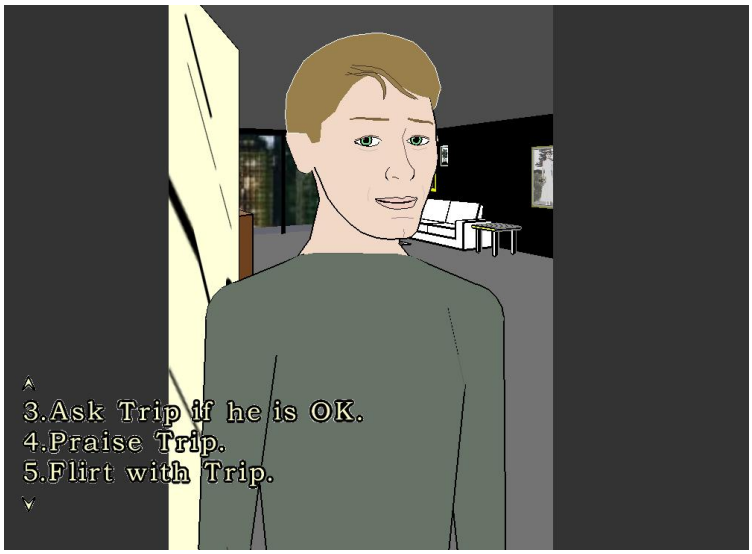


Figure 15. Abstract-response version of Façade.

Reactive-pause

As part of our studies into dialogue systems in games, we also implemented variations on the original NLU system of *Façade*. One of those variations paused the game indefinitely as soon as the player started typing, giving the player infinite time to type in a response. The user could resume the game either by typing an utterance and pressing the enter key, or by clearing the text buffer.

Prompt-pause

In the prompt-pause version, the user was only allowed to interact at fixed points, and the game specifically prompted the user to enter a response (Figure 16). The player can again take an infinite amount of time to enter a response. This version was also identical to the menu-based versions in when it allowed player interaction.

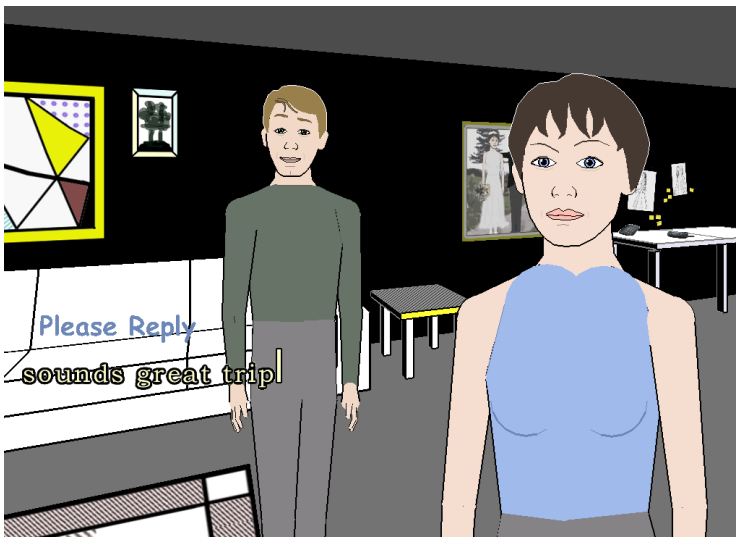


Figure 16. Prompt-pause version of Façade.

STRUCTURES AND SUPPORT FOR METRIC COLLECTION AND AI LOGS

In addition to the qualitative studies we conducted, we supplemented *Façade*'s logging system with additional information that was not previously available in the log files. We added the player's chosen name and gender to the log files, and annotated each line with timestamps. In addition, we added special markers to gameplay logs so that we can parse log files more easily. The statistics and metrics we collected from those log files will be discussed in depth in Chapters 6 and 7,

where we discuss our visualizations of story traces, and our quantitative framework for evaluating dialogue systems in games.

TERMINOLOGY OF STATISTICAL METHODS USED

Cronbach's alpha

One important desired property of survey instruments is that they are internally consistent and reliable. Reliability ensures that if the survey was applied to a similar population sample, it would give similar results, or if the questions were replaced with similar questions, the results would be close to the original survey. These imply that the survey measures a single underlying or latent construct. Cronbach's alpha (Cronbach 1951) is a mathematical method that measures the internal consistency of a set of survey items in a survey instrument. Higher alpha values imply a more reliable instrument. In statistics, as a rule of thumb, a survey instrument is said to be reliable if the alpha value is above 0.7 (Nunnally 1978).

Factor analysis

The main purpose of factor analysis is seeking simpler patterns in relationships between variables in a dataset. In an experiment, variations on observed variables might be due to variations in a fewer number of unobserved variables. Factor analysis aims to simplify data by trying to uncover those latent variables called factors. For example, research by Yee at al. (Yee 2006) into what motivates players to play online games started with a 40-item survey and through factor analysis uncovered three strong factors: achievement, social and immersion. Factor analysis

assumes that the values of observed variables are linear combinations of the values of underlying unobserved variables. The coefficients of those factors are called factor loadings.

Non-parametric statistical significance tests

Statistical significance tests measure whether differences between the values of observed variables from different tests are due to an actual difference or just chance. The results of significance tests reflect the probability that two sets of values of observed variables might have come from the same distribution as a result of chance. When this number is really small, we deduce that it's highly unlikely that the difference in observed variables occurred due to coincidence – in other words, it's highly likely that these two samples come from actually different distributions. The selection of which significance test to use depends on assumptions and knowledge about the data. In this work, we chose to apply non-parametric significance tests since we didn't have reason to assume that our variables were normally distributed. Additionally, non-parametric significance tests tend to be more robust since fewer assumptions about the data are made.

DISCUSSION & CONCLUSION

We believe *Façade* is an ideal platform for experiments such as ours. While most games interleave conversational interaction with other forms of gameplay, *Façade* is an interactive drama in which almost all player activity is conversational. This means that player responses to non-conversational elements of play do not have the

potential to color our results—and that a 15-minute play session includes an amount of conversation that would take much longer to achieve (and evaluate) in a game that also included combat, world exploration, and so on. In addition, precisely because most player actions take place through conversation, the specifics of the dialogue interface are likely to have a noticeable impact on gameplay. At the same time, it is also worth noting that our results may be influenced by the specifics of *Façade*, which remains a very unusual game, though one that may represent a potential future for games in which dialogue is a central element.

4

STUDYING EFFECTS OF DIFFERENT INTERFACES

In our first study, we compared menu-based interfaces with the NLU interface in a controlled study. Prior to our study, the only guidance available to dialogue interface designers has been their own intuition and the assertions, sometimes contradictory, of other game designers and writers. Our study helps situate earlier received wisdom that we reviewed in the related work section. The following sections detail our study goals, methodology, and results.

STUDY GOALS

The first goal of our study was to uncover how different conversation modalities affect the user experience in an otherwise identical game. We wanted to gain insight into what design considerations actually manifest when we test those interfaces in a controlled study. This provided us with important design guidelines and considerations that might otherwise go undiscovered.

Our goals, however, went deeper. While most of the existing discussion on those interfaces deal with the differences on the surface level, such as how the user selects an utterance to speak, what mode of input is used, or how often the user is allowed to interact, we also wanted to know how these different interfaces shape system understanding, and how this understanding affects player experience. Our results provided further evidence and support for more complex models of agency that also take into account system understanding and user perception.

METHODOLOGY

Stimulus material

We ended up selecting three interfaces to employ in our comparative study: The first interface is the sentence-selection interface, which emphasizes authorial control over what the options presented to the user will be, and guides the user along a more controlled path through the experience since interaction is only allowed at fixed, but significant plot points. The abstract-response version forgoes some authorial power in favor of making the outcome of each option much clearer, thus emphasizing player control over the experience. The NLU version, on the other hand, emphasized player freedom with the possible cost of inevitable failures and misunderstandings. We specifically wanted to employ those interfaces in our first study as the sentence-selection and abstract-response versions are most commonly employed in today's mainstream games, as discussed in the related work section, while NLU systems, in our opinion, show great promise for the future of dialogue systems in games. We

also felt this selection represented a good range of systems and enough variation to emphasize the different aspects of gameplay and perceptional qualities we want to study.

Recruitment

For our experiment, we recruited 42 students from an introductory game design class taught at UCSC. Data from 7 sessions had to be thrown out due to problems in the testing session, leaving us with 35 participants. The class was open to all majors. Participants were compensated with extra credit in the class. We only recruited native English speakers with gaming experience so that language ability and familiarity with game interfaces were not influencing factors. Each participant was required to play all three versions in random order to account for learning effects. After each play session, the participants filled out surveys aimed at measuring flow and presence. After the participant played all three versions, we conducted semi-structured interviews designed to explore how the different dialogue interfaces influenced players' sense of presence, control, engagement, agency and enjoyment, which are all frequently discussed concepts in the game design community.

Measurement Instruments

We used the Flow Short Scale by Rheinberg et. al. (Vollmeyer and Rheinberg 2006) to measure flow (Figure 17). We adapted Kim & Biocca's telepresence survey (T. Kim and Biocca 1997) to measure presence (Figure 18). Both these surveys have previously been used in gaming studies and proven to be reliable and useful. Some

sample questions from our interviews can be found in Table 3. The final version of our interview questions can be found in the appendices A and B. We analyzed the results from our interviews by encoding subjects' responses into common categories that emerged from the data as we conducted the interviews, creating new ones if necessary. We discuss our results in those categories below. We also asked participants to rank the different versions (with ties allowed) across the dimensions we considered such as presence, engagement and enjoyment.

F1. I felt just the right amount of challenge.	1	2	3	4	5	6	7
F2. My thoughts/activities didn't run fluidly and smoothly.	1	2	3	4	5	6	7
F3. I didn't notice time passing.	1	2	3	4	5	6	7
F4. I had no difficulty concentrating.	1	2	3	4	5	6	7
F5. My mind was completely clear.	1	2	3	4	5	6	7
F6. I was totally absorbed in what I was doing.	1	2	3	4	5	6	7
F7. The right thoughts/movements occurred of their own accord.	1	2	3	4	5	6	7
F8. I didn't know what I had to do each step of the way.	1	2	3	4	5	6	7
F9. I didn't feel that I had everything under control.	1	2	3	4	5	6	7
F10. I was completely lost in thought.	1	2	3	4	5	6	7

Figure 17. Flow Short Scale.

P1. When the game ended, I felt like I came back to the "real world" after a journey. (1: Strongly disagree – 7: Strongly agree)	1	2	3	4	5	6	7
P2. The game came to me and created a new world for me, and the world suddenly disappeared when the game ended. (1: Strongly disagree – 7: Strongly agree)	1	2	3	4	5	6	7
P3. While playing the game, I felt I was in the world the game created. (1: Never – 7: Always)	1	2	3	4	5	6	7
P4. While playing the game, my body was in the room, but my mind was inside the world created by the game. (1: Never – 7: Always)	1	2	3	4	5	6	7
P5. While playing the game, the game-generated world was more real or present for me compared to the "real world." (1: Never – 7: Always)	1	2	3	4	5	6	7
P6. While playing the game, I NEVER forgot that I was in the middle of an experiment. (1: Never – 7: Always)	1	2	3	4	5	6	7
P7. The game-generated world seemed to me only "something I saw" rather than "somewhere I visited." (1: Never – 7: Always)	1	2	3	4	5	6	7
P8. While playing the game, my mind was in the room, not in the world created by the game. (1: Never – 7: Always)	1	2	3	4	5	6	7

Figure 18. Presence Survey.

Table 3. Sample questions from our interview.

Q1. Would you like to play this game again? Why\why not? Which version did you enjoy the least? Which one did you enjoy the most? Why?
Q5. How engaged were you in different versions of the game? Can you rank them in terms of engagement?
Q8. Which interface variation made you most motivated to move the story forward? Why?
Q13. How much influence did you feel over the story using the different versions?
Q15. How did you form strategies and make decisions? How easy or hard was it to execute your strategy?

RESULTS

Qualitative results

The following sections describe results from our interviews. We coded player responses into categories that are consistent with existing definitions for dimensions such as agency, engagement, sense of control or story involvement. We also asked participants to rank different versions in terms of those dimensions. We allowed ties or non-answers, and only used rankings when we could reliably determine that the subject was consistent in the definition of the dimension being talked about.

Engagement

Despite the dismissal of this interface option by the mainstream game design community, slightly more than half of our participants (54.3%) reported being most engaged in the natural language understanding (NLU) version. 45.7% reported that the menu-based versions were more engaging. Of those, half felt most engaged in

the sentence-selection version, while the other half found the abstract-response version more engaging (Figure 19).

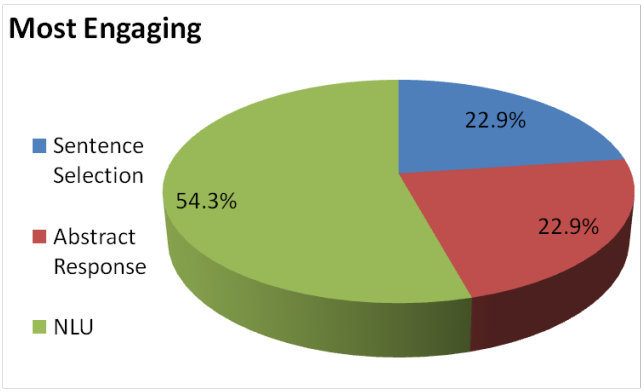


Figure 19. Engagement results. 54.3% of participants found the NLU version the most engaging.

The participants noted that the NLU version was engaging because it allowed them to say whatever they wanted to say, even though they had difficulty making their statements understood by the game. One participant said he found the NLU version to be the most engaging one because:

> *"I was able to actually talk and give my own words and [I didn't have] to deal with random dialogue that someone else had generated. I felt like I was actually controlling the character more instead of just 'here's five thoughts, pick one.' "*

Another reason that participants thought the NLU version felt more engaging was it required considerably more attention than the other versions. Since the players had

to figure out when to interject and how, they had to maintain a constant level of focus and stay alert. One participant noted:

> *"[The NLU version is more engaging] because you are focused on trying to get the right answer more, and you are focused on what different possibilities do I have, and [it requires] more critical thinking. You just have to think more when you have more options."*

While the original *Façade* managed to be engaging by establishing a high sense of attachment between the player character and the player and by requiring constant attention, the menu-based versions achieved engagement by involving the players more in the dramatic situation conveyed by the conversation through more accurate interpretation of what they chose to say. One participant noted:

> *"The menu system had more conversational engagement with me because I could choose from more explicit options in which way the conversation would go [….], instead of just taking a shot in the dark and saying placating things like you would do in an everyday social situation. I felt myself saying more things that represented what I was really feeling about the characters and the situation, so I got more wrapped up in the dramatic aspects of the interaction between Grace and Trip and myself."*

As we will discuss further below, the difference between the two menu-based versions was more defined by the target of control. Some participants enjoyed

having control over what effect their actions would have, whereas others preferred having control over what specifically the player character would say. Knowing the expected outcome of a line of dialogue made the game feel easier, whereas knowing what the player character would say allowed a better association with the character, resulting in a better sense of engagement. When asked which version was the most engaging, a participant noted:

> *"The [sentence selection] version afforded many more options in terms of how you wanted to [play a specific way], because you had lots of possible dialogue to choose from. [In the NLU version] I felt like the limitations of the computer program in turn limited the gameplay aspect, so I couldn't really utilize the freedom of speech as much as I felt would be possible. And the [abstract response version] was the most limiting of them all, because it was just giving you categories of dialogue instead of specific sentences [that allow you] to try and hit a specific emotion."*

Another participant noted the following regarding the other versions:

> *"For the [abstract response version] the difference [from the sentence-selection version] was the text was really unnatural […] and it made it too obvious what was going on in the background. And for the [NLU version] it seemed like everything I said had no effect whatsoever – like they just ignored everything I said pretty much, unless it was 'I agree', 'I disagree.' "*

Participants who felt the abstract-response version was more engaging thought knowing the outcome of their conversation choices made moving through the experience easier. As a result, they felt more engaged in the story. One participant offered this comparison between the abstract-response version and the other versions:

> *"[In the abstract-response version] it was just easier to figure out how you are going to have an effect on the story. In the [NLU] version I felt like they can't understand what I was saying. I was limited to certain things. In the [sentence-selection version] it's not certain what you are going to say means exactly what to them."*

Challenges with different interfaces

According to 71.4% of our participants, the free-form text entry version was the most challenging to learn and use (Figure 20). As one would predict from conventional game design wisdom, the participants frequently complained that the game didn't understand them. They struggled to figure out how to phrase their responses so that the game would correctly interpret and react to them. One participant said the NLU version had "too much freedom" and "it was too difficult to know when you can actually say something and what you are supposed to say." Another felt that the freedom was actually an inhabiting factor: "You can't make snap judgments. It disrupts the flow." Another participant stated:

"The most frustrating was typing in my own responses. I guess I'd type in something that had a keyword in it, so [Trip and Grace] would take that keyword and use it how they were programmed to respond to that keyword but not in the context of my sentence."

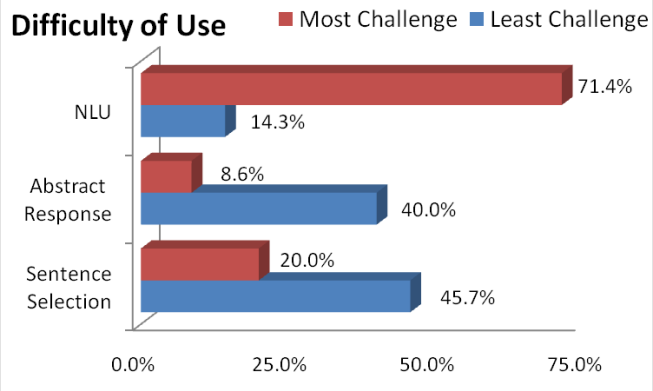


Figure 20. Challenge level. A significant majority of our participants found the NLU interface the most challenging.

As the NLU system in *Façade* maps user utterances onto a limited number of discourse acts, the sense of freedom inevitably broke down in some instances. Figuring out what utterances the NLU system supports was an essential step towards using the NLU version efficiently. Participants who were able to overcome this challenge were more likely to enjoy the system, whereas some participants weren't able to make the transition from the constraints and affordances of real-life conversation to the modeling of conversation in the game. One participant reported feeling as if she was *"behind a glass wall"* the entire time she was using the NLU version, because Trip and Grace didn't understand and didn't respond to her.

Another participant noted that while it was fun to watch Trip and Grace interact with each other, it felt like they didn't want to interact with him.

For some participants, unconstrained typing also suggested informal communication strategies that were inappropriate. One participant noted that he tried to use slang and everyday college language, while another tried abbreviations such as "u" instead of "you." Both styles of discourse were not forms understood by the NLU system.

The NLU version also does little to suggest when and what player actions are appropriate. Most of the participants stated that the game should more clearly indicate to them when their input is expected—the suggestions ranged from implementing subtle prompts that signal to the players when their input is expected to the game pausing entirely to wait for input. Some participants expressed a need for a tutorial, or better feedback on how their input was processed and understood by the system. Although one of Mateas and Stern's goals when developing the NLU system for *Façade* was to make sure the system never says, "I don't understand"(Michael Mateas 2004), a few participants actually wanted the system to somehow inform them that it didn't understand, so that they wouldn't miss the opportunity to interact.

Despite the fact that it most approached the natural flow of conversation, pace and timing were also found to be significant drawbacks with the NLU version. The participants felt that the time required for deciding on a response, then formulating

that response so that the game would understand it, and finally actually typing in that response, was too long. By the time the player was done responding to Trip or Grace, the characters would have already moved on to another topic, and the player input would no longer make sense in the original context: the opportunity to interact would be lost.

> *"The prompts definitely helped, not so much with what I needed to say, […] but just the fact that it let me know when I could respond to what I was supposed to. […] In the [NLU] version, you are trying to type in a response, and you are trying to think of something to say, and they have already moved beyond the question. It's over before you can get anything out."*

Participants reported developing several strategies to cope with these problems. One participant tried to anticipate events and have a response typed in so that he could press enter to submit it at the correct moment. Another reported only typing in very short and simple phrases and sentences that he was sure the game would understand. While these are valid strategies, they also defeat possible goals of having a realistic conversation system, as player utterances are reduced to simple keywords that cannot possibly capture the nuances of real-life dialogue, and pacing becomes an inhibiting, rather than enabling, factor for presence.

A relatively minor number of participants reported issues using the menu-based versions. The most significant challenge that our participants experienced with the

sentence selection version was that the options didn't give the player a clear sense of what the outcome of speaking that line would be. Despite the fact that the time consuming nature of reading full responses is seen as the major drawback of this interface in the game design community, only two out of 35 participants complained about having to scroll through and read all the options.

We also asked our participants what improvements should be made to the interface. A more accurate NLU system and a prompt that informs the player when his or her input is expected were the most popular answers for the typing version. Participants wanted to see more options on the screen in the menu-based versions, reducing the need for scrolling, which is in fact one of the additional advantages of the interface introduced by *Mass Effect*, which arranges menu options radially.

Sense of Control

The three interfaces in this experiment offer different paths to giving the player a sense of control over the game world. The abstract-response version gives the player more control over the ultimate outcome whereas the NLU version gives the player more direct control over the avatar. This may explain some seemingly-contradictory statements about agency and control in the related literature. The sentence-selection version, on the other hand, may be seen as aiming for a more balanced approach by allowing some degree of both authorial control and player freedom, allowing the player to choose among authored responses.

Our participants reported having the strongest sense of control in the abstract-response version (Figure 21). Knowing the outcome of speaking a certain line of dialogue made players feel more influential in the game world.

While participants still enjoyed having total control over what the player character said, ultimately the difficulty of the interface coupled with the interpretation problems inherent to NLU resulted in a loss of control. In fact, 65.7% of our participants reported that they felt least in control using the NLU version.

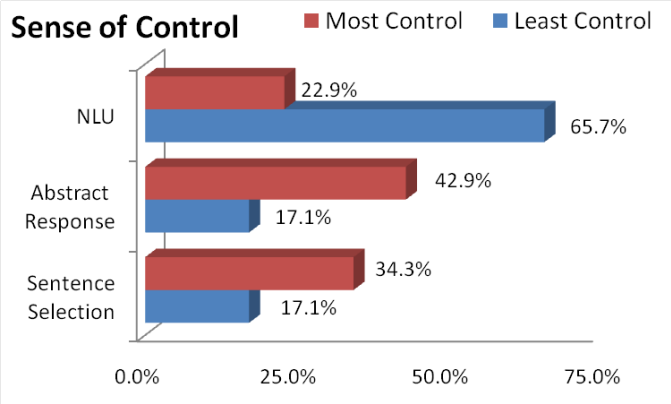


Figure 21. Sense of control. Participants felt most influential using the abstract-response version, and least influential using the NLU version.

Looking at our results, as predicted by our earlier work on agency, it seems players experience a greater sense of control when interfaces make the outcomes of their actions clear, rather than offering an illusion of greater control that isn't entirely supported by the underlying system. Some participants even reported feeling overwhelmed by the freedom offered by the NLU version. They felt that so many things seemed possible that they had no idea how to choose a particular thing to say.

Realism & Presence

During our interviews, we also tried to gain further insight on how a conversation system can feel realistic and natural (given this is a stated goal for many game designers) and how well our three different systems can support and maintain a sense of realism and naturalness. Participants who found the NLU version unnatural mostly complained of interpretation errors and felt that despite *Façade*'s attempts to the contrary, the limitations of the program were still very visible. One participant noted that "interacting with [Trip and Grace] still felt like interacting with software." Another said "in [the NLU version] you can still tell the program is trying to hit keywords in a database."

Most participants thought that the abstract-response version was an unrealistic model of conversation, because it strayed too far away from the realities of day-to-day conversation. Players are accustomed to conversation at the level of words and sentences, not discourse acts. Another factor that decreased presence in the abstract-response version was, as one would predict from prior game design wisdom, the mismatch between player's intent when selecting a short, abstract response and the actual line of dialogue spoken by the player character as a result. One of our participants said:

> *"[In the sentence-selection version] you can get more into the character's shoes. [In the abstract-response version] if you agree with [either Trip or Grace] your character might end up saying something […] rude to the*

other person and that might not have been what you intended. [For example] if you disagree, [the spoken line] might make it sound more rude than you'd actually say."

The main facilitator of presence was control over the player character's statements. Participants noted having a stronger sense of control over what the player character will say in the NLU version, which potentially allowed them to be themselves in the world of *Façade*.

This increased sense of presence came with some trade-offs. Some participants felt that the increased sense of presence resulted in too much responsibility. They felt that the fate of Trip and Grace's marriage was entirely in their hands and as a result the experience stopped being enjoyable.

Interestingly, we also observed that our participants felt they were more bound by social norms and conventions when using the NLU version. A participant noted that while she was trying to play a more difficult character in the abstract-response version and the sentence selection version, she definitely tried to be nicer in the NLU version. She stated:

"… [the NLU version] was more like a social situation than a multiple choice test. I was less inclined to say those things [I said in the other versions] that I wouldn't normally say."

Although on the surface free-form text entry seems like the most natural model of conversation, games present a system that players expect to be able to understand and influence more directly than real-life interactions. Some participants felt that with the NLU version, formulating and executing a plan was almost impossible, which resulted in a loss of control over the experience.

Story involvement

Our participants reported that they felt significantly more motivated to move the story forward using the sentence-selection version (Figure 22). One participant stated that "[the menu] was already there for me, it was easier for me to see what I wanted to do." Another participant said:

> *"[in the sentence selection version] you had a lot more variety, a lot more range [compared to abstract response version], you had a lot more leeway, you can somewhat agree or somewhat disagree, whereas in the [abstract response] version you had to either go with this person or that person … In the [NLU] one I was just so out of control that I just felt stranded a lot."*

Interestingly, this perception of more range and leeway was just an illusion: the sentence-selection and abstract-response versions were identical except for what was displayed in the selection menus. Still, this participant felt that with the sentence-selection version, she could relate more to the options and not feel trapped into taking a certain path in the game—having the actual lines of dialogue entirely

written out gave the participant an illusion of more range and variety even though they were mapped to the same discourse acts as in the abstract-response version.

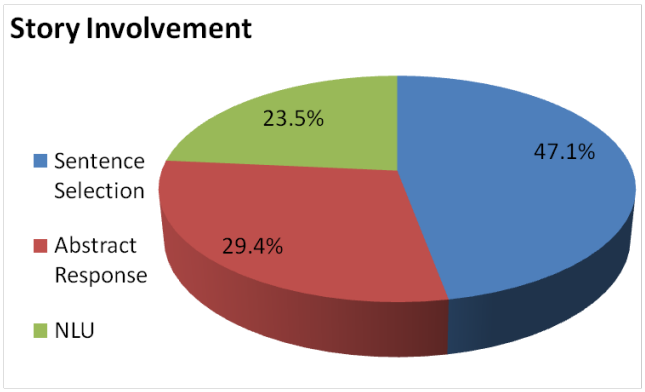


Figure 22. Story involvement. Participants were more involved in and more motivated to move the story forward in the sentence-selection version.

The immediate reaction from one of our participants when he saw the option to flirt with Trip in the very first menu was *"Now I'm tempted to try this!"* Some participants wanted to go back to the game after finishing and try other options to see how the characters would react. Many participants stated that with the sentence-selection version they felt more like a character within the plot, although that character wasn't them. None complained of being forced to play a particular character, though this is one of the concerns about this interface option expressed in the game design community. One of the participants noted that having pre-scripted options made even Grace and Trip feel more fleshed out and well-developed, and as a result made him more involved in the game world with a higher sense of purpose. Participants felt the menu-based versions better placed their character in context with the game's dramatic events and the network of social relations between Trip, Grace and the

player character. While free-form text entry, coupled with the game's relative lack of back story for the player character, provided a blank canvas for the player to reflect his/her personality on, some participants didn't feel that they had enough compelling reasons to care for Grace and Trip and their marriage. As a result, their actions felt meaningless.

Enjoyment

Despite all the significant drawbacks that they mentioned, and flying in the face of conventional wisdom, more than half of our participants still reported enjoying the NLU version the most (Figure 23). The participants particularly enjoyed being able to say whatever they wanted and interrupt the characters at any time, in contrast to the discrete and limited interaction possibilities offered by the menu-based versions. When asked which version she liked the most, one participant noted:

> *"The [NLU version] for sure. You had a lot more freedom in what you could say. If you want to put in your opinion about something while they are talking, it felt like you could do it then rather than just waiting for this [menu] to pop up with limited choices of what you want to say. Maybe it's not something you want to say but you don't have a choice. I guess it's the freedom and the real-time thing too. [The NLU version] is not like wait-go-wait-go."*

Participants also noted that while the NLU version had its problems, it was a very fulfilling experience when it actually worked. As noted by a participant:

"[I enjoyed the NLU version the most] when they threw me out. I told Trip to shut up ... I didn't really expect [to be thrown out]. Seeing the reaction kind of quickly made [the NLU version] engaging."

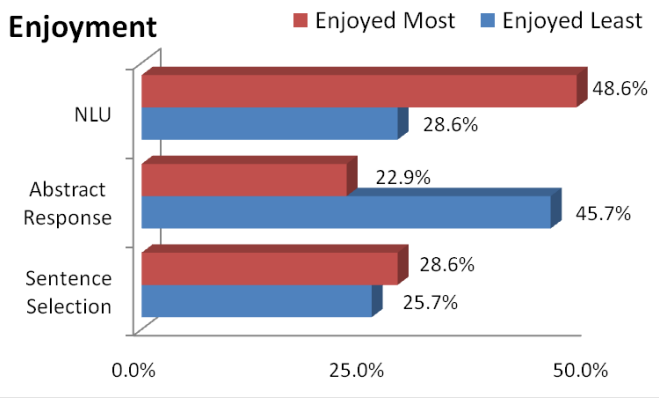


Figure 23. Participants noted enjoying the NLU version the most, and the abstract-response version the least.

A participant even noted that the NLU version gave him the most sense of control, because it enabled him to take a back seat and let Trip and Grace work out their problems on their own. When playing the NLU version, he just answered enough questions so he didn't get thrown out. He stated:

"[In the NLU version] I thought maybe if I just listened they would figure things out themselves. It was clear from the first two [playthroughs] that what I said wasn't helping them.... [In the typing version] I had more control, even though I didn't control anything because I didn't influence them at all I felt like I was more in control and more able to help them ... I influenced them by not doing anything and just listening, and letting them work things out themselves."

Our subjects noted that although they realized the limitations of the system, the NLU version still gave them an illusion of freedom that is absent from other versions; as one participant noted, when using the NLU version it "felt like everything was possible." Another said that his "imagination just died in the [menu-based] ones."

This illusion of freedom, however, did not translate to enjoyment for all participants. Some felt it was too difficult to figure out what to say since too many things seemed possible at any given moment. As a result, players' sense of agency suffered. Participants who enjoyed the sentence-selection and abstract-response versions usually noted that the game allowed them to think, make decisions, and execute their plans, in contrast to the NLU version which suffered from misinterpretation and timing problems, along with too much unguided freedom which made it difficult to figure out what actions to take to move the story forward.

Quantitative results

A survey instrument's internal consistency and reliability is typically measured with the reliability coefficient Cronbach's alpha (Cronbach 1951). This coefficient measures how reliably a survey instrument measures a single construct by looking at whether the items in the survey produce similar scores. Cronbach's alpha values for each of the surveys we used are listed in Table 4. These values show that our measurement instruments are acceptably reliable.

Table 4. Reliability scores for our flow and presence surveys.

Version	Survey	Cronbach's alpha
NLU	Flow	0.884
	Presence	0.900
Sentence-selection	Flow	0.812
	Presence	0.786
Abstract-response	Flow	0.842
	Presence	0.873

Flow

To analyze our results, we used the Friedman test. The Friedman test is a statistical significance test used for computing the significance of observed differences between repeated measures on non-parametric data (Milton Friedman 1937). Using the Friedman test, our results show that the survey item F3, "*I didn't notice time passing.*" is significantly different across the three conditions for the Flow measures (C^2 = 12.689, df=2, p=0.002). To determine from which pair or pairs of conditions the significance arises from, one then has to run pairwise significance tests. The Wilcoxon matched pairs test is a non-parametric test that can be used for that purpose (Wilcoxon 1945). For our data, the Wilcoxon's posthoc tests show that the significance was for the NLU and abstract-response versions (p=0.002) and abstract-response and sentence-selection versions (p=0.042), but not between sentence-selection and NLU versions (Table 5).

Table 5. Results of Friedman's test on Flow Survey Item 3: *I didn't notice time passing*.

Test Statistics[a]

N	35
C^2	12.689
df	2
p	.002

a. Friedman Test

Version	Mean Rank
Sentence-selection	2.07
Abstract-response	1.63
NLU	2.30

Test Statistics for post-hoc pairwise Wilcoxon signed ranks test

	Orig. vs sentence-sel.	Abstract-resp. vs sentence-sel.	Orig. vs abstract-resp.
Z	-1.480[a]	-2.029[a]	-3.087[b]
Asymp. Sig. (2-tailed)	.139	.042	.002

a. Based on positive ranks.

b. Based on negative ranks.

As our results indicate, using the sentence-selection and the NLU version participants reported losing their sense of time significantly more than the abstract-response version. Based on the qualitative data, most participants reported feeling that the abstract-response version was almost too easy – one of the participants stated that he felt like he was commanding a robot. The relative lack of interesting writing in the abstract-response version also resulted in a more bland experience, whereas through the help of interesting and tempting options in the sentence

selection version, they felt more involved in the story and the game world, and wanted to go back to the game and explore the outcome of those options. As mentioned above, a participant noted feeling that the sentence-selection version offered more nuance in its options in the menus, although by our design both menu-based versions were identical except for the text displayed in the menus.

One of the most common issues raised with the NLU version was that it required more cognitive engagement from the players, since they not only had to figure out what to say from a seemingly infinite number of options, but also when to say it. The NLU version also allowed participants to talk to the characters anytime they wanted, which might have resulted in a better sense of flow with regards to time spent in game. As expected, that version also felt more "real" – this was an advantage for some players, but for others it made the game feel more tense and they felt more bound by social norms – a finding that nonetheless suggests that NLU interfaces provide a higher degree of presence and engagement.

We also conducted factor analysis on the flow survey results to see if the variations we observed can be explained by latent factors that we haven't considered and measured. Since we had only 42 participants, following Gorchush' advice that at least 5 participants per variable is required (Gorsuch 1983), we only considered the first 8 items of the flow survey. For the sentence selection version, we found evidence suggesting that for this version flow mainly splits into two factors (Table 6).

Table 6. Results from factor analysis on the sentence-selection results.

	Factor 1 Absorption	Factor 2 Clarity
F6. - I was totally absorbed in what I was doing.	.913	
F3 – I didn't notice time passing.	.896	
F2 – My thoughts/activities ran fluidly and smoothly.	.694	
F4 – I had no difficulty concentrating.	.683	
F7K – The right thoughts/movements occurred of their own accord.		.780
F8K – I knew what I had to do each step of the way.		.775
F5K – My mind was completely clear.		.713
F1K – I just felt the right amount of challenge.		.605

We call these factors absorption (items 6, 3, 2 and 4) and clarity (items 1, 5, 7 and 8). This suggests that players' scores were mostly grouped and influenced across those two dimensions, which are likely to be important design parameters for dialogue systems. The first factor is mostly related to being absorbed in the game, suggesting a deep embodiment of the player character played a role in players' preferences about the sentence-selection version. The second factor, clarity, is mostly related to a clear understanding of how to operate the computational system through the interface. These two factors explain a significant portion of players' preferences about the sentence-selection version.

Presence

According to our results, using Friedman's test, survey items P1, P2, P5 are significantly different across the three conditions for the presence measures. The Wilcoxon's posthoc tests show none of the P1 pairs are actually significant. For the second item, there was a significant difference between the NLU version and both the sentence-selection and the abstract-response versions in favor of NLU (Table 7). For the fifth item, we found significant evidence only for a difference between the NLU version and the sentence-selection version (Table 8).

Both of these items were related to the arrival dimension of the presence survey. The results show that the NLU version offered the participants a higher sense of presence than the menu-based versions. This fact is also corroborated by the qualitative data. In particular, participants enjoyed the freedom offered by the NLU version despite the significant drawbacks such as accuracy problems or the difficulty of figuring out when to interrupt Trip and Grace to enter input. In addition, we also observed that the greater degree of freedom offered by the NLU version also allowed more leeway in interpreting the characters' reactions to player statements.

Table 7. Statistics for the presence survey item *P2: The game came to me and created a new world for me, and the world suddenly disappeared when the game ended. (1: Strongly disagree – 7:Strongly agree)*

Test Statistics[a]

N	35
C^2	8.970
df	2
Asymp. Sig.	.011

a. Friedman Test

Ranks

Version	Mean Rank
Original	2.33
Sentence-selection	1.94
Abstract-response	1.73

Test Statistics for Wilcoxon pairwise post-hoc tests

	Orig. vs sentence-sel.	Abstract-resp. vs sentence-sel.	Orig. vs abstract-resp.
Z	-2.065[a]	-1.596[a]	-2.789[a]
Asymp. Sig. (2-tailed)	.039	.111	.005

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Table 8. Statistics for the presence survey item *P5: While playing the game, the game-generated world was more real or present for me compared to the "real world." (1: Never – 7:Always)*

Test Statistics[a]

N	35
C^2	7.252
Df	2
Asymp. Sig.	.027

a. Friedman Test

Ranks

Version	Mean Rank
Original	2.31
Sentence-selection	1.77
Abstract-response	1.91

Test Statistics for Wilcoxon pairwise post-hoc tests

	Orig. vs sentence-sel.	Abstract-resp. vs sentence-sel.	Orig. vs abstract-resp.
Z	-2.822[a]	-1.169[b]	-1.731[a]
Asymp. Sig. (2-tailed)	.005	.242	.083

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

In addition to the above items, we also questioned participants on enjoyment, their empathy capabilities, their personality types, and their computer usage and gaming habits. We found no evidence suggesting that enjoyment differed between those three different versions. We also didn't find any correlation between empathy or personality traits or computer use and gaming habits in our results, except hours

spent gaming was significantly correlated to experience using computers - indicating that gaming still remains mostly exclusive to computer literate demographic.

Discussion & Conclusion

Our direct comparison of different game dialogue interfaces has some surprising findings. In particular, players enjoy most the interface approach—natural language understanding—that also made them feel least in control and often produced frustrating errors.

However, this does not necessarily indicate that game designers should choose NLU dialogue interfaces. We find sentence-selection interfaces, which appear to be in rapid decline within the mainstream game industry, most effective for producing a sense of story involvement—and common critiques (that they take too long to read and put words in the player's mouth) are not supported by our findings. Abstract response menu interfaces, in contrast, which have been praised for preserving natural conversation flow, are found most unnatural in our results, though most effective at producing a sense of control relative to the game system. Perhaps most fundamentally, our study demonstrates that game dialogue interfaces have a profound impact on the experience of gameplay, even when all other aspects of the game are held steady, something of which all designers should be aware.

In this study, we compared interfaces that differ in both pacing and mode of interaction. In a following study, which I'll discuss in the next chapter, we kept the

NLU interface constant, but introduced different pacing options to the interface to study the effects of these artificial modifications to player experience. In the conclusion chapter, we will return to results from both these studies for a more extensive discussion, summarizing what we have learned on dialogue systems.

5

STUDYING EFFECTS OF PACING

In our first study, discussed in the previous chapter, we compared menu-based interfaces with the NLU interface. In this study we explore the NLU interface option in more depth, addressing issues of pacing that came up in our prior work by introducing artificial pauses into the NLU system of *Façade*. Our results provide important guidance on how to design for creating particular types of experiences, and provide further insights into game design.

DESIGN GOALS

In our previous study, we found that the NLU version maximized engagement and presence, despite participants reporting many issues with pacing. While it'd be tempting to conclude that it was the realism offered by the NLU interface that maximized those aspects, we wanted to actually test this assumption in a controlled study. In the game design community, realism is a frequently pursued goal, assuming it maximizes engagement and presence. The *Holodeck* has long been

assumed to be the ultimate interactive experience. Designers and scholars have sometimes found themselves at different ends of a discussion that debates the role of realism in game design, with Salen and Zimmerman criticizing –and in fact, labeling as *immersive fallacy* - this point of view as being too focused on presentational aspects and attributes of games and assuming immersion to be an intrinsic property of games (Salen and Zimmerman 2004) . According to Salen and Zimmerman, immersion occurs through the process of play itself as well, and should take into account how the game functions in relation to the player. In this experiment, we were able to gain insight into these issues within the context of interactive drama. Our results hint that the Holodeck is not necessarily the end-goal, and designers can introduce explicit mediations to shape player experience. The following sections detail our methodology and explain our results in more detail.

METHODOLOGY

Stimulus material

The three interfaces we compare in this study are as follows: First version we consider is free-form text entry (as in the original *Facade*) in which the user can type anything at any time. In this version, *Façade's* NPCs — Trip and Grace — will move on at their own conversational pace. A response is only registered when the player hits return at the end of typing an utterance. The pacing of this version most closely resembles real-life conversation – when the player starts speaking, Trip and Grace

will allow the player some time to enter a response, but will move on leaving behind even direct questions to the player character if the player takes too long to respond.

Second, in what we call the *reactive-pause* version, Trip and Grace will wait indefinitely when the player starts typing an utterance, as though the player character has signaled that they wish to take a conversational turn at the player's first keypress (and as though Grace and Trip are infinitely deferential and patient). The game resumes when the player is done typing or clears the text buffer.

Third, we consider the *prompt-pause* version, which displays a prompt that explicitly asks the player to type an utterance when conversational interaction is expected by the game, e.g. when Trip and Grace asks the player's opinion on an issue or asks the player a question. The game again waits indefinitely for player input.

Recruitment

We conducted within-subject controlled experiments. We randomized the play order to account for learning effects. Participants were recruited from the university undergraduate population and from local members of their online communities (mean age = 20.0 years (standard deviation, 2.34)). There were 23 participants in total; data from 2 sessions had to be excluded from the analysis due to problems during the session. We only recruited native English speakers with gaming experience so that language ability, possible common character traits of gamers and familiarity with game interfaces were not influencing factors. Participants were instructed on study procedures before they began. They were also made aware that

they could opt out of the study at any time before or during. Refreshments were provided for each participant. Sessions lasted between 90 and 120 minutes.

The study began with participants filling out a general demographic survey. After the participants played all three versions, a semi-structured overall experience interview was conducted. The interview consisted of a mixture of ranking questions, and more open-ended questions designed to elicit the underlying reasons for the ranking choices made by the player. This survey was based on our survey from the study discussed in the previous chapter, with added questions that emphasized the effect of pacing options on player perception. The semi-structured nature of our interview allowed us to pursue participants' statements about the various dimensions we considered so that correct mappings could be achieved even if participants had different definitions or interpretations of the concepts we interviewed them about.

RESULTS

Engagement

According to our results, slightly over half of our participants (52.4%) found the reactive-pause version the most engaging. The original version was the most engaging for 28.6% of our participants whereas 19.1% found the prompt-pause version the most engaging (Figure 24).

The original version requires constant interaction and presents a more realistic model of conversation. However, in our interviews we found that the appeal of this "naturalism" wore off quickly when players felt they lost control of the conversation. Some participants reported becoming less engaged with the conversation when they were too busy trying to keep up. They often indicated that the original version moved on too fast for them to respond. In fact, over half of our participants indicated that their inability to keep up with the conversation negatively affected their experience. The original *Façade* interface featured limited pausing during player interaction because the designers believed this would maintain dialog momentum and thus contribute to a heightened sense of dramatic intensity. Clearly these engagement results are contrary to this original design intention.

Participants who indicated the reactive-pause version was the most engaging noted that having enough time to formulate and enter their responses was a positive contributing factor to their engagement with the game. Participants also noted that being able to speak without talking over the characters improved their engagement with the game. One participant replied that he was "able to keep up with the world [instead of] talking over people." In the original version, many participants reported not wanting to be rude to Trip and Grace by interrupting them, and therefore reported holding off on responding until they felt there was a break in the conversation or an explicit invitation or question from the characters. On the other hand, in the versions that paused for player input, participants reported feeling that

when they spoke they were being listened to. Engaging during the back and forth accusations and arguments between the characters didn't feel as challenging socially as in the original version, and for some participants the slower pace of the pausing versions made the situation feel less tense and allowed more involvement.

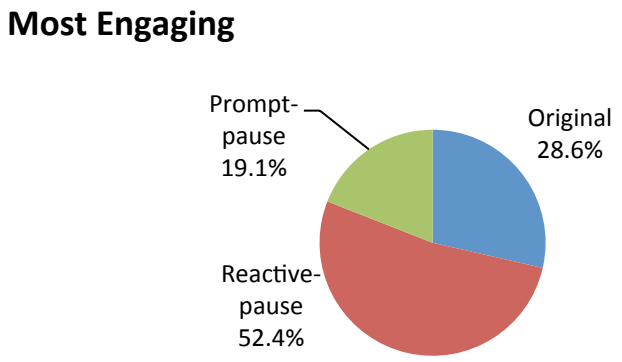


Figure 24. Engagement. Our participants reported feeling most engaged in the reactive-pause version.

In contrast, participants who felt the prompt-pause version was more engaging reported that knowing when their interaction was required made the experience more like a game. As one participant noted, "[the] prompt reminded me when to interact, it was less of a movie." Another noted, "I found the prompt one most engaging cause it gave me the amount of time that I needed to type what I wanted to say and it also told me when I could comfortably say things without cutting off what the characters were saying."

Sense of control

The three different versions we employed in our experiment give varying degrees of control over the pace of conversation. The original version allows players to completely control when to speak, but gives players limited time to interact with the characters — Trip and Grace will move on if the player takes too long to complete her utterance. The reactive-pause version also gives the players complete control of when to interact through conversation, but it pauses the game indefinitely when the player is typing an utterance. The prompt-pause version limits interaction to predefined prompt points, with the game pausing indefinitely at these points for the player to enter input.

According to our results, 52.4% of participants reported having the strongest sense of control using the prompt-pause version (Figure 25). Participants reported that having a prompt not only helped them figure out when their interaction was required, but also made them feel that their actions had a bigger influence on the game. One participant noted, "The prompt definitely helped me keep up with the characters because [in the other versions] sometimes you're [wondering] 'Should I say something here?'" Although the prompt-pause version took away some of the control from the players over when they can speak, the more structured interaction it offers resulted in a higher sense of control.

However, one of the main complaints about the prompt-pause version was that sometimes there was a mismatch between when the player wanted (or didn't want)

to interact with Trip and Grace and the timing of the prompts. Sometimes a prompt would show up when the player didn't want to speak or thought interrupting the characters was unnecessary, and other times the game wouldn't prompt them to speak when they wanted to.

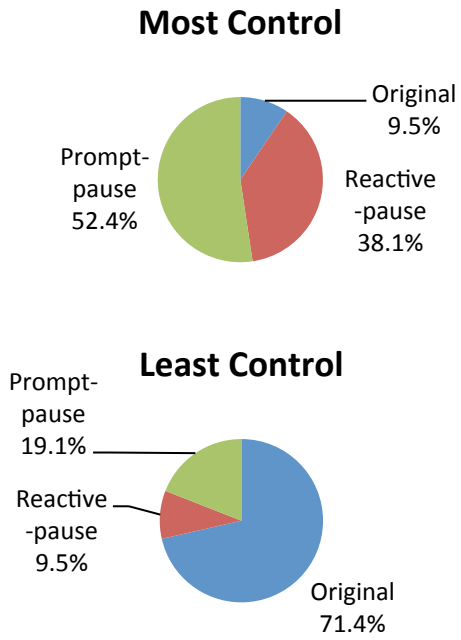


Figure 25. Sense of control. Participants reported feeling most influential using the prompt-pause version, and they felt the least sense of control using the original version.

The reactive-pause version also provided a relatively strong sense of control with 38.1% reporting the strongest sense of control and only 9.5% reporting the least sense of control. Participants most often noted that the ability to take as much time as they wanted allowed them to formulate their responses properly. Having enough time to

think of how they wanted to act, formulate a response accordingly, and edit it before entering their utterance increased their sense of control over the game.

The original version gave the players the least sense of control. Only 9.5% of participants responded they had the strongest sense of control using the original version while 71.4% said it gave them the least control. The most prominent complaint was that they did not have enough time to formulate proper responses to the conversation.

Difficulty of use

A strong majority (71.4%) of our participants reported that the original version was the most difficult to use. The prompt-pause version was reported as the easiest to use by 52.4% of our participants (Figure 26).

Not having enough time to enter a response was the major drawback of the original version. Some participants said that when they finished their response, the characters had already moved on to another conversation topic before they could even press enter. When asked whether or not the pausing was helpful one participant responded:

> *"Yeah, because I can't type as fast as I can talk. So it in essence kind of made it more realistic because […] you could respond without them cutting you off."*

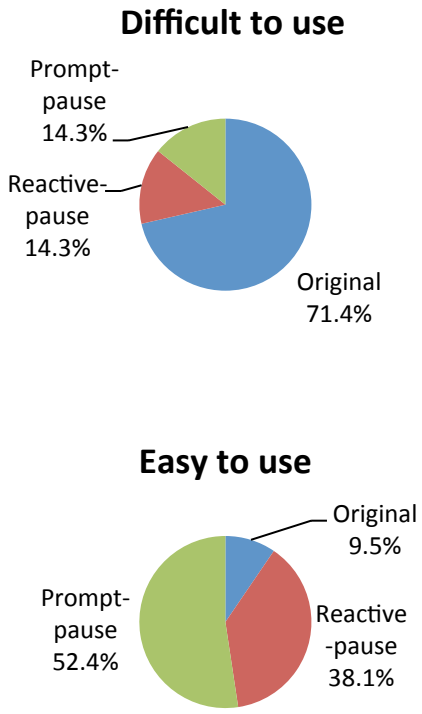


Figure 26. Difficulty of Use. Participants found the original version the most challenging, and the prompt-pause version the easiest to use.

The lack of time in the original version also made it hard to correct typos or edit responses when they changed their minds. Often participants would miss critical interaction points while they were editing their response, or they would simply give up altogether by clearing the text buffer. One of the participants reported developing the following strategies to cope with the challenges presented by the original version:

"I would try and anticipate when they were about to stop saying something, I would start typing right before that and I would definitely

try and type faster. I wouldn't worry about… sometimes I would get more typos I noticed like two or three typos where as I wouldn't get those in the pauses. And that would throw me off a little bit. It was manageable but it felt a little bit harder to keep my head above water."

The pause gave players an infinite amount of time to edit their responses, which also made dealing with input length limit easier. As a participant noted:

"I wasn't talking over conversations or ah… running into difficulty with character limitation as to how long my sentences could be. It's an improvement because when I'm editing myself in the [original] version, I'm running out of time or I'm sitting there waiting for things and the characters will either stare at me or keep talking to one another. And the conversation moves beyond what I was going to comment on while I'm typing."

The prompt-pause version was reported as the interface easiest to use, the main reason being that it explicitly alerted users to opportunities of interaction where their input could have significant meaningful impact. One participant remarked that the prompt "gave a necessary structure and polish to conversation." Another participant noted:

"The prompt improved it because it made clearer when they were at a junction, when they wanted feedback so they know where to go. You don't [have to] think about it while playing the game; just think 'oh I should

talk now.' Because [in the original version] sometimes you just forget to respond and just sat back and watched them talk."

However, a small number of participants still preferred the more naturalistic model of conversation offered by the original version. As one participant noted:

"the prompt version I enjoyed the least. The prompt made me feel very aware of every time it paused game play, and I always felt obligated to answer it. The free form nature of inserting comments whenever one wants (as opposed to answering when prompted) makes Façade a lot more fun."

Story Involvement

During the interviews we also asked participants which version made them feel most motivated to move the story forward. Among all the different versions, the prompt-pause version offered the highest degree of story involvement as reported by 52.4% of the participants (Figure 27).

As expected, participants noted that the prompt indicated a good time to respond to the characters and they used that as a cue to act in the game. One participant told us, *"the prompt allowed me to think of a more proper response to their discussions and I feel as if I got more engaged due to it."* Other participants mentioned in the interviews that they tried to use the prompt cues to explore other topics or shift the conversation, thus turning conversation into a gameplay activity with more explicit story exploration power.

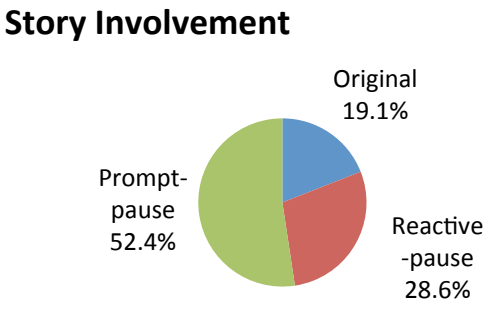


Figure 27. Story Involvement. Participants reported feeling more involved in and more motivated to move the story forward in the prompt-pause version.

Presence & immersion

We also asked our participants to compare the prompt-pause and reactive-pause versions to the original version separately, to get a sense of how these modifications made a difference to the gameplay experience in terms of immersion. The results from our interviews show that 71.4% of participants felt that the reactive-pause version was a more immersive experience than the original version. Some participants, however, commented that the almost too explicit nature of interactions in the prompt-pause version made the experience too easy and took away from the immersion. Their interactions didn't feel as situated within the game world as in the other versions. As one participant noted *"[the] prompt made it less of an interaction"*. Another commented, *"It also broke the sense of presence. Having the prompt there would remind me that I was playing a game instead of trying to resolve a marriage conflict."*

Participants who felt the prompt-pause version was the more immersive of the three versions reported that knowing when their input would have a significant impact resulted in a higher ability to have more meaningful interactions with the game. This

in turn made the characters and the game world more approachable and understandable. An overwhelming 66.7% of participants favored the prompt-pause version over the original version, which is interesting given that the original version offers the most natural model of conversation.

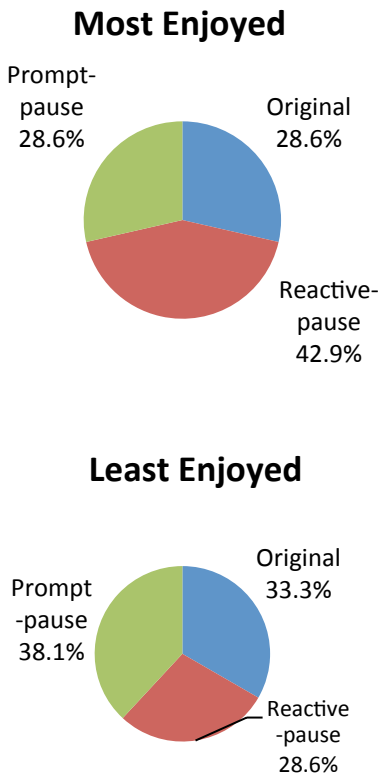


Figure 28. Enjoyment. Participants reported enjoying the reactive-pause version the most.

Enjoyment

According to our results, the participants enjoyed the reactive-pause version the most, with 42.9% of the participants picking this version as their favorite (Figure 28). This is despite the fact that the prompt-pause version (not the reactive-pause

version) was seen to offer the most control, be easiest to use, and engender the greatest story involvement. However, reactive-pause had the greatest engagement, which may point to this as the most important factor for enjoyment. Additionally, when comparing with the original, players often mentioned the difficulty of keeping up with the game when there was no time to enter an utterance.

The prompt-pause version was reported to be the least enjoyable; the combination of the prompt and the game pausing was often described as too distracting. However, as we noted above, participants who enjoyed this version reported that they liked having a prompt guiding them on when to interact with the characters, whereas in the other versions they felt lost in the conversation. Having the game help guide them through made it easier to play and actively participate in the conversation.

People who reported enjoying the original version the most liked how that interface most closely imitated real-life conversation. They felt the pauses created unnatural breaks in the experience and disrupted the flow of the game. In contrast to most participants, they preferred a natural model of conversation to one where they were constantly reminded that they were playing a game where the characters explicitly waited or prompted them for their interactions.

Discussion & Conclusion

The results from our experiment point to important design guidelines for dialogue systems. First and foremost, we have found that the different mediations introduced

to the original NLU system of *Façade* provide different design choices and trade-offs. The original version offers the highest degree of naturalism and freedom, but might result in an experience that feels too rushed and tense – an experience that might still be preferable for some designers and players. The prompt-pause version offers a clear guide to the player of when their intervention is required — maximizing sense of control, ease of use, and story involvement — but also disrupting the experience in a manner less engaging and enjoyable than the reactive-pause version. The reactive-pause version seems to offer the best compromise in terms of control between these two versions, as it allows players more time to formulate and execute a conversational strategy, while still offering control over when to speak. However, it is a less realistic model of conversation, which might result in a less immersive experience for some players. Our findings suggest that combining the pausing and prompted conversation aspects of menu-based interfaces with the high degree of freedom offered by NLU systems on what to say might result in promising future directions for dialogue system design in games.

Our results also point to deeper insights into game design and player behavior. Interestingly, the most "realistic" or "natural" imitation of real-life conversation was less appealing to most players than versions that behave in "unnatural" ways on dimensions ranging from engagement to the perceived sense of control and story involvement. As mentioned above, in our previous study, when we compared the original version with different menu-based dialogue interfaces, we found that

despite all the challenges participants reported using it, the original version was found to maximize presence and engagement. One might be tempted to conclude that this is due to the naturalism offered by the NLU interface. In this study, however, we found that it's not the most natural version of NLU that maximizes these aspects of player experience. The participants felt that artificially being able to get a word in edgewise was very satisfying. This finding suggests that rather than striving for methods of interaction that seem most natural, designers should instead consider, for almost every experience goal, what types of explicit mediation to strategically introduce to shape the player experience.

Another important finding is that participants felt a greater sense of agency when the interface actively limited when they could take action in the game and instead guided them more about when their actions would be meaningful. This is further proof that we need to move beyond a naïve account of agency as free will that emphasizes giving the players the ability to do what they want when they want, and towards deeper accounts that take into consideration how players come to understand what actions it would be significant to take (and when it would be significant to take them) to impact the underlying computational system. In our case, even though the performance of the NLU system remained the same across different versions, it seems knowing when their actions would have meaningful impact made players more strongly experience conversation as a material for action.

Finally, all the players in our study had experience with playing videogames. While this means that all players came in with an understanding of how to navigate *Façade's* 3D world, it also means that the players came in with a bias toward current design conventions for non-player character interaction. As interactive drama becomes better known as a genre, new conventions for interacting with NPCs will arise, possibly changing these results over time.

NLU interfaces are promising systems that offer high degrees of freedom to players, and are of significant importance to the future of games as expressive and highly interactive media. In this study, we explored different pacing options that can be employed in NLU dialogue systems in games and their effects on gameplay, player experience and behavior. Our results indicate that the Holodeck dream for interactive narrative, that interactive worlds should strive to be as realistic and natural as possible, is not the ultimate end-goal for every type of interactive experience. Instead, different mediation conventions emphasize different aspects of gameplay experience; designers should consider what type of interaction is most appropriate for the experiences they want to create. Our results also provide further support for the idea that agency is not simply free will, but rather a complex phenomenon that involves creating desires and understandable opportunities for the player to act in a manner (and at a time) that will impact the underlying computational system. We believe our observations should be of interest to game designers who may want to explore using NLU interfaces in their games.

6

USING INFORMATION VISUALIZATION TO UNDERSTAND INTERACTIVE NARRATIVE

The future of video games promise highly interactive and flexible worlds which the players can shape to their desires and have a high degree of influence not only on the outcome, but also how each outcome is reached, how the story unfolds and how the player experiences the narrative. In designing such worlds it's of utmost importance that we are also able to study how people interact with them, and how different design decisions can influence players' interaction patterns in these worlds.

In the preceding sections, I discussed our initial exploratory studies on dialogue systems and their effects on player behavior in depth. In both experiments, we studied how different mechanics for dialogue systems change player perception and behavior in an otherwise similar interactive experience. While these experiments gave us invaluable insights into the deeper design issues in dialogue systems and games in

general, over the course of our studies, it became clear that while we had a firm grasp on how players' perceptions along dimensions such as flow, presence, engagement, sense of control and agency changed, we still didn't know if and how those changes in perception translated to or manifested themselves in in-game behavior. As a result, we started exploring players' gameplay logs and transcripts in more depth. The need for more and better tools to study in-game behavior in interactive, story-oriented experiences became apparent. For that purpose, we developed a log analysis and visualization tool that allowed us to easily visualize this information in a more digestible manner. The following sections describe this tool in more detail, and present two case studies using this tool: Our first case study is on different versions of *Façade*, and the second case study is on Prom Week, which is a recently released social game by Josh McCoy et al. (Josh McCoy et al. 2012). *Façade*, as discussed above, is a highly interactive game where player interaction can influence what topics the characters bring up, what they choose to reveal to the player about their problems and history, and ultimately what the fate of Trip and Grace's marriage will be. Parts of the story may also be enacted differently depending on player's past actions and affinity towards Trip or Grace. This flexibility results in a huge variability on how players experience the narrative of *Façade* – changes are possible not only on what is being told, but also how it's being communicated to the player. Presenting information from many player logs is a challenging task, which we show, can be made easier by using information visualization techniques. The subject of our second case study, *Prom Week*, is a recently released game by Josh McCoy et al (Josh

McCoy et al. 2012). In Prom Week, the player can interact with any of the various characters in the game world through a variety of social moves such as flirt, insult, bully, etc. to achieve various goals dictated by the campaign. These goals might range from being friends with at least three people by the end of the campaign to breaking up some character's relationship in order to date that character. *Prom Week* tries to merge the richly-realized characters found in heavily-authored games such as *Mass Effect* and *Façade*, and the richness and variation in social interaction possibilities found in more "simulation-heavy" games such as *the Sims* Series (Maxis 2000). I believe those two case studies, taken together, provide substantial evidence towards the usability of information visualization techniques for studying game design and interactive narratives.

RELATED WORK

While information visualization techniques have been used as creative tools for input in some works of digital media (Romero, Pousman, and Mateas 2008; Pousman et al. 2008; Holmquist and Skog 2003), they are very rarely employed as functional tools to study digital games. While interest in using information visualization techniques have definitely grown recently, with many scholars and experts in the field pointing to the importance of analytics in game studies and design (Drachen and Canossa 2009; Medler 2012; J. H. Kim et al. 2008) and a possible important role for information visualization to aid in such analyses (Bowman, Elmqvist, and Jankun-Kelly 2012), actual tool implementations and evaluations remain rare.

One example is the work on *the Restaurant Game* by Orkin et al. (Orkin and Roy 2007). *The Restaurant Game*, by the authors' own definition, is a "minimal investment multiplayer online" role-playing game where the player can assume the role of a customer or a waitress, and try to achieve a simple goal such as earning money or having dinner. The authors collected sequences of actions and utterances from players' interactions with *the Restaurant Game*, which they then used to build statistical models of expected patterns of behavior and language for particular goals (Orkin and Roy 2009). In their work, they also present superimposed visualizations of physical actions of players from many logs. Another work that is focused on visualizing physical actions in virtual worlds is the work by Hoobler et al. (Hoobler, Humphreys, and Agrawala 2004) in which team strategies and player behavior patterns are visualized on level maps in the game *Return to Castle Wolfenstein: Enemy Territory*. Similar work has been done on *Halo 3* maps by Microsoft (Clive Thompson 2007) and by Chittaro et al (Chittaro, Ranon, and Ieronutti 2006). BioWare's SkyNet (Georg Zoeller,2010) uses heatmaps to indicate where crashes in game maps occur. A more recent work is Data Cracker by Medler et al (Medler 2012), which is a visual game analytics tool that presents visual representations for various gameplay metrics collected from Dead Space 2 (Visceral Games 2011).

The visualization of Choose Your Own Adventure (COYA) books by Christian Swineheart (Christian Swineheart) is also a relevant work that uses information visualization techniques to gain insight into the evolution of these books over the

years. Swineheart produced visualizations of twelve COYA books published between 1979 and 1998 by color-coding the endings from great to catastrophic with respect to the desirability of each, and looked at how the number of endings and the linearity of the plot changed over time, along with visualizations of the different paths that the story can take. While story-oriented digital games can be likened to COYA books, a COYA book offers a fixed number of choices at every step, and therefore it's actually possible to construct all possible paths that a player can take while exploring a narrative, while software is a much more expansive medium that presents a unique set of challenges for visualization. As an example, *Façade* offers many choices to the player at each step, along with global mixins that can be initiated by the player through interactions with objects, by referring to certain topics or through certain actions.

Perhaps the closest work in the literature is Playtracer by Andersen et al (Andersen et al. 2010). Playtracer is a generalized heatmap that can be used to visualize state changes in games with discrete states. It employs clustering and multidimensional scaling techniques so that games with many states can be practically visualized. Player experience is modeled as transitions between states, which makes Playtracer independent of the state structure of the game. The main focus of Playtracer is also player activity and state changes based on that activity. For story-based experiences, while player actions are usually limited in number, each possible sequence of story beats represents a different state. Furthermore, Playtracer doesn't capture the

temporal aspects of player experience. For a story-based experience, we are also interested in when players experience certain story sequences, what story bottlenecks are, what sequences are likely to occur frequently, and so on.

In contrast to the work cited above, which mostly focus on studying player actions in games where the main gameplay activity is navigation, shooting or object interaction, our work is focused intensely on analyzing interactive narratives for which shaping the narrative is the ultimate gameplay activity. Our tool was also designed to be more utilitarian: Our primary goal is providing more functional tools to actually study how players interact with story-oriented experiences, and inspire in designers how they can devise and study important quantitative metrics based on graph formalisms which would otherwise be difficult to come up with. The resulting tool provides designers with not only good looking visualizations, but also with graphs and formalisms on which, as we will show below, further complicated analyses can be carried out.

FAÇADE LOG ANALYSIS AND VISUALIZATION TOOL

A screenshot of our tool is given in Figure 29. While the tool we developed was initially aimed at analyzing *Façade* gameplay logs and visualizing metrics calculated from these logs, over the course of our work on it, we realized its potential for analyzing other works of interactive narrative, and started brainstorming within a more general paradigm as to what visualizations might be useful to interactive

narrative designers and authors in general. This tool was developed using Visual C#, the open source GraphSharp llibrary (*Graph#* 2009), the Javascript InfoVis Toolkit (Fekete 2004) and custom code.

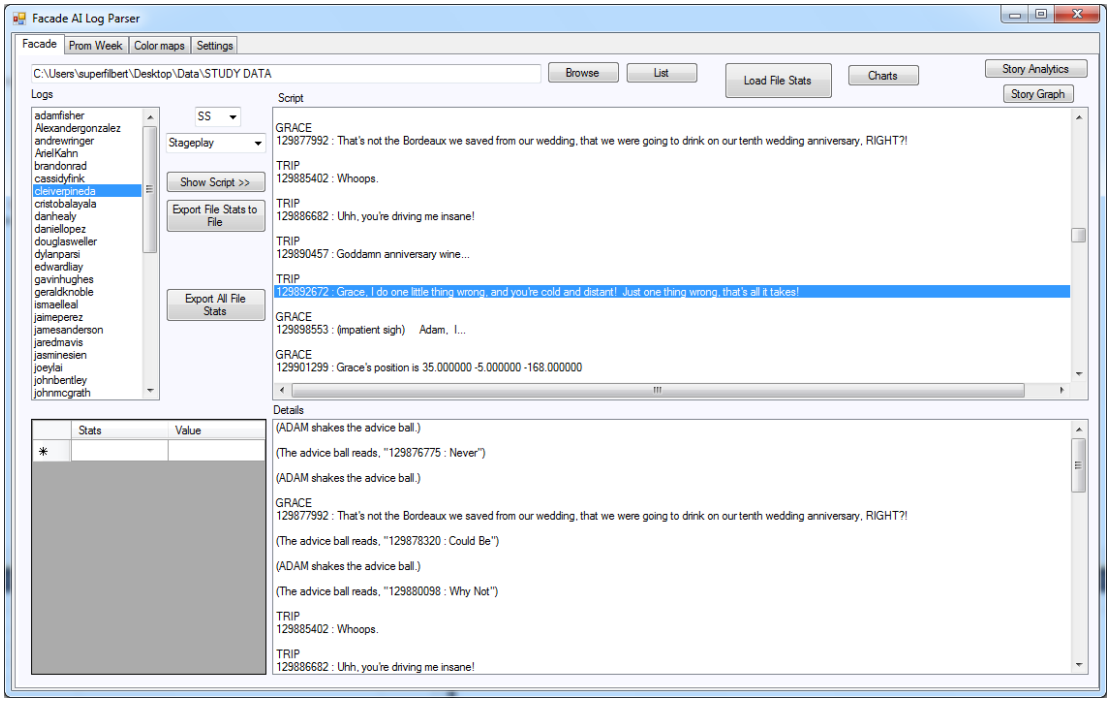


Figure 29. A screenshot of our analysis and visualization tool.

The tool consists of a parser layer that can parse a game's output and convert it into a proprietary format from which the plotting layer of our tool can then build the specific visualizations we propose. I will discuss the visualizations we came up with in more detail in the following sections.

Visualization techniques employed

Heat-maps

Our visualizations make use of the heatmapping technique frequently used in information visualization. Heatmapping is a century-old technique based on the idea of color-coding data depending on values for a more understandable representation (Friendly 2009). We make extensive use of this technique in our visualizations, and our tool supports creating custom color scales for each visualization, with customizable color schemes and different data range-color mappings (Figure 30).

Figure 30. Our visualization tool supports creating custom color scales. This is a squarified tree-map of the various discourse acts players used in dialogue.

Squarified tree-maps

Originally proposed by Johnson and Schneiderman (Johnson and Shneiderman 1991), treemapping is a widely used information visualization technique used to

visualize hierarchical data in a space efficient representation. Treemapping techniques mostly differ by the tiling algorithm they use, which change the shape and size of the rectangles in the resulting graph. In our work we chose to use squarified treemapping algorithm (Bruls, Huizing, and Wijk 1999), which aims to make the rectangles as square in shape as possible. This makes for a more understandable and compact representation, and also makes interacting with them using the mouse easier. Our treemaps allow both hierarchical and single-level representation (Figure 30).

Story-sequence view

We also developed a custom visualization that presents a comparative view of individual gameplay traces (Figure 31). Each column is a representation of one particular gameplay sequence, and the sequence of nodes in each column represents the story pieces (in varying level of granularity) that the player experienced in that play-through (Figure 37). There's a line connecting nodes in column i and i+1 if the story piece occurs in both play traces i and i+1, and the color of each node changes along the color scale depending on how many subsequent play traces that particular story piece occurs. This visualization allows us to see which story pieces are more constantly experienced by players, as well as seeing where more variation occurs along the story lines player encounter.

While I developed this visualization independently, it has afterwards come to my attention that similar visualization techniques exist for other domains – History

Flow, proposed by IBM Reseach labs (Fernanda B. Viégas and Martin Wattenberg 2003), was used to visualize the evolution of Wikipedia articles over time as a result of collaborative edits. Parallel coordinates by Inselberg (Inselberg 2009), used to visualize relationships among different dimensions in multivariate data, results in similar visualizations.

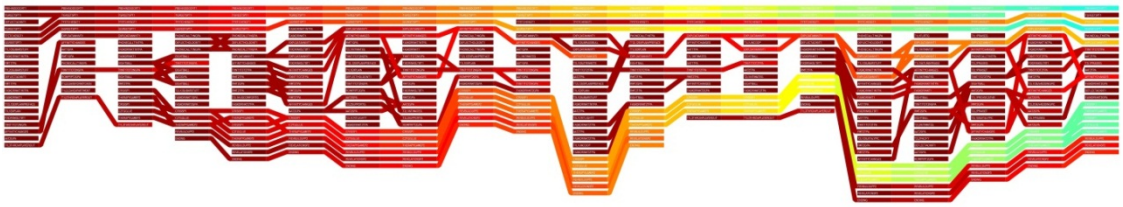


Figure 31. Story-sequence view.

Story graph

To visualize story space coverage we use a graph that we call a "story graph". We formally define a story graph as follows:

1. Each node represents a story unit.

2. There's an edge <u,v> from node u to node v, if beat v follows beat u in some gameplay log.

3. Each Node u is labeled <StoryUnitId indegree outdegree>, where beat id is a unique identifier for u, indegree is the number of incoming edges to node u, and the outdegree is the number of outgoing edges from node u.

4. Each node is colored based on its indegree.

This visualization allows us to see at a glance which story beats occur most frequently, which ones represent "story bottlenecks" and approximately what percentage of the story space the player was able to discover.

Story tree

The story tree structure is a variation on the graph introduced in the previous section. Instead of having just one node for every beat, it also takes into consideration the sequencing of those story pieces – so each node is instead defined by a combination of both the story beat label, and in what order it occurs in the gameplay trace. The resulting visualization allows us to see the variation among different gameplay traces in a lot more depth.

CASE STUDY I: *FAÇADE*

Our first case study was focused on the usability of our tool both for gaining insight into player behavior and for performing comparative analysis of different versions. For our comparative study, we used gameplay logs from two different versions of *Façade* – the sentence selection version and the NLU version (Figure 32). Our participants were mostly from the undergrad population at UC Santa Cruz. They played both versions in random order to account for learning effects, and we then used the data from their gameplay logs to perform a study of those interfaces using our tool.

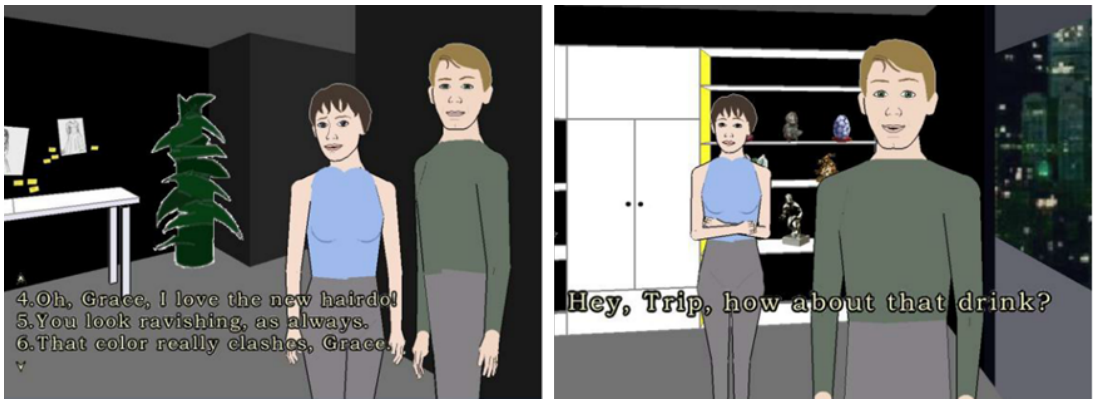


Figure 32. Versions we used for our case study. Sentence-selection (left) and the NLU version (right).

Discourse act coverage

As discussed previously, *Façade* maps each player utterance to a discourse act, which aims to capture the semantics of the line in the context of the game world. For example, greeting utterances like "Hi, Trip!" or "How are you doing, Grace?" get mapped to "Greet Trip" or "Greet Grace" discourse acts respectively. *Façade*'s NLU version maps player utterances to approximately 30 main discourse act categories and each of those categories can have a varying number of parameters. The player's expressiveness is directly related to how many of those discourse acts they are able to employ using different interfaces when conversing with the characters. We designed a representation based on the squarified treemapping technique to visualize the average percentage each discourse act is addressed over all gameplay logs. In our representation, the color of each rectangle is determined by the average percentage each discourse act is addressed (Figure 33).

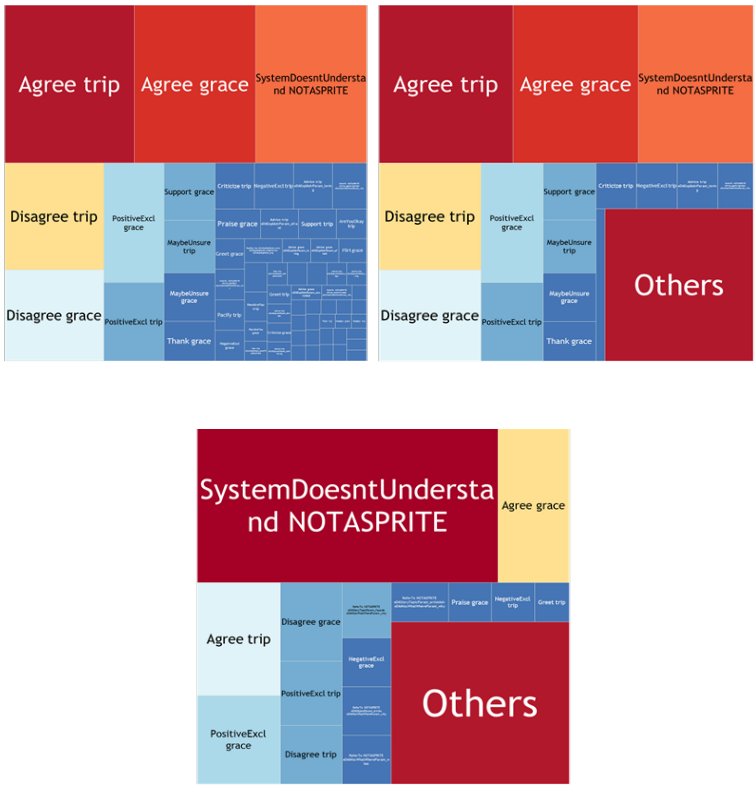


Figure 33 Squarified treemaps for discourse act usage patterns for (a) the sentence-selection version with all the discourse acts, (b) the sentence-selection version with discourse acts addressed below 1% grouped in others category, and (c) the NLU version with similar grouping.

The most striking difference between the representations above is the relatively high percentage of the "System Doesn't Understand" category in the NLU version, which means the NLU module was unable to map the player utterance to any of the discourse acts it understands.

Another interesting point to note is how the players were inclined towards empathetic responses in both versions. Positive discourse acts such as Agree, Positive Exclamation, and Support make up a much higher percentage of discourse

acts employed in both versions than negative responses, especially when we look at the discourse act distribution without considering the parameters (Figure 34).

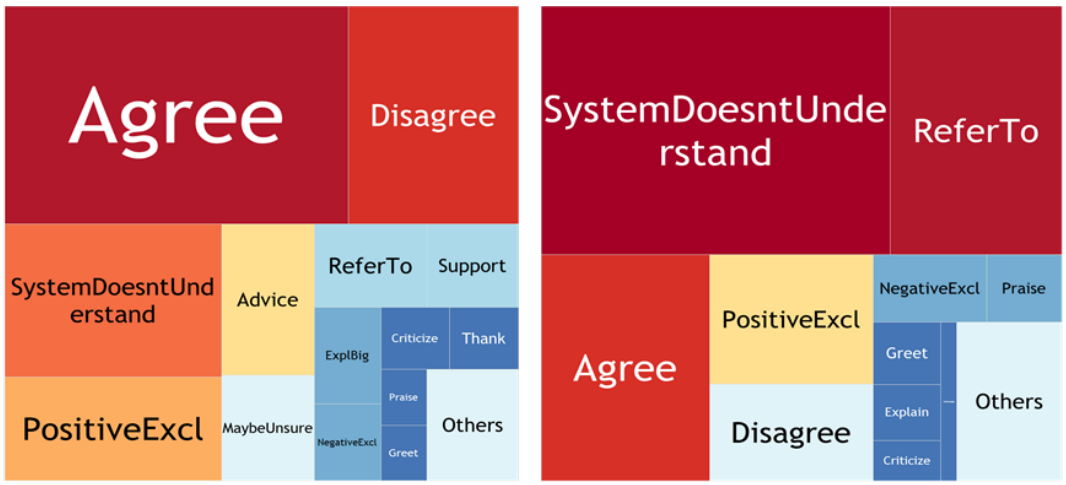


Figure 34. Squarified treemaps for discourse act usage patterns not considering the parameters for (a) the sentence-selection version and (b) the NLU version.

We also looked at how many unique discourse acts (including the parameters) the players were able to discover in a single playthrough. That number ranged from 14 to 35 for the NLU version, and 15 to 33 for the sentence-selection version, which is very similar and very low given the huge set of discourse acts available to the players. However, given a large percentage of the discourse acts were not understood by the system in the NLU version as evidenced by the treemaps above, we can deduce that the players were able to enact a higher number of influential actions in the sentence-selection version compared to the NLU version.

Story space coverage

A close-up of the story graph from 10 logs from the NLU version is shown in Figure 35. In the example below in Figure 35, we can clearly see that players were more likely to encounter the story piece where there's a discussion over which drinks to have (FASKDRINKT1NTPA) than the beat where Grace complains about the apartment's decoration (AAT1GPA). The full versions of this graph for both NLU and sentence selection versions are given in appendix C.

Looking at the graphs, it's also evident that the graph for the NLU version has more nodes than the graph for the sentence-selection version. Players were able to discover 57 different beats in total across 10 playthroughs in the NLU version. Playing the sentence selection version, they were only able to go through 43. This implies that the variation in story among different playthroughs was higher in the NLU version compared to the sentence selection version, which is immediately visible when we compare the graphs. Even though a large percentage of the discourse acts were not understood by the system in the NLU version as evidenced by the treemaps we discussed above, players were able to strategically insert discourse acts using the NLU version, which resulted in more story content being revealed to the player. A look at the average in-degree of nodes also reveals that the sentence-selection version graph has slightly higher average in-degree (2.37) than the NLU version graph (2.2).

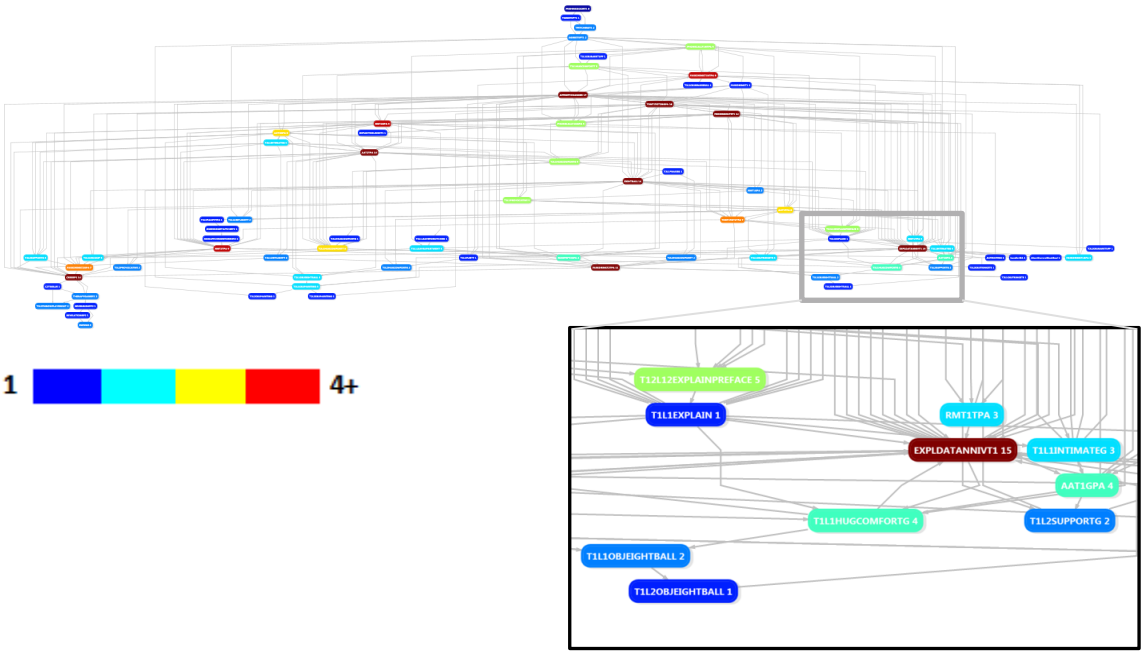


Figure 35. A close-up view of the story graph.

We also used the same technique for visualizing the average expected time players are likely to spend in each beat (Figure 36). The results across two versions were quite similar, with participants spending slightly more time on average in the therapy game beat in the sentence-selection version compared to the NLU version. This visualization also reinforces the fact that players are likely to encounter more variation in story using the NLU version, as the number of tiles in the map for the NLU version is clearly higher than the map for the sentence-selection version.

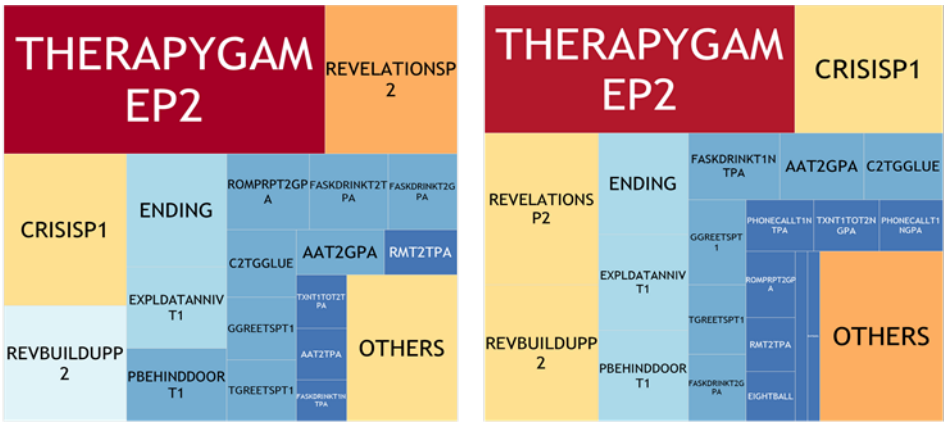


Figure 36 Expected average time spent in each beat for (a) the sentence-selection version, and (b) the NLU version.

Story timelines

As seen in Figure 37, at the story beat level, in terms of generic story structure, players experience a lot of variation between the greeting and revelation beats – most of the time, two gameplay traces almost entirely differ in content in between those sections. The length of gameplay session also varies considerably.

However, this graph doesn't tell the whole story. *Façade*'s therapy game and revelation beats are structured in such a way that a single beat handles the selection among a big pool of possible questions or issues that can be discussed with the player and many different revelations that Trip and Grace can make depending on player's actions up to that point. To take this inquiry further, we created the same visualization in the level of therapy game mixins and revelations combined, and individual utterances (Figure 38). The resulting graphs still show that players encounter a relatively large amount of variation in what therapy game mixins are

played and what revelations Trip and Grace make. Even in the utterance level the stories experienced by the players are quite different. Most of the commonly occurring story pieces are generic utterances that connect mixins, such as “Let’s talk about of us” during the therapy game, or “You are driving me insane” at the beginning of the crisis section.

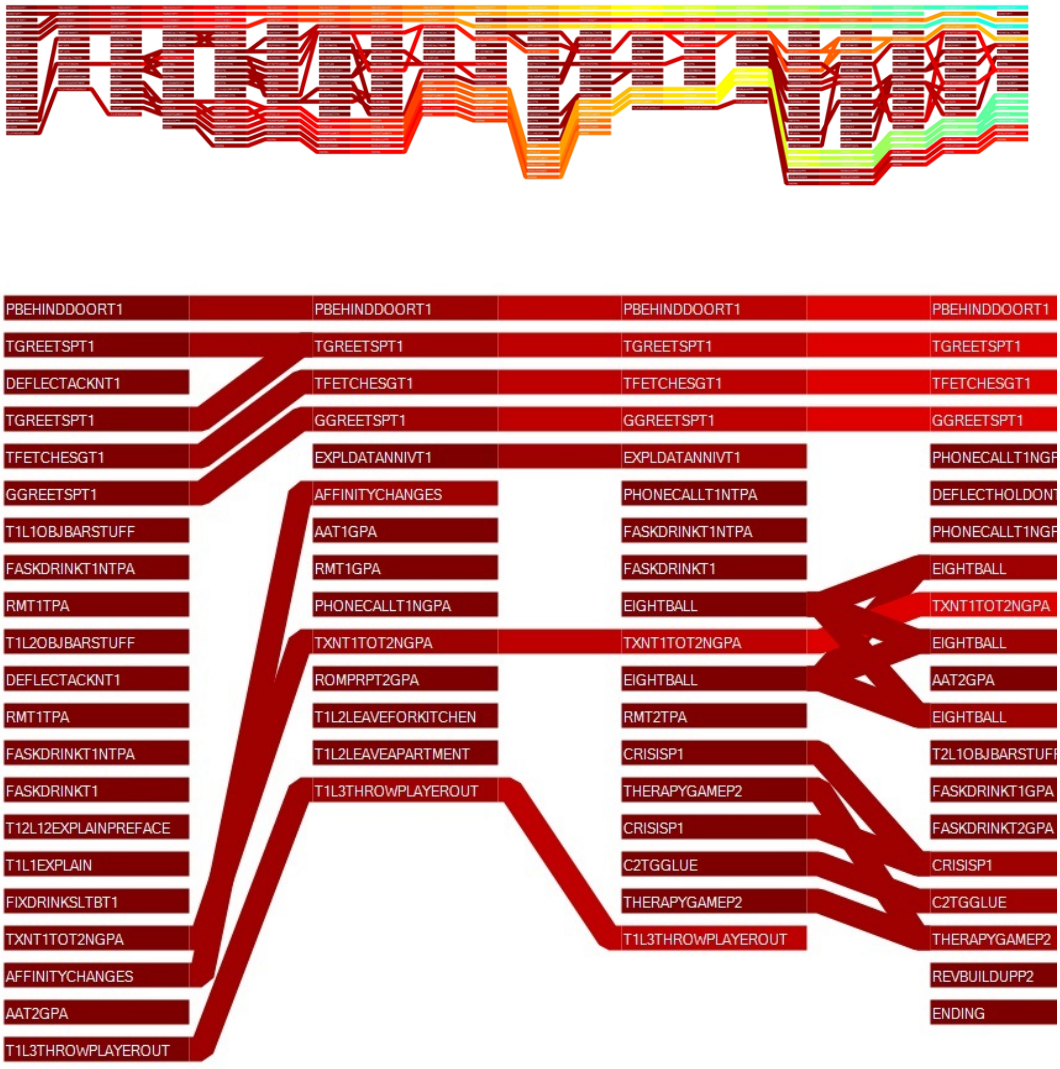
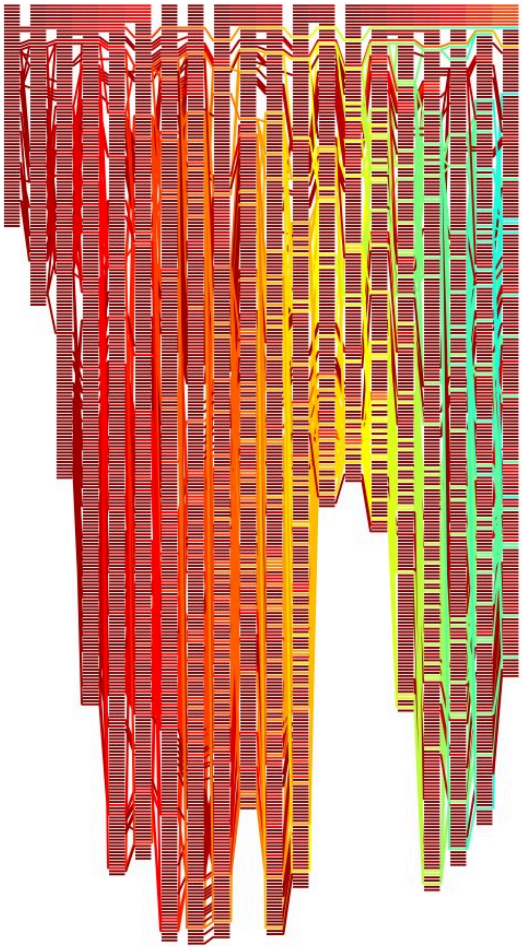


Figure 37. Story timeline in story beat level (top) and a zoomed-in version (bottom).

(a)



(b)

Figure 38. The story timeline in (a) therapy game mixin and revelations and (b) utterance level. While there are still similar therapy game mixins, revelations and utterances the players encounter in almost every gameplay trace, the experiences are mostly unique.

CASE STUDY II: *PROM WEEK*

Our second case study is on *Prom Week*. Prom Week has been recently released to critical acclaim, and has been nominated a finalist for technical excellence at IGF 2011. *Prom Week* is powered by the social physics engine *Comme il Faut* (*CiF*) developed by Josh McCoy (McCoy et al. 2010). *CiF* attempts to enable a new type of gameplay experience in which players can engage with a complicated simulation of social relationships that is as mallable as simulations of physics in games, and player actions have considerable, cascading impact on the state of the emulated social network of social relationships between the characters in the game world. Using *Prom Week* as one of our case study subjects allowed us to gain insights into the usability of our tool for analyzing such complicated design spaces and with large datasets.

In order to get a sense of how *CiF*'s simulation and *Prom Week*'s gameplay impact the actual choices presented to the player, level traces were analyzed and visualized to form an understanding of how players were interacting with the release version of *Prom Week*. In *Prom Week*, the player can choose among many possible social exchanges such as bully, annoy, share interest, ask out or brag. Every social exchange is associated with many instantiations that consist of multiple lines of dialogue exchanged between the characters. In our visualizations, we collapsed multiple instantiations of the same social exchange together since we were more interested in differences in the level of social exchanges. Even though the player has many options of social exchanges to choose from, it is not clear without evaluation

that there are enough paths through the story space to satisfy the whims of each individual player. Furthermore, story goals, level casts, and the desires of the characters themselves may restrict the options available in such a way that many players will be forced down a narrow few paths in their pursuit of story goals.

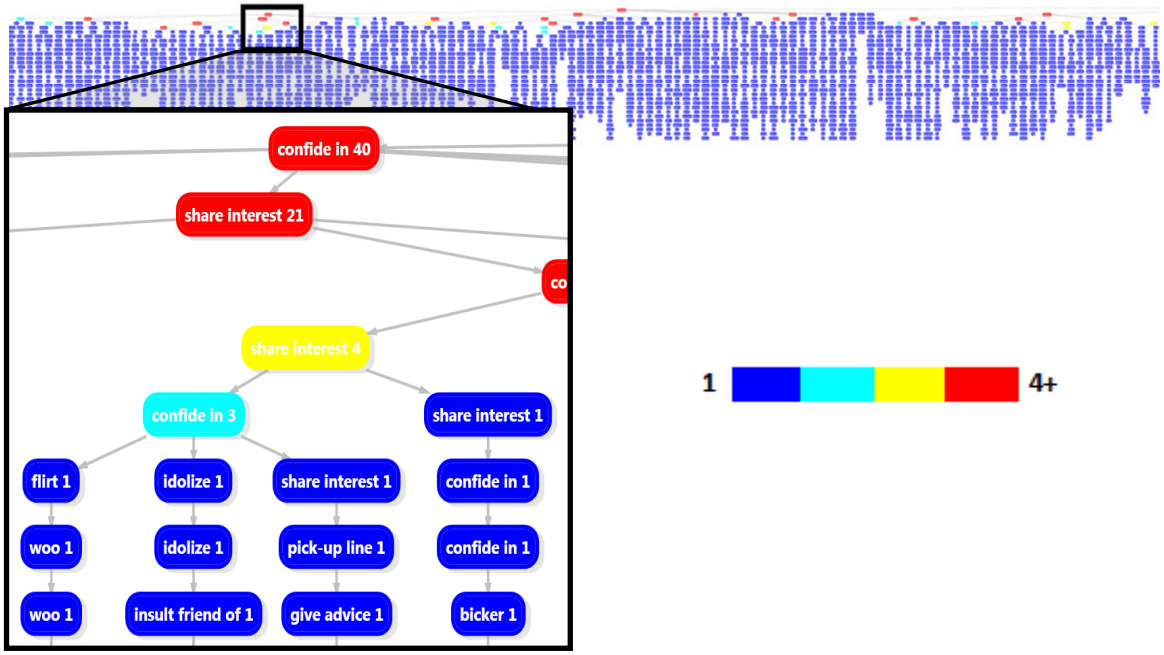


Figure 39. A play trace graph showing how often each distinct path through Simon's story was taken (shown by the color and number associated with each node). The large band of nodes seen at the top of the diagram represents approximately one third of the total size of the complete map. The cutout shows a section of the map in detail including examples of social exchanges (like "pick-up line" and "confide in") that appeared in more than one play trace. The majority of play traces are unique.

Our tool allowed us to discover that there was a very large degree of variation in the way that players navigated the social space. Examining a tree map representing the social moves selected during the final level of Simon's campaign reveals that, of the

263 unique playthroughs we analyzed, no two were exactly alike; the space was rich enough to allow for an entirely unique play trace per player. Figure 39 is a tree graph of the play traces analyzed for Simon's campaign. Each node represents a selected social exchange, each of which results in changes to the game state (e.g. relationships starting or ending). A path through the tree is the sequence of social exchanges a player made from the starting state in the first level (the root), to an ending (a leaf). Although there are a fixed amount of maximum turns in Simon's campaign, not all paths in the tree are the same length as players have the option of skipping remaining turns and jumping ahead to the next level. The color of the nodes is a heat map indicating frequency of node visitation along that specific path; red is frequently visited (i.e. several players followed that exact same route up to the point of that node), and dark blue means visited only once (i.e. the route to that node was experienced by only a single player). For readability purposes, the names of the nodes have been collapsed to the names of social exchanges selected, when in actuality gameplay moves are identified by the social exchange and the two characters involved as initiator and responder in that social exchange. Including this differentiator would have further increased the branching of the tree, but we claim that it is already branchy enough for the purposes of validating the high variability in Prom Week.

The average indegree (times a node was encountered by a player) of a node in this graph is approximately 1.11; though as mentioned above there are a few nodes

towards the beginning that are selected many times--"share interest" and "confide in" are popular starting moves, happening 91 and 40 times respectively—the vast majority of nodes are visited precisely once.

The analysis and visualization tool we developed also allows us to carry out n-gram analysis: Performing this analysis revealed some interesting statistics on the patterns of sequences of social moves played (this analysis is explored in more detail in the next section). Using 1-gram analysis, there are 38 unique social moves that players employed on this level, out of a total possible 39 social moves that exist in the game. Using 3-gram analysis, we have 2521 unique patterns, of which only 80 appear more than 10 times. With 6-gram analysis, there are 5066 unique patterns of

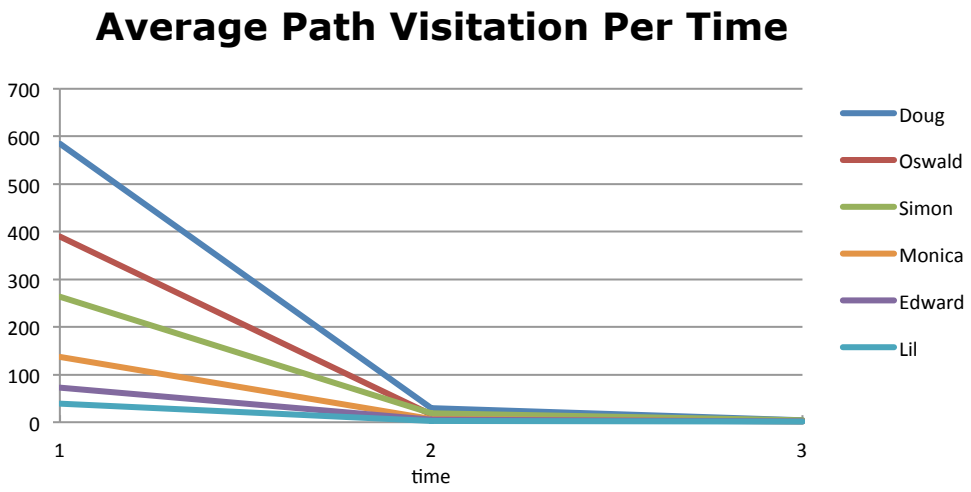


Figure 40. This plot shows how unique each player's path through the story space is as time progresses. The x-axis is time, or number of turns, and the y-axis the average of how many times a story path has been visited.

social exchanges, one of which occurred 16 times, another 10 times, and all the rest less than 5 times. The fact that so many separate patterns exist, with so little repetition, indicates that players were able to find their own way through the story space. Moreover, the n-grams that have the most repetition are situations in which the same social exchange was played multiple times in a row. Though apparently there is a player type that relies on a strategy of brute force (for example, attempting to 'woo' six times in a row), they are dwarfed by the number of other patterns exhibited.

We discovered another interesting point by examining the tree graph of player choice. The sheer breadth of the tree gives a positive view of just how much variability there is in player choice; not only does the system allow for variability, but players are taking advantage of it as well. Additionally, though there are only 11 nodes that players chose for the first move, there are 79 different nodes selected for the second, and 143 for the third. By the fourth turn, nearly every gameplay trace is unique (see Figure 40). Even traces with subtle differences in gameplay actions (for example, the sequence of social actions "reminisce", "confide in", "ask out" as opposed to "confide in", "reminisce", ask out") can result in remarkably different traversals through the social state, as *Prom Week* keeps track of the specific social exchanges and instantiations that the user has seen and incorporates them into future social exchange selection. Moreover the specific ordering of social changes

also impacts the formulation of which social exchanges characters want to play with each other, thus even seemingly similar play traces can be considered unique.

The general trend of paths becoming unique can be seen across the stories and is even more prevalent in the more difficult stories of the late game. Take Oswald's story as an example, which has 390 level traces that all begin in the same starting state. Twenty-five different opening moves were selected with an average indegree of 15.6. After the second turn the average drops to 2.36. The average dips to 1.27 after the third turn, and hits 1.07 after the fourth.

The above analysis supports our claim that our tool can be used across different games to analyze the variability in how players experience the story of Prom Week. Using our tool, we were able to come up with quantifiable measurement of changes in player experience inspired by established graph formalisms for a game as different from Façade as Prom Week, which in turn allowed us to prove the successful implementation of two important design goals for it. First, we saw that the low average indegree indicates that Prom Week can create a completely unique playthrough experience for each player; the low valued n-grams indicate that these unique playthroughs consist of different patterns of play; and the rapid branching factor means that the little overlap that does exist between players quickly separates into distinct traces. Given these results, with the help of our tool, we can claim, through quantitave results and analysis, that *Prom Week* was successful in providing a game space with large amounts of variability, even if, as we see below, players

selected between only a handful of the total possible options on the first turn. Second, our tool also allowed us to show that *Prom Week* is specifically providing large variability in the service of making stories playable. The relatively low variability seen during the first turn is actually positive evidence for this second hypothesis. There are five characters in Simon's first level, and each character wants to engage in five possible social exchanges with each other character (the top five social exchanges character A wants to perform with B given the desires computed by *CiF* for character A). Since the player picks a unique initiator and responder, this means that there are at least 100 potential opening social exchanges (the actual number is a little higher, as players can spend story points to unlock additional options).

The fact that, of these hundred starting options, only eleven were ever pursued between all of the gameplay traces implies that players have been attempting to accomplish story goals. The beginning of each level provides framing text that contextualizes the characters' relationships to each other with relation to campaign goals, and offers small hints about how to accomplish the goals. The hints take the form of advising the player on which characters to form relationships with, but offer no advice on which specific social exchanges to try. This means that player actions are being motivated by story goals without being dictated by them, providing a solid foundation for our second hypothesis.

Strategy Driven Play

In addition, we will also show that our tool can be used to determine if players actually engage in strategy-driven gameplay – that is, their actions in the game are actually motivated by a desire to accomplish the goals the game asks them to. To determine if *Prom Week* promotes strategic play, we analyzed the player-driven paths through *Prom Week* with respect to the successful completion of story goals. To be seen as an indicator for strategic play, large portion of the story paths - variable though they may be - need to lead to successful goals. Story goals in *Prom Week* represent story states for the player to make true in the storyworld. For example, in Simon's campaign, the player is tasked with accomplishing five distinct goals, including having Simon make five friends, having Simon begin dating someone, and giving Simon an "ideal rival" by making him friends and enemies with the same person. The combination of goals accomplished determines which ending for the campaign the player receives. Though endings are mostly pre-written to leverage authorial control, there still exists template dialogue within endings that allows for explicit references to specific social exchanges that were chosen by the player throughout the course of gameplay. This gives every choice the player makes—and not just goal completion—an impact on the campaign's climax.

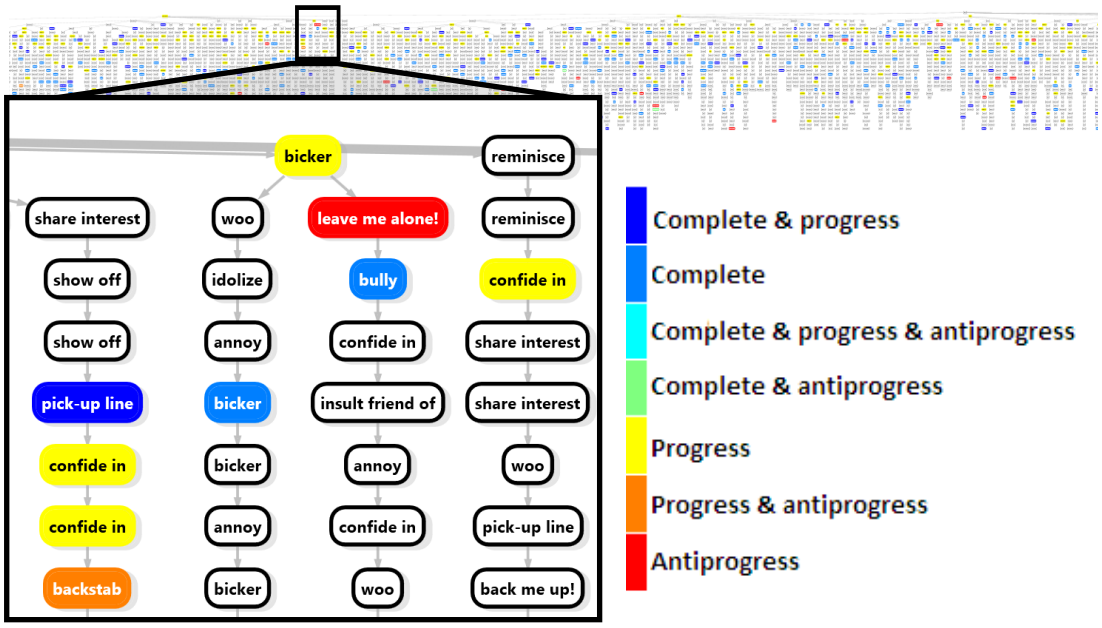


Figure 41. A tree displaying the amount of progress towards goals in Simon's campaign. The color of the nodes represents the type of goal progress. There are three types of goal progress that can be combined in any way. Complete (Blue) means a goal was completed, progress (yellow) means that one aspect of a goal was made true, and antiprogress (red) means that an aspect of a goal that used to be true was made false. White nodes mean that no progress (or antiprogress) was directly made by making that social exchange, though the social state was still changed which could lead to progress in future turns. The large band of nodes along the top still represents about 1/3 of the total play traces of Simon's story.

Story goal completion

Figure 41 shows another view of the 263 traces that start at Simon's first level and progress their way through the end of his campaign. In this graph the color of the nodes shows the impact with respect to the story goal of that node. Story goal completion ranges from dark blue to green, progress toward the goals is in the range of light blue to orange, and moving the social state away from the story goal (antiprogress) is colored orange or red. This data was generated by taking the same

level traces used to generate Figure 39 and running them through *CiF*, keeping track of the goal accomplishments at each game turn.

Simon's campaign is the third non-tutorial level in *Prom Week* and is of intermediate difficulty. Though some goals can be accomplished in just a single turn (across all 263 traces for Simon's campaign, only 13 completed a goal on the first turn, and only 17 completed a goal on the second), the rest take several turns to complete. As seen in Figure 41, the story goals were completed by players at many points along the story paths. Of all of Simon's traces, only a single one did not contain any goal progress. All others exhibited at least some amount of effort towards achieving story goals.

Even though Simon's campaign is of intermediate difficulty, players still displayed an aptitude for achieving goals. Between all of the play traces, goal completion (on any of Simon's five goals) was reached a total of 610 times (average of 2.32 goals per player). If every trace from every file had accomplished all five goals, the total would be 1,315, which means that around 46% of all possible Simon goals were achieved. Goal progress was made a total of 837 times (average of 3.18 times per player), and goal antiprogress was made a total of 44 times (average of 0.18 times per player).

A concern when designing goals is that *Prom Week*'s gameplay—manipulating social relationships within a setting of cascading social influences in the pursuit of story goals—is fairly unique. Since *Prom Week* serves as an introduction to this genre of social puzzle game for most players, figuring out the nuances of the system to make

story progress could have proven to be a challenge. Although the goal completion rate is perhaps a little low for a campaign of only intermediate difficulty, the results are encouraging because not only were players motivated to pursue story goals, they were also able to create a strong enough internal model of the storytelling system to be able to pursue story goals with some amount of success.

DISCUSSION & CONCLUSION

In this chapter, we have presented a visualization tool that aims to enhance our current toolset for studying interactive narratives. We have also demonstrated the usability of our visualizations in two case studies: one using two different versions of *Façade*, and another using Prom Week. In story graphs and story trees we introduced unique visualization structures that aid in analyzing how people experience the story in both games with respect to the sequential nature of story-based experiences. Our analysis allowed us to use graph formalisms such as average indegrees and outdegrees to get an idea, expressed in numbers and metrics, of how much variation a game allows players to experience and how different choices in design can influence this variation, how quickly players' experiences branch out and become unique, and to what degree a game inspires strategy-driven gameplay.

In contrast to other visualization tools, our tool follows a utilitarian design approach: It's designed to both present information in a clear, easily-digestible manner, and also inspire further analysis based on graph formalisms and structures. Instead of navigation and combat, it's mostly aimed at visualizing story-oriented

experiences in which the player experiences a narrative that is likely to vary deeply depending on player interaction, and it's meant to be more useful in this domain, in contrast to state-based tools like Playtracer, or map visualizations heavily employed in first-person shooters and other action games. Our tool is mostly aimed at experiences in which shaping the story is the most important gameplay activity rather than navigation and object interactions, and it focuses on visualizing player actions and story events rather than states. It lets designers see beyond the theoretically possible story space and get an idea of the feedback loop between players and systems on how they behave with respect to each other. It should be noted, however, that our tool assumes the player experiences a progressive narrative along a timeline. State-based approaches might still be a better fit for studying navigation, combat or environment interaction, or keeping track of cumulative gameplay metrics.

Our tool and visualizations allow us to formulate and evaluate metrics related to those measures and carry out more quantitative analysis. Our results from our two case studies show that information visualization techniques have the potential to be useful in the field of game studies where highly interactive and flexible story worlds are poised to become a norm.

7

EXPLORING QUANTITATIVE METRICS OF PLAYER SATISFACTION IN DIALOGUE SYSTEMS

The insights we gained from the visualization tool discussed in the previous chapter, combined with our analysis of qualitative data gained from our interviews in our previous studies, allowed us to study quantifiable aspects of player behavior in more depth. In this chapter, I will discuss the results from another study we conducted in which we focused more on quantitative, countable metrics that we thought is of possible importance to player satisfaction from a dialogue system in a game. In this study, we compared the sentence-selection and NLU interfaces in a within subject experiment. Based on our exploratory analysis, we developed a 12-item survey which the participants filled out after playing each version. During the gameplay session, we also collected gameplay metrics pertaining to how users interact with a dialogue system in a game. The rest of this chapter discusses those metrics and the results we obtained from mining players' gameplay logs.

STUDY DESIGN

For this study, we recruited 48 people in total, most from an undergraduate game design class open to all majors. Data from two sessions had to be thrown out due to problems in the testing session, leaving us with forty-six participants. The mean age of our participants was 20.1, with a standard deviation of 2.2. As in our previous studies, we recruited only native speakers with previous gaming experience. A further breakdown of our participant demographic can be found in the appendices section.

Measurement instrument

In order to measure player satisfaction, we developed a survey that consisted of twelve seven-item Likert scale questions. We constructed the questions in this survey based on the results from our first two exploratory studies, which I described in the previous chapters. Our questions are given in Table 9.

Collected metrics

After the study was finished, we analyzed players' gameplay logs and collected the following metrics:

Number of revelations: Towards the end of the game, depending on player interaction up to that point, Trip and Grace can make a number of revelations to each other.

Table 9. Our survey for dialogue systems in games. The questions in this survey was constructed based on our results from the previous two exploratory studies, the results of which I discussed in the preceeding chapters.

Q#			1 2 3 4 5 6 7	
1	It was easy to decide what I want to say using this interface.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
2	Once I decided what to say, it was easy to communicate that to the characters using this interface.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
3	The responses of the characters to my dialogue actions made sense to me.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
4	I felt in control of the conversation.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
5	The choices I wanted to make were present in the interface.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
6	The ending made sense to me.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
7	The number of interactions the interface allowed me to have was satisfactory.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
8	The interface gave me enough time to make decisions on what to say.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
9	The interface was easy to learn.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
10	The interface was easy to use.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
11	I enjoyed using this interface.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
12	I felt the ending was a direct result of my interactions.	I disagree.	⬭ ⬭ ⬭ ⬭ ⬭ ⬭ ⬭	I agree.
			1 2 3 4 5 6 7	

Ending score: We came up with a scoring scheme based on the desirability of endings. We detail our scoring scheme below. We also calculated a cumulative score that is the sum of ending score and number of revelations.

Number of deflects: When Façade's NLU system doesn't understand a player utterance, or cannot make sense of a player action in the current context, Trip and Grace choose to deflect the player. This variable reflects how many times a deflect reaction occurred during a player's gameplay session.

Number of "System Doesn't Understand" discourse act mappings: This variable stores how many times the NLU system wasn't able to map the player utterance to any of the discourse acts it understands.

Number of interactions: This variable keeps the total number of interactions players perform. It's a sum of dialogue actions and gesture interactions such as kiss, hug, or pickup object.

Number of gesture interactions: This metric keeps track of how many gesture interactions the player performed.

Number of dialogue interactions: This metric keeps track of how many times a player chose to speak during the game.

Average time between dialogue interactions: This metric keeps track of how much time elapsed on average for each player between dialogue interactions.

Average time to enter input: For the NLU version, this variable keeps track of the elapsed time between when the player started typing and the time they finished entering input. For the sentence-selection version, it keeps track of how long the player took to make a selection among the options presented.

Number of affinity changes in favor of Trip and number of affinity changes in favor of Grace: During the first half of the time, Trip and Grace put players in a series of situations that force them to take sides. These two variables store the number of times when players' affinity switched in favor of one of the characters.

Number of Trip utterances and number of Grace utterances: These two variables keep track of how many times Trip and Grace chose to spoke.

Task success rate: We kept task success rate as 1 for the sentence-selection version, and through our manual annotation found it to be 0.73 on average for the NLU version.

Number of unique discourse acts discovered by the player: This metric keeps track of how many unique dialogue actions players were able to discover. We calculated two versions of this metric: one that doesn't take into account the parameters of the discourse act which define the more granular details such as who the utterance was addressed to, or what object was referred to, and so on. The other version also took these parameters into account.

Game time: We also calculated how long each play session took.

RESULTS

Cronbach's alpha values for our survey are given in Table 10. The results indicate that our survey instrument is reliable and precise, with alpha values ranging from 0.849 to 0.886.

Table 10. Cronbach's alpha values for our survey instrument. The values show that our survey instrument is reliable and consistent.

Reliability Statistics

	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items	Number of samples included
Sentence-selection	.886	.887	12	48
NLU	.849	.853	12	46

The results from our survey indicate a significant difference in favor of the sentence selection interface in six of the items. The results are given in Table 11.[2]

Table 11. Significant differences between the sentence-selection and NLU version. Labels in the first row refer to version followed by the item number.

	nlu2 - ss2	nlu3 - ss3	nlu4 - ss4	nlu8 - ss8	nlu9 - ss9	nlu10 - ss10
Z	-2.205[b]	-3.679[b]	-3.047[b]	-5.665[b]	-2.251[b]	-2.690[b]
Asymp. Sig. (2-tailed)	.027	.000	.002	.000	.024	.007
Exact Sig. (2-tailed)	.027	.000	.002	.000	.023	.006
Exact Sig. (1-tailed)	.013	.000	.001	.000	.011	.003
Point Probability	.000	.000	.000	.000	.001	.000

Here's a breakdown of those 6 items:

[2] We also conducted power analysis for insignificant results. See Appendix D.

Q2. Once I decided what to say, it was easy to communicate that to the characters using this interface.

Our results indicate a strong preference for the sentence-selection interface for this question. Although the sentence-selection version actively limits what the players can say, since it doesn't suffer from interpretation problems the participants found that communicating with the characters was made easier. It should be noted that although we expected otherwise, we didn't find evidence pointing to a significant difference in the previous question, "It was easy to decide what I want to say", between the different versions. This might indicate that the participants didn't feel particularly bothered by the limited options that the menu-based sentence-selection version offered them, and we were still able to invoke a sense of agency through the well-placed and well-written options in the menus, and through better communicating to the player when their input will have meaningful impact. Even though they could say whatever they wanted in the NLU version, that didn't feel particularly liberating and empowering on deciding what to say. That was echoed in our earlier qualitative study as well – the limited nature of sentence-selection interfaces might result in players getting a better sense of what actions are available to them in the game, what their status is with respect to the world created by the game, and what their relationship to the characters is. In contrast, as mentioned in the previous chapters, the freedom offered by the NLU version might result in

players feeling lost or like they are *"taking a shot in the dark"* when engaging in conversation with the characters.

Q3. The responses of the characters to my dialogue actions made sense to me.

Participants reported that the NPCs' reactions to player input made more sense in the sentence-selection version than in the NLU version. This finding confirms that while players can contextualize NLU system's interpretation failures within the game world as noted in earlier studies, these failures still hurt players' sense of immersion and perception of how the game world *"makes sense"*. In our previous studies, more constrained forms of dialogue were surprisingly found to offer more story involvement. Coupled with this finding, players might have felt a greater sense of story involvement as a result of both the interface guiding players more strictly on when their input would have meaningful impact and the fact that the characters' responses made more sense – in another words, the system was more accurate.

Q4. I felt in control of the conversation.

Players are likely to experience a greater sense of agency when they are able to figure out how to operate the underlying computational system efficiently. Operating the rules, the cranks and the dials of this machine result in a greater sense of agency when players are able to figure out how to do so, learn how to do it efficiently and consistently. In our previous studies, participants often mentioned how menu-based versions or the versions that pause for player input offer a greater sense of control,

and result in a more structured, polished mode of conversation. Combined with the results from this study, this suggests that players are likely to feel a stronger sense of agency if the game offers them more guidance on when their input would have a meaningful impact and what that impact would be. The versions that waited for player input offered players more time to reason about the rules of the game world and the dramatic affordances of the story, and they learned to how to better and more efficiently operate the game. The result was a game that behaved more expectedly.

Q8. The interface gave me enough time to make decisions on what to say.

As expected, the sentence-selection version was found to be better in this aspect than the NLU version. Mirroring our findings from earlier studies, the NLU version resulted in a rushed, tense experience in which the players struggled to formulate and enter a response in time.

Q9. The interface was easy to learn. & Q10. The interface was easy to use.

The participants felt the sentence selection version was both easier to learn and easier to use than the NLU version. While we expected participants to rate the sentence-selection version as being easier to use, since it allows players to take all the time they want to enter a response, it's surprising that it was also found to be easier to learn, as both interfaces in our study are quite easy to learn and operate on surface level. This finding points to important insight into how players approach interfaces

in games: Beyond being just a way to enter their input into the game, when using an interface players also reason about how that interface helps them operate the underlying computational model. In the sentence-selection version, it's clear that each line of dialogue is associated with a particular outcome that the game knows how to respond to. On the other hand, the players had difficulty figuring out how to operate the system in accordance with their goals, as the inner workings of the system are not as clear and transparent. In our previous studies, during the interviews, we noticed participants often remarking about their deductions on how the NLU system works. They continuously tried to figure out how the system works so that they can operate it efficiently, and once they did they developed various strategies such as entering shorter responses that they felt are among the keywords the game responds to, avoiding typos and slang terms, or entering a response ahead of time to make sure they can exert control over the game. This might be a possible indication that in games, "learning" how to use an interface is a more complex phenomenon that is more intertwined with how the interface exposes and teaches the underlying game mechanics and computational model to the player and how it affects player's understanding of the system.

Player behavior metrics

As part of our study, we also analyzed gameplay logs and collected various metrics related to player behavior. In this section, we look at how those metrics change depending on the interfaces in our study.

Unique discourse acts discovered

As part of our study, we also collected the number of unique discourse acts that the players were able to discover during their entire game session. The results are given in Table 12. Figure 42 and Figure 43 are plots of how many unique discourse acts players were able to discover using both versions.

Table 12. Significant differences in gameplay metrics between sentence-selection and NLU versions.

Metric	N	Mean	Std. Deviation	Minimum	Maximum
NumUniqueDiscActNLU	46	15.5000	4.08112	4.00	25.00
NumUniqueDiscActWithParamsNLU	46	33.7609	10.70241	4.00	51.00
NumUniqueDiscActSS	46	12.3261	1.56424	10.00	15.00
NumUniqueDiscActWithParamSS	46	18.8696	2.86441	12.00	24.00

	NumUniqueDiscActSS - NumUniqueDiscActNLU
Z	-4.092[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

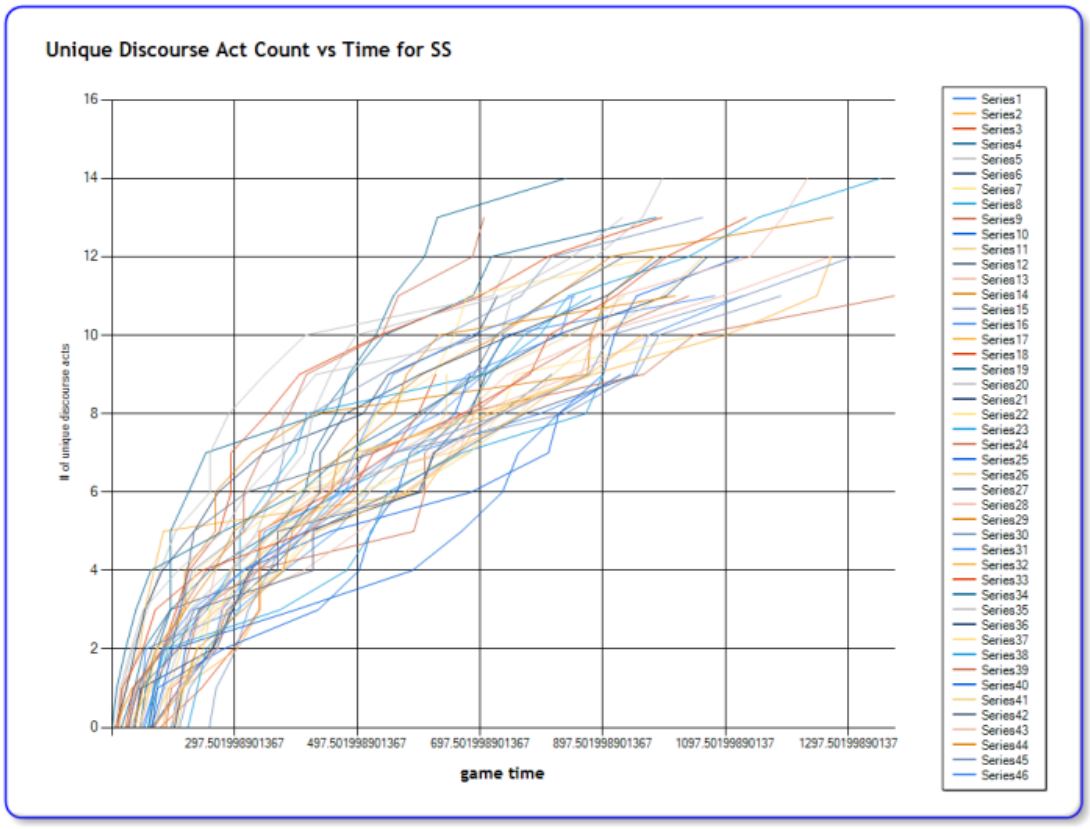


Figure 42. Number of unique discourse acts vs time for the sentence-selection version.

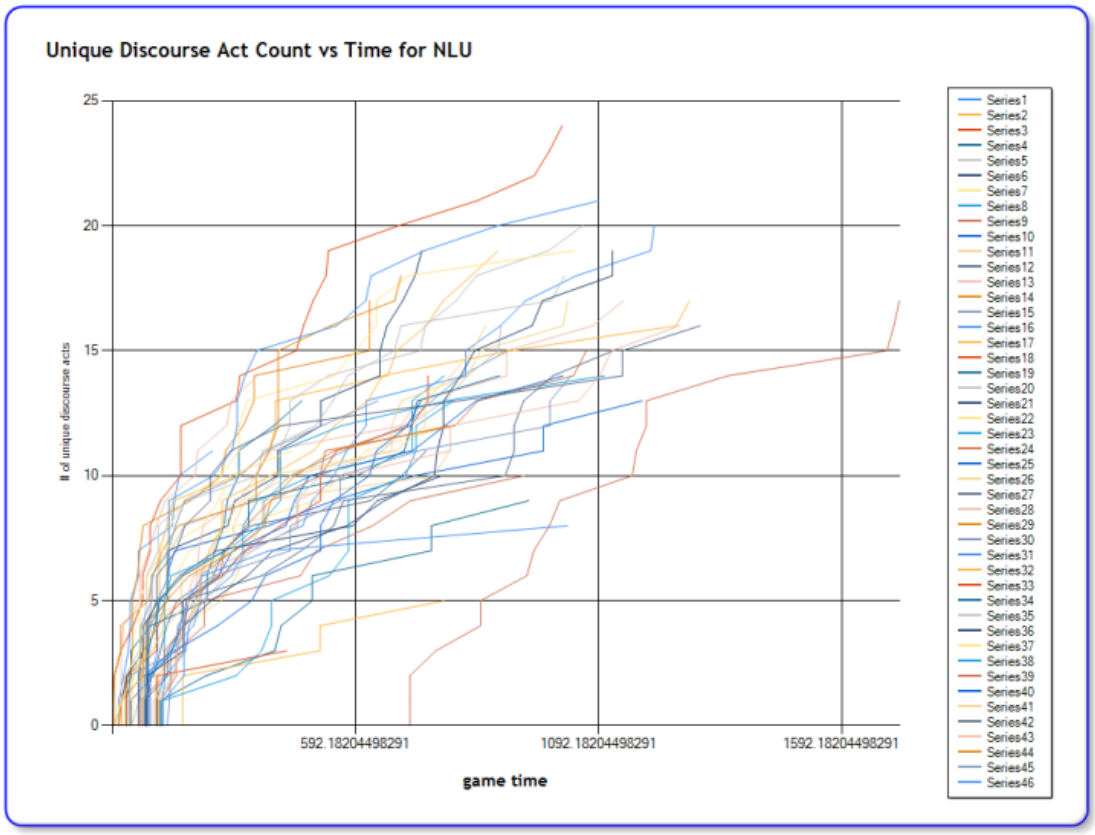


Figure 43. Number of unique discourse acts vs time for the NLU version.

Input time

Our results indicate that players took significantly less time to enter input in the NLU version (Table 13). This fact was echoed in our earlier studies as well, with the NLU version being reported as the more tense and rushed version.

Table 13. Significant differences in time to enter input between the sentence-selection and NLU versions.

Ranks

		N	Mean Rank	Sum of Ranks
avgTimeInputNLU - avgTimeInputSS	Negative Ranks	41[a]	23.41	960.00
	Positive Ranks	5[b]	24.20	121.00
	Ties	0[c]		
	Total	46		

a. avgTimeInputNLU < avgTimeInputSS
b. avgTimeInputNLU > avgTimeInputSS
c. avgTimeInputNLU = avgTimeInputSS

Test Statistics[a]

	avgTimeInputNLU - avgTimeInputSS
Z	-4.583[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

Ending & revelations

In order to examine whether it was easier to get a better ending in one of the versions versus the other, we developed a scoring scheme for the endings.

According to this scheme, participants received a score of zero for the ending if they were kicked out, a score of one if Trip and Grace decided to stay together, but didn't make any revelations, a score of two if either Trip or Grace (but not both) makes a number of revelations and one of them decides to leave, and a score of three for the best ending where Trip and Grace both reveal secrets but they decide to stay together and work on saving their marriage. We also calculated the total number of revelations from each gameplay log.

Although we didn't find statistically significant evidence that players were able to reach a more satisfying ending using one version versus the other, possibly due to the size of our sample pool, our results certainly point towards a trend: Participants often uncovered more revelations and better endings in the sentence-selection version, with a significance value of 0.06. The difference became more pronounced when we calculated a cumulative score as the sum of ending score and number of revelations. This finding coincides with participants experiencing a higher sense of control using the sentence-selection version, as reported in our previous experiment.

Number of interactions

Our results show important significant differences on how participants chose to interact with Façade using different dialogue systems (Table 14). Participants interacted significantly more with the characters using the sentence-selection version. This difference however mostly stemmed from a higher number of gesture interactions using the sentence-selection version than in the NLU version.

Table 14. Differences in number of interactions and number of gesture interactions between the sentence-selection and NLU versions.

	noInteractionsSS - noInteractionsNLU	**noGestureInteractionsSS - noGestureInteractionsNLU**
Z	-2.322[b]	-3.409[b]
Asymp. Sig. (2-tailed)	.020	.001
Exact Sig. (2-tailed)	.019	.000
Exact Sig. (1-tailed)	.010	.000
Point Probability	.000	.000

This is definitely an interesting result. The most probable reason is that due to the limited interaction opportunities offered by the sentence-selection version, participants spent the time in-between the dialogue prompts trying to express themselves through the gesture system. However, we didn't find statistical evidence to suggest that there is a significant difference between how many utterances players typed or how often they typed them. This might suggest that both interfaces offered a similar experience in terms of pacing – at least, the different wasn't as pronounced. This might point to the need for other explanations on why the sentence-selection interface prompted a much higher One of the possible reasons might be that participants are seeking additional support from the gesture system to communicate with the characters due to the limited number of dialogue interaction opportunities offered by the sentence-selection version. In our previous studies, we also found that the NLU system maximized flow and presence. This might indicate that the players are more likely to enact physical actions with the characters when using interfaces

that limit those aspects of gameplay. It should obviously be noted that these are only hypotheses that need to be tested in future experiments.

Game time and other metrics

Table 15. Differences in number of utterances by Trip and game time between the sentence-selection and NLU versions.

	noTripUtterancesSS - noTripUtterancesNLU	**gameTimeSS - gameTimeNLU**
Z	-3.480[b]	-5.151[b]
Asymp. Sig. (2-tailed)	.001	.000
Exact Sig. (2-tailed)	.000	.000
Exact Sig. (1-tailed)	.000	.000
Point Probability	.000	.000

We found participants had significantly longer game sessions using the sentence-selection version (Table 15). Number of lines of dialogue spoken by Trip was also significantly higher in the sentence-selection version.

Observed correlations

We also calculated Spearman's rank order correlation coefficient for the metrics in our dataset[3]. The following is a very brief summary of the significant correlations we found for different versions that we believe to be important.

[3] Spearman's rank order correlation coefficient is a correlation test for non-parametric data. See (Spearman 1987) for details.

Sentence-selection

For the sentence selection version, we found positive correlations between the first item in our survey, "It was easy to decide what I want to say using this interface" and the number of revelations, number of unique discourse acts addressed by the player and ending score. This demonstrates that the ease of use provided by the interface actually translates to, or is affected by in-game success. Ease of use, when it comes to game interfaces, doesn't just concern operating the interface easily on its surface, but operating the underlying computational model easily as well. This provides further proof for our more nuanced definition of agency.

There was a significant negative correlation between the total number of interactions and the number of dialogue interactions. This finding is consistent with players' complaints that the pausing nature of menu-based interfaces might chunk the experience in a way that might lead players to express themselves in other ways.

There was also a significant positive correlation between the item "The ending made sense to me" and the ending score, showing that players actually wanted to reach the better endings and found them more reasonable.

NLU version

There was a significant negative correlation between the score for the item "Once I decided what to say, it was easy to communicate that to the characters using this interface." and the number of deflects. Deflects in the NLU version occur when the system fails to understand a player utterance, but doesn't want to disrupt the

experience in an unpleasant way by attempting a repair by using typical repair utterances (i.e. "I didn't understand you") or explicit feedback. Additionally, ending score was significantly positively correlated with the score for the item "The responses of the characters to my dialogue actions made sense to me." Once again, in both those items, players correlated and evaluated ease of use with respect to understanding how system mechanics deeper than the surface level work.

For item 4, "I felt in control of the conversation", there was a significant positive correlation between the ending score and item score. Since correlation doesn't imply causality, a different experiment needs to be run to see if the participants perceived they had more control because they got a good ending, or they got a good ending because the interface actually allowed them more control[4]. Nonetheless, the result is interesting in that it shows sense of control is actually reflected in or affected by the ending score. The score for this item was also significantly positively correlated with the number of unique discourse acts that the player was able to discover. This was also an interesting result since the sense of control actually increased as players were able to employ a higher *range* of discourse acts – the perceived control was enforced by seeing not only correct responses, but also a higher variety of reactions from the NPCs.

[4] An experiment can be designed such that one group consistently receives good endings regardless of their interactions whereas the other group receives the endings that actually resulted from their interactions, but this was beyond the scope of this dissertation.

Another interesting result was the negative correlation between the average time it took the player to enter input versus the score for the item "The ending made sense to me." Average time to enter input was significantly negatively correlated with items "The interface was easy to learn." and "The interface was easy to use.". These findings correlate with the findings from our previous studies in which the players frequently mentioned the time-sensitive nature of a realistic conversation model frustrating.

Regression Analysis

Finally, in order to determine the contributions of each metric to our usability scores, we followed the procedure outlined in the original PARADISE paper (M. A. Walker, Litman, Kamm, and Abella 1997) and performed multiple linear regression, using scores from our usability surveys as the dependent variables, and our metrics as independent variables. Each metric was normalized to avoid problems with different scales used when calculating the metrics.

Using this procedure, we were able to construct a model for the first and fifth items for the sentence selection version, and the mean score from all items. It should be noted that while we had many significant models with high predictive power, most of the time none of the variables significantly predicted the variance in the item score on its own, possibly due to our small sample size. We don't include such cases here, but they are included in the appendices. The numerical details of the models are given in Tables 16 - 20 in the following pages.

Table 16. Regression model for the item "It was easy to decide what I want to say using this interface."

Equation	Sig.	Predictor Sig.
SS1 = 5.492 +	0.01	
10.561*NumInteractions		0.02
-11.121*NumGestureInteractions		0.02
-2.210*numDialogueInteractions		0.03

For the first item, while the total number of interactions had a positive weight, surprisingly the number of dialogue interactions had a negative weight. This might be due to the fact that our menu implementations included interaction points too frequently. The number of gesture interactions had a negative weight as expected, since players are more likely to resort to using gestures if they can't make use of the dialogue system.

Table 17. Regression model for the item "The choices I wanted to make were present in the interface."

Equation	Sig.	Predictor Sig.
SS5 = 4.583 +	0.02	
+0.729*endingScore		0.010
-3.572*FrequencyOfInteractions		0.006
-3.062*FrequencyOfGestureInteractions		0.015
-0.353*avgTimeBetweenInteractions		0.015
-0.651*NumGraceAffinityChanges		0.007
+0.587*FrequencyOfUniqueDiscourseActs		0.026

For item 5, ending score has a positive weight, as expected. Coupled with the finding above, the fact that the frequency of interactions and frequency of gesture interactions were both calculated as costs is not surprising either – it's more likely

that players won't be able to find the choices they wanted to make in every menu that pops up. Average time between interactions was also calculated as a cost. Unsurprisingly, frequency of unique discourse acts was a contributing factor to higher scores.

Table 18. Regression model for the mean usability score for sentence-selection version.

Model Summary

Model	R	R2	Adjusted R2	Std. Error of the Estimate
1	.842[a]	.709	.376	.7964419

a. Predictors: (Constant), NumUniqueDiscActNoParamFreqSS, noDeflectsFreqSS, noNLUDontUnderstandSS, avgTimeBwInteractionsSS, noTripAffSS, noGraceAffSS, noGraceUtterancesFreqSS, noPCUtterancesSS, noTripUtterancesFreqSS, NumUniqueDiscActSS, noInteractionsSS, EndingScoreAdjustedSS, noRevelationsSS, avgTimeInputSS, noPCUtterancesFreqSS, noDeflectsSS, NumUniqueDiscActFreqSS, gameTimeSecsSS, noGraceUtterancesSS, noTripUtterancesSS, noNLUDontUnderstandFreqSS, endingScoreSS, noGestureInteractionsFreqSS, NumUniqueDiscActNoParamSS

Table 19. Statistics for the regression model for the mean usability score for sentence-selection version.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	32.394	24	1.350	2.128	.042[b]
	Residual	13.321	21	.634		
	Total	45.715	45			

a. Dependent Variable: MeanSS

b. Predictors: (Constant), NumUniqueDiscActNoParamFreqSS, noDeflectsFreqSS, noNLUDontUnderstandSS, avgTimeBwInteractionsSS, noTripAffSS, noGraceAffSS, noGraceUtterancesFreqSS, noPCUtterancesSS, noTripUtterancesFreqSS, NumUniqueDiscActSS, noInteractionsSS, EndingScoreAdjustedSS, noRevelationsSS, avgTimeInputSS, noPCUtterancesFreqSS, noDeflectsSS, NumUniqueDiscActFreqSS, gameTimeSecsSS, noGraceUtterancesSS, noTripUtterancesSS, noNLUDontUnderstandFreqSS, endingScoreSS, noGestureInteractionsFreqSS, NumUniqueDiscActNoParamSS

While our model significantly captured the variances in mean user satisfaction score, no variable significantly captured the difference by itself except number of deflects and frequencies of deflects. In the sentence-selection version, we made use of deflect reactions to create comedic opportunities where the player character would say something weird or out of place and the characters would react to the player awkwardly, and ignore what was said as if not to embarrass the player. We detected an interesting interplay between the coefficients of these variables: While participants enjoyed getting a deflect reaction every once in a while, overall they didn't like experiencing too many of them.

Table 20. Coefficients for the metrics for the model.

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	5.418	.117		46.143	.000
noRevelationsSS	.033	.845	.033	.039	.969
EndingScoreAdjustedSS	.221	2.157	.219	.102	.919
endingScoreSS	-.142	2.718	-.141	-.052	.959
noDeflectsSS	-3.326	1.505	-3.300	-2.210	.038
noDeflectsFreqSS	4.404	1.852	4.370	2.378	.027
noNLUDontUnderstandSS	3.384	1.932	3.358	1.751	.094
noNLUDontUnderstandFreqSS	-3.811	2.320	-3.781	-1.642	.115
noInteractionsSS	2.215	2.467	2.198	.898	.379
noGestureInteractionsFreqSS	-2.621	2.835	-2.601	-.925	.366
avgTimeBwInteractionsSS	-.065	.210	-.065	-.310	.760
avgTimeInputSS	-.117	.401	-.116	-.291	.774

noTripAffSS	.064	.188	.064	.341	.737
noGraceAffSS	-.190	.185	-.188	-1.026	.316
noTripUtterancesSS	-3.219	2.010	-3.193	-1.601	.124
noTripUtterancesFreqSS	2.317	1.518	2.299	1.526	.142
noGraceUtterancesSS	-1.035	2.333	-1.027	-.444	.662
noGraceUtterancesFreqSS	1.365	2.087	1.354	.654	.520
noPCUtterancesSS	-2.507	2.064	-2.487	-1.215	.238
noPCUtterancesFreqSS	1.434	1.612	1.423	.890	.384
gameTimeSecsSS	5.191	2.925	5.150	1.775	.090
NumUniqueDiscActSS	-2.026	2.885	-2.010	-.702	.490
NumUniqueDiscActFreqSS	3.185	3.751	3.160	.849	.405
NumUniqueDiscActNoParamSS	3.844	3.123	3.814	1.231	.232
NumUniqueDiscActNoParamFreqSS	-4.102	3.326	-4.070	-1.233	.231

We weren't however able to construct a significant model for the NLU version. When selecting our metrics, we tried to choose metrics that generalize to a wide spectrum of dialogue systems found in games. While it's quite possible that the main reason for not being able to construct a linear model was the small size of our sample pool, another possibility is that user satisfaction in NLU systems is affected by metrics that are very specific to those systems that we haven't been able to determine.

8

CONCLUSION

In this dissertation, I have explored the complicated design space of dialogue systems in games in depth. I believe the results I presented provide important insights not only into the design issues related to those systems, but also into important aesthetic and experiential properties of games and how those effect player perception. In the following sections, I will summarize our contributions, and discuss future directions for this work. First we will go over the different design trade-offs our studies uncovered between the systems we considered. Then we summarize our results related to aspects that are more intertwined with how people experience and participate in games. We believe our process is also important as future guidelines for researchers interested in doing evaluation work in this domain – so our insights from the process will also be discussed. Finally, I will present ideas for future work.

CONTRIBUTIONS

Dialogue system design

Our results lead to several interesting conclusions. It might be tempting to assume that interfaces that are easy to use and feel as transparent as possible to the player will result in the highest degree of engagement. However, in our first study, we found that despite the problems players reported using the original NLU interface in both studies, they found the NLU version to be the most engaging when compared to the sentence-selection and abstract-response versions. Players were impressed by the level of freedom offered by that interface, and enjoyed the interface greatly when it worked. It also required constant attention so that players can figure out when and how to interject and get their word in, whereas engagement in menu-based interfaces seemed to stem more from deeper dramatic involvement with the situation. The sentence-selection version, by giving more control over what the player character will say, provided a better association with the player character, whereas the abstract-response version offered stronger control over the outcomes of the actions and resulted in a higher sense of success. The results of our pacing study, on the other hand, showed that players preferred the reactive-pause version of the NLU interface the most – that version might have resulted in the most-balanced version in terms of pacing and freedom, as players can still type in anytime they want, but they have all the time they want to formulate their responses as well. Overall, our results show that engagement can be achieved in different ways, resulting in different types of experiences.

Another design guideline that our studies seem to refute is the belief that if an interface gives players a lot of freedom, they will feel a high sense of control. In our first study, we found that players reported feeling the strongest sense of control using the abstract-response version. In the following study on pacing options, players reported feeling the highest sense of control in the prompt-pause version, surpassing both reactive-pause and original versions, even though the prompt-pause version was the most limiting in terms of when players can interact with the game, in effect taking some of the control away from the players. In both cases, the original NLU version was reported as offering the least sense of control. Surprisingly, in both studies, players reported feeling a higher sense of control when the interface actively limited them. This points to another important insight: Players experience a higher sense of control when the interface actively guides them on how to impact the underlying computational model.

Another naïve assumption in game design is that realism is the ultimate end goal if we want to maximize flow and presence. Having more control over the player character made players feel more present as themselves in the game world, and they felt more responsibility and more bound by social norms and conventions. The abstract-response version was found to be too unnatural as a model of conversation; the mismatch between intent and outcome was jarring. Players also reported that the prompt-pause version chunked the experience in unappealing ways.

However, the realism offered by the NLU version didn't translate to more story involvement. In the first study, we found the sentence-selection version offered the highest story involvement. Participants liked exploring the story through the interesting and tempting options in the menus. They felt that the player character, authored in part for them by us, was better situated within the game plot. As a result they cared more for Trip and Grace and felt a higher sense of purpose in contrast to the blank canvas offered by the original version. Fully realized options in the sentence-selection version also made them feel as if there was more nuance in the available options in the menus, despite the fact that sentence-selection and abstract-response versions were identical except the option texts. In the pacing study, we found that the prompt-pause version, which was reported to be the least realistic of all versions, offered the highest degree of story involvement, as the prompts indicated a good time for the player to respond to the characters, and they could use these cues to make more meaningful impact in the game. The prompts structured the experience in a way that gave players more story exploration power, as they knew when their input would have meaningful impact. Taken together, the results from our two studies show that players feel more story involvement when their interaction is more structured.

Overall, our results show that different interfaces have different affordances and they maximize different aspects of gameplay. Sentence-selection interfaces, through well-written options and by presenting a clearer path through the experience on

when players are more likely to have meaningful impact, evoke in participants a high sense of story involvement. Abstract-response interfaces give players a high degree of control, as players not only know when their input is going to have meaningful impact, but also have a better idea of what the outcome of their dialogue actions will be. NLU interfaces maximize presence, engagement and flow, but have been found to be more difficult to use than other versions, with participants reporting difficulties with interpretation problems, pacing and a loss of agency due to the extreme freedom they allow. While it would be tempting to conclude that it's the realism offered by the NLU system that maximized flow, presence and engagement, in a following study we were able to test this assumption by implementing artificial prompts and pauses in Façade's NLU system. Surprisingly, we have found that participants felt a higher sense of control and higher agency when we actively limited when they can interact with the game in the prompt-pause version. However, this version resulted in an experience that hurt immersion and presence for some participants. The reactive-pause version offered a good compromise in terms of control between the NLU version and the prompt-pause version, but it was still less realistic than the NLU version. The NLU version offered the highest degree of realism and freedom, but this resulted in an experience that felt too tense and rushed for some participants – which might still be a desirable goal for some designers. Combining pausing and prompting features of menu-based interfaces with NLU might be a promising future direction for dialogue systems. Perhaps most importantly, we have shown that dialogue interfaces significantly

impact gameplay experience beyond the surface level, and play a significant role in how players come to understand the underlying computational model.

Game design

Our findings reveal important design trade-offs on the surface level, but they also point to deeper insights into game design: Interestingly, we found that the most *"realistic"* or *"natural"* versions were less appealing to most players than versions that behaved in *unnatural* ways from dimensions ranging from engagement to sense of control and story involvement. This indicates that the Holodeck dream for interactive fiction, the belief that interfaces should strive to be as realistic as possible, is not the ultimate end-goal for every type of experience: Designers should instead consider what types of explicit mediation to introduce to shape player experience.

Another important insight we gained from our studies points to further evidence that agency is a complex phenomenon that is related to how players come to understand and rationalize the inner workings of the underlying computational system, learn what actions it would be significant to take and when. This was echoed both in our study in which we compared different pacing options, as participants experienced a higher sense of control when we actively limited when they can take action, and in our last study in which we compared the sentence-selection version with the NLU version, with participants reporting the sentence-selection version as the easier interface to learn and use. These findings suggest that players do not simply try to learn how to operate an interface – they also try to continuously figure

out how to operate the underlying computational model as reflected by the interface. These observations point to the need to move away from a naïve interpretation of agency as free will, i.e. letting players do whatever they want whenever they want, and more toward a formulation that takes into account this process of *learning* and *manipulation:* agency involves creating desires and understandable opportunities to act in a manner (and at a time) that will impact the underlying computational system.

Evaluating dialogue systems

The work presented in this dissertation also resulted in the development of a survey instrument, presented in Chapter 7, which can be used to test the usability of a variety of dialogue systems found in games in general. This survey is statistically reliable and precise. I believe it should be useful to researchers willing to work in this domain.

I believe this complicated intersecting domain of dialogue systems and story-based experiences presents many challenges in user evaluation – therefore we attempted to attack the problem in several fronts. Our first two studies were mostly qualitative and exploratory, and focused on the effects of changes in interfaces on player perception. These studies allowed us to question and test the assumptions of collective wisdom on the subject, and move toward a more quantitative approach in a more informed manner. The visualization tool we developed was a major stepping-stone in this transition – I believe the two case studies we presented present

convincing evidence towards the usefulness of information visualization techniques when studying this domain.

Finally, in our last study, we proposed a set of metrics for evaluating dialogue systems in games. I believe most of those metrics are applicable to different dialogue systems and games as they are, and the ones that aren't are easily modifiable to apply to those systems as well. We showed that differences in player behavior do really express themselves in quantifiable ways in those metrics, sometimes in very significant degrees, showing that those metrics are indeed relevant to user satisfaction from a dialogue system in a game. Finally, through regression methods, we showed how to calculate relative contributions of those metrics to scores for different survey items, hopefully providing some guidance to designers and researchers on how to carry out evaluation work in this complicated domain that will also give them insight into how to design to influence certain aspects of player experience.

FUTURE WORK

While I believe the contributions discussed above are still significant, they only represent a first foray into understanding this complicated design space. First and foremost, there are still many more possible studies that can be carried out. When we initially started this work, in our very first meetings, we brainstormed about seven to ten different possible interfaces that emphasize different experiential aspects such as pacing, input modality, mixed modality or mixed pacing options, and so on. While

we settled in the end on interfaces that we believed are most relevant to current state of the domain in games, testing these interfaces in controlled studies are likely to result in deeper insights and further refine our results. I'd personally be interested in testing mixed or reactive pacing options that pause the game depending on how long a player takes to enter input, or mixed modality systems in which players can both type and select from a menu. It remains to be seen how the trade-offs we discovered for those interfaces combine and change player experience.

Another thread I'd like to explore further is changes in player perception depending on player demographic. Unfortunately we didn't have the resources to recruit participants from different demographics due to our limited participant pool, but it remains to be seen how factors such as gender differences, familiarity with certain interfaces, or genre preferences affect player choices for those interfaces.

It should also be noted that the specifics of Façade might have influenced our results. Its' tense situation and dramatic structure might not be to everyone's taste. However, it still remains one of the most conversation-oriented games where players can navigate a space and interact with the game in ways familiar to them from the current generation of games. Furthermore, over the span of 20 minutes it presents the players with a complete story arc, with significant conversational interaction that has deep impact on what ending they get. However, it'd still be interesting to run the same experiments with a different game so that we can observe the effects of the specifics of the game on the results.

The visualization tool we developed also presents promising opportunities for future work. Further case studies are likely to reveal the need for different visualizations that might be useful for studying story-based experiences, as well as further modifications to existing visualizations. I believe the game design community should benefit from a public release of this tool in the future.

Finally, further studies should be conducted to determine the applicability of our metrics and our survey instrument to other dialogue systems. While they seemed to capture some aspects of user satisfaction in our study, it'd be interesting to see which metrics generalize across different systems, which need to be modified, and which don't generalize at all. While I believe the complicated nature of this domain makes it incredibly difficult to come up with strong predictive models, further regression algorithms can also be applied to the data to get a better fit with more predictive power.

REFERENCES

Adams, Douglas. 1998. *Starship Titanic*. The Digital Village.

Adams, Ernest. 2005. "You Must Play Façade, Now!" http://www.designersnotebook.com/Columns/075_You_Must_Play_Facade_Now/075_you_must_play_facade_now.htm.

Andersen, Erik, Yun-En Liu, Ethan Apter, François Boucher-Genesse, and Zoran Popović. 2010. "Gameplay Analysis Through State Projection." In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, 1–8. FDG '10. New York, NY, USA: ACM. doi:10.1145/1822348.1822349. http://doi.acm.org/10.1145/1822348.1822349.

Beringer, Nicole, Ute Kartal, Katerina Louka, Florian Schiel, Uli Türk, Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, and Uli Türk. 2002. "PROMISE - A Procedure for Multimodal Interactive System Evaluation." http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.2767.

Bioware. 2003. *Star Wars: Knights of the Old Republic*. Bioware.

———. 2007. *Mass Effect (series)*. Bioware.

———. 2010. *Dragon Age: Origins*. Bioware.

Bowman, B., N. Elmqvist, and T.J. Jankun-Kelly. 2012. "Toward Visualization for Games: Theory, Design Space, and Patterns." *IEEE Transactions on Visualization and Computer Graphics* 18 (11) (November): 1956 –1968. doi:10.1109/TVCG.2012.77.

Brown, Emily, and Paul Cairns. 2004. "A Grounded Investigation of Game Immersion." In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 1297–1300. CHI EA '04. New York, NY, USA: ACM. doi:10.1145/985921.986048. http://doi.acm.org/10.1145/985921.986048.

Bruls, Mark, Kees Huizing, and Jarke van Wijk. 1999. "Squarified Treemaps." In *In Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization*, 33–42. Press.

Cater, John, Rob Dubbin, Eric Eve, Elizabeth Heller, Jayzee, Kazuki Mishima, Sarah Morayati, et al. 2009. *Alabaster*.

Chittaro, L., R. Ranon, and L. Ieronutti. 2006. "VU-Flow: A Visualization Tool for Analyzing Navigation in Virtual Environments." *IEEE Transactions on Visualization and Computer Graphics* 12 (6) (December): 1475 –1485. doi:10.1109/TVCG.2006.109.

Christian Swineheart. "One Book, Many Readings." http://samizdat.cc/cyoa/.

Clive Thompson. 2007. "Halo 3: How Microsoft Labs Invented a New Science of Play." *WIRED*. http://www.wired.com/gaming/virtualworlds/magazine/15-09/ff_halo?currentPage=all.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge Academic.

Crawford, Chris. 2008. "Deikto: An Application of the Weak Sapir-whorf Hypothesis." In *Proceedings of the Hypertext 2008 Workshop on Creating Out of the Machine: Hypertext, Hypermedia, and Web Artists Explore the Craft*, 1–4. Creating '08. New York, NY, USA: ACM. doi:10.1145/1379153.1379155. http://doi.acm.org/10.1145/1379153.1379155.

———. *Storytron*. www.storytron.com.

Cronbach, Lee. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3): 297–334.

Crowther, Will. 1975. *Adventure*.

Csíkszentmihályi, Mihály. 2008. *Flow: The Psychology of Optimal Experience*. HarperCollins.

Danieli, Morena, and Elisabetta Gerbino. 1995. "Metrics for Evaluating Dialogue Strategies in a Spoken Language System." In *In Proc of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*. Vol. 16. American Association for Artificial Intelligence. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Metrics+for+evaluating+dialogue+strategies+in+a+spoken+language+system#0.

David Levy. *Do-Much-More*. Intelligent Toys Ltd. http://www.worldsbestchatbot.com/.

Dow, Steven, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. "Presence and Engagement in an Interactive Drama." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1475–1484. CHI

'07. New York, NY, USA: ACM. doi:10.1145/1240624.1240847. http://doi.acm.org/10.1145/1240624.1240847.

Drachen, Anders, and Alessandro Canossa. 2009. "Towards Gameplay Analysis via Gameplay Metrics." In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, 202–209. MindTrek '09. New York, NY, USA: ACM. doi:10.1145/1621841.1621878. http://doi.acm.org/10.1145/1621841.1621878.

Ellison, Brent. 2008. "Defining Dialogue Systems." *Gamasutra*. http://www.gamasutra.com/view/feature/3719/defining_dialogue_systems.php.

Emily Short. 2009. "'Homer In Silicon': Sub-Façade." *GameSetWatch*. http://www.gamesetwatch.com/2009/05/column_homer_in_silicon_the_co_1.php.

Fekete, Jean-Daniel. 2004. "The InfoVis Toolkit." In *Proceedings of the IEEE Symposium on Information Visualization*, 167–174. INFOVIS '04. Washington, DC, USA: IEEE Computer Society. doi:10.1109/INFOVIS.2004.64. http://dx.doi.org/10.1109/INFOVIS.2004.64.

Fernanda B. Viégas, and Martin Wattenberg. 2003. "History Flow - How It Works." http://www.research.ibm.com/visual/projects/history_flow/explanation.htm.

French, Robert M. 2000. "The Turing Test: The First 50 Years." *Trends in Cognitive Sciences* 4 (3) (March 1): 115–122. doi:10.1016/S1364-6613(00)01453-4.

Friendly, Michael. 2009. "The History of the Cluster Heat Map." *The American Statistician*.

Georg Zoeller. "Game Development Telemetry in Video Games Projects" presented at the Game Developers Conference, 2010. http://gdc.gulbsoft.org/talk.

Gorsuch, Richard L. 1983. *Factor Analysis*. Psychology Press.

Graph#. 2009. graphsharp.codeplex.com.

Holmquist, Lars Erik, and Tobias Skog. 2003. "Informative Art: Information Visualization in Everyday Environments." In *Proceedings of the 1st International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*, 229–235. GRAPHITE '03. New York, NY,

USA: ACM. doi:10.1145/604471.604516. http://doi.acm.org/10.1145/604471.604516.

Hoobler, Nate, Greg Humphreys, and Maneesh Agrawala. 2004. "Visualizing Competitive Behaviors in Multi-User Virtual Environments." In *Proceedings of the Conference on Visualization '04*, 163–170. VIS '04. Washington, DC, USA: IEEE Computer Society. doi:10.1109/VISUAL.2004.120. http://dx.doi.org/10.1109/VISUAL.2004.120.

Infocom. 1980. *Zork (series)*. Infocom.

Inselberg, Alfred. 2009. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer.

Johnson, Brian, and Ben Shneiderman. 1991. "Tree-Maps: a Space-filling Approach to the Visualization of Hierarchical Information Structures." In *Proceedings of the 2nd Conference on Visualization '91*, 284–291. VIS '91. Los Alamitos, CA, USA: IEEE Computer Society Press. http://dl.acm.org/citation.cfm?id=949607.949654.

Josh McCoy, Mike Treanor, Ben Samuel, and Aaron A. Reed. 2012. *Prom Week*. promweek.soe.ucsc.edu.

Kim, Jun H., Daniel V. Gunn, Eric Schuh, Bruce Phillips, Randy J. Pagulayan, and Dennis Wixon. 2008. "Tracking Real-time User Experience (TRUE): a Comprehensive Instrumentation Solution for Complex Systems." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 443–452. CHI '08. New York, NY, USA: ACM. doi:10.1145/1357054.1357126. http://doi.acm.org/10.1145/1357054.1357126.

Kim, Taeyong, and Frank Biocca. 1997. "Telepresence via Television: Two Dimensions of Telepresence May Have Different Connections to Memory and Persuasion.[1]." *Journal of Computer-Mediated Communication* 3 (2). http://dx.doi.org/10.1111/j.1083-6101.1997.tb00073.x.

Knickmeyer, Rachel Lee, and Michael Mateas. 2005. "Preliminary Evaluation of the Interactive Drama *Facade*." In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1549–1552. CHI EA '05. New York, NY, USA: ACM. doi:10.1145/1056808.1056963. http://doi.acm.org/10.1145/1056808.1056963.

Lombard, M, and T Ditton. 1997. "AT THE HEART OF IT ALL: THE CONCEPT OF PRESENCE." *Journal of Computer Mediated Communication* 3 (2). http://jcmc.indiana.edu/vol3/issue2/lombard.html.

Lucas Arts. 1990. *Monkey Island (series)*. Lucas Arts.

Martin, Patricia Yancey, and Barry A Turner. 1986. "Grounded Theory and Organizational Research." *The Journal of Applied Behavioral Science* 22 (2) (April 1): 141–157. doi:10.1177/002188638602200207.

Mateas, Michael. 2001. "A Preliminary Poetics for Interactive Drama and Games." In , 51–58.

Mateas, Michael, and Andrew Stern. 2002. "A Behavior Language for Story-Based Believable Agents." *IEEE Intelligent Systems* 17 (4) (July): 39–47. doi:10.1109/MIS.2002.1024751.

Maxis. 2000. *The Sims (series)*. http://thesims.com/en_us/home.

McCoy, Josh, Mike Treanor, Ben Samuel, Brandon Tearse, Michael Mateas, and Noah Wardrip-Fruin. 2010. "*Comme Il Faut 2*: a Fully Realized Model for Socially-oriented Gameplay." In *Proceedings of the Intelligent Narrative Technologies III Workshop*, 10:1–10:8. INT3 '10. New York, NY, USA: ACM. doi:10.1145/1822309.1822319. http://doi.acm.org/10.1145/1822309.1822319.

McTear, Michael. 2004. *Spoken Dialogue Technology: Towards the Conversational User Interface*. Springer.

Medler, Ben. 2012. "Play With Data: An Exploration of Play Analytics and It's Effect on Player Experiences". Georgia Institute of Technology. http://hdl.handle.net/1853/44888.

Mehta, Manish, Steven Dow, Michael Mateas, and Blair MacIntyre. 2007. "Evaluating a Conversation-centered Interactive Drama." In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, 8:1–8:8. AAMAS '07. New York, NY, USA: ACM. doi:10.1145/1329125.1329135. http://doi.acm.org/10.1145/1329125.1329135.

Michael Mateas. 2004. "Games and Natural Language Understanding." http://grandtextauto.org/2004/09/17/games-and-natural-language-understanding/.

Michael Mateas, and Andrew Stern. 2005. *Facade*. www.interactivestory.net.

Mikel Reparaz. 2011. "L.A. Noire Review." http://www.gamesradar.com/la-noire-review-10/.

Milton Friedman. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32 (200) (December 1): 675–701. doi:10.2307/2279372.

Minsky, Marvin. 1980. "Telepresence." *Omni* (June): 45–51.

Murray, Janet Horowitz. 1998. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. MIT Press.

Nunnally, Jum C. 1978. *Psychometric Theory*. McGraw-Hill.

Nutt, Christian. 2007. "Gamasutra - News - AGDC: BioWare Charts Writing For Mass Effect." *Gamasutra*. http://www.gamasutra.com/php-bin/news_index.php?story=15406.

O'Brien, Heather Lynn. 2008. *Defining and Measuring Engagement in User Experiences with Technology*. Dalhousie University (Canada).

Orkin, Jeff, and Deb Roy. 2007. "The Restaurant Game: Learning Social Behavior and Language from Thousands of Players Online." *Journal of Game Development* 3 (1) (December): 39–60.

———. 2009. "Automatic Learning and Generation of Social Behavior from Collective Human Gameplay." In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, 385–392. AAMAS '09. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=1558013.1558065.

Pinar Saygin, Ayse, Ilyas Cicekli, and Varol Akman. 2000. "Turing Test: 50 Years Later." *Minds and Machines* 10 (4): 463–518. doi:10.1023/A:1011288000451.

Pousman, Zachary, Mario Romero, Adam Smith, and Michael Mateas. 2008. "Living with Tableau Machine: a Longitudinal Investigation of a Curious Domestic Intelligence." In *Proceedings of the 10th International Conference on Ubiquitous Computing*, 370–379. UbiComp '08. New York, NY, USA: ACM. doi:10.1145/1409635.1409685. http://doi.acm.org/10.1145/1409635.1409685.

Quantic Dream. 2005. *Indigo Prophecy*. Quantic Dream.

———. 2010. *Heavy Rain*. Quantic Dream.

Reed, Aaron. 2008. *Blue Lacuna*.

Richard Wallace. *A.L.I.C.E. (Artificial Linguistic Internet Computer Entity)*. http://www.alicebot.org/.

Rollo Carpenter. *Jabberwacky*. http://www.jabberwacky.com/.

Romero, Mario, Zachary Pousman, and Michael Mateas. 2008. "Alien Presence in the Home: The Design of Tableau Machine." *Personal Ubiquitous Comput.* 12 (5) (June): 373–382. doi:10.1007/s00779-007-0190-z.

Rouse, Richard, and Steve Ogden. 2005. *Game Design: Theory & Practice*. Wordware Pub.

Salen, Katie, and Eric Zimmerman. 2004. *Rules of Play: Game Design Fundamentals*. MIT Press.

Sheldon, Lee. 2004. *Character Development and Storytelling for Games*. Thomson Course Technology.

Short, Emily. 2006. *Glass*.

———. 2012. "Conversation." *Emily Short's Interactive Storytelling*.

Spearman, C. 1987. "The Proof and Measurement of Association Between Two Things. By C. Spearman, 1904." *The American Journal of Psychology* 100 (3-4): 441–471.

Team Bondi. 2011. *L.A. Noire*. Team Bondi.

Telltale Games. 2009. *Tales of Monkey Island*. Telltale Games.

Thomas, Len. 1997. "Retrospective Power Analysis." *Conservation Biology* 11 (1) (February): 276–280. doi:10.1046/j.1523-1739.1997.96102.x.

Victor Gijsbers. 2010. *'Mid the Sagebrush and the Cactus*.

Visceral Games. 2011. *Dead Space 2*.

Vollmeyer, Regina, and Falko Rheinberg. 2006. "Motivational Effects on Self-Regulated Learning with Different Tasks." *Educational Psychology Review* 18 (3): 239–253. doi:10.1007/s10648-006-9017-0.

Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. "PARADISE." In , 271–280. Association for Computational Linguistics. doi:10.3115/976909.979652. http://portal.acm.org/citation.cfm?doid=976909.979652.

Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, Ace A. Kamm, and Alicia Abella. 1997. "PARADISE: A Framework for Evaluating Spoken Dialogue Agents." In , 271–280.

Walker, Marilyn, Candace Kamm, and Diane Litman. 2000. "Towards Developing General Models of Usability with PARADISE." *Nat. Lang. Eng.* 6 (3-4) (September): 363–377. doi:10.1017/S1351324900002503.

Wardrip-Fruin, Noah. 2009. "Expressive Processing: Digital Fictions, Computer Games, and Software Studies" (September 30). http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11872.

Wardrip-Fruin, Noah, Michael Mateas, Steven Dow, and Serdar Sali. 2009. "Agency Reconsidered." In *Breaking New Ground: Innovation in Games, Play, Practice and Theory: Proceedings of the 2009 Digital Games Research Association Conference*, ed. Atkins Barry, Kennedy Helen, and Krzywinska Tanya. London: Brunel University. http://www.digra.org/dl/display_html?chid=09287.41281.pdf.

Weibel, David, Bartholomäus Wissmath, Stephan Habegger, Yves Steiner, and Rudolf Groner. 2008. "Playing Online Games Against Computer- Vs. Human-controlled Opponents: Effects on Presence, Flow, and Enjoyment." *Comput. Hum. Behav.* 24 (5) (September): 2274–2291. doi:10.1016/j.chb.2007.11.002.

Weizenbaum, Joseph. 1966. "ELIZA - a Computer Program for the Study of Natural Language Communication Between Man and Machine." *Commun. ACM* 9 (1) (January): 36–45. doi:10.1145/365153.365168.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1 (6) (December 1): 80–83. doi:10.2307/3001968.

Witmer, Bob G., and Michael J. Singer. 1998. "Measuring Presence in Virtual Environments: A Presence Questionnaire." *Presence: Teleoper. Virtual Environ.* 7 (3) (June): 225–240. doi:10.1162/105474698565686.

Yee, Nick. 2006. "Motivations for Play in Online Games." *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society* 9 (6) (December): 772–775. doi:10.1089/cpb.2006.9.772.

APPENDIX A. USER STUDIES FOR COMPARING DIFFERENT DIALOGUE SYSTEMS : PROTOCOL & SURVEY

A.1 Protocol

We recruited participants from an introductory game design class taught at UCSC and the game design undergraduate major. All participants were 18 or older, and all were native English speakers. We conducted within-subject experiments. Participants filled out a survey on their gaming habits, a personality survey and an empathy survey before the experiment. After playing each version, the participants also filled out a survey aimed at measuring flow, presence, sense of control and enjoyment, and details of their experience with the game. After the play sessions, we conducted semi-structured interviews with the participants focused on more detailed investigations in the same issues.

A.2 Survey

Part 1. Pre-study Survey

FAÇADE DIALOGUE SYSTEMS SURVEY

Participant #: _____

Date: _____

Gender: **M F**

Occupation: _____

Age*: _____

***All participants in this survey are required to be 18 years of age or older.**

1. Experience using computers:

No experience 1 2 3 4 5 6 7 Expert User

2. Estimated hours using computers per week: _____

3. Estimated hours playing video games per week: _____

4. Favorite kind of games (check all that apply)

___ Action Adventure

___ Role-Playing

___ First Person Shooters

___ Strategy Games, including Real-Time Strategy

___ Adventure

___ Sports

___ Puzzle

___ Simulation games (e.g. SimCity, RollerCoaster Tycoon)

___ Massively Multiplayer

___ Casual games (web based)

___ Other _____

5. Estimated hours watching TV/movies per week: _____

6. Favorite kind of TV/movies (check all that apply)

___ Action

___ Drama

___ Comedy

___ Mystery

___ Detective Stories

___ Documentaries

___ Love Stories

___ Science Fiction

___ Thriller

___ Art films

___ Westerns

7. I play games on the following platforms:

___PC

___XBOX 360

___Playstation 3

___XBOX

___Playstation 2

___Nintendo Wii

___Nintendo DS

___PSP

___Other _____

8. How many times have you played *Façade*?

a. Never

b. Once

c. 2-5

d. 6-9

e. 10 or more

Please answer the following questions before playing the different versions of Facade.

(1: Strongly disagree – 7: Strongly agree)

1. When I watch a good movie, I can easily put myself in the place of a leading character.	1	2	3	4	5	6	7
2. When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.	1	2	3	4	5	6	7
3. After seeing a play or movie, I have felt as though I were one of the characters.	1	2	3	4	5	6	7

4. I am usually not objective when I watch a movie or play, and I often get completely caught up in it. 1 2 3 4 5 6 7

5. I really get involved with the feelings of characters in a novel. 1 2 3 4 5 6 7

Please select the set of statements that best describes your personality (if even just a little better). Choose an entire column based on whom you really are, not how you wish you were, or have to be at work.

a: (choose either the left group or right group based on how well it matches your personality)

Have high energy	Have quiet energy
Talk more than listen	Listen more than talk
Think out loud	Think quietly inside my head
Act, then think	Think, then act
Like to be around people a lot	Feel comfortable being alone
Prefer a public role	Prefer to work "behind-the-scenes"
Can sometimes be easily distracted	Have good powers of concentration
Prefer to do lots of things at once	Prefer to focus on one thing at a time
Outgoing & enthusiastic	Self-contained and reserved

b: (choose one column below)

Focus on details & specifics	Focus on the big picture & possibilities
Admire practical solutions	Admire creative ideas
Notice details & remember facts	Notice anything new or different
Pragmatic - see what is	Inventive - see what could be
Live in the here-and-now	Think about future implications
Trust actual experience	Trust my gut instincts
Like to use established skills	Prefer to learn new skills
Like step-by-step instructions	Like to figure things out for myself
Work at a steady pace	Work in bursts of energy

c: (choose one column below)

Make decisions objectively	Decide based on my values & feelings
Appear cool and reserved	Appear warm and friendly
Most convinced by rational arguments	Most convinced by how I feel
Honest and direct	Diplomatic and tactful
Value honesty and fairness	Value harmony and compassion
Take few things personally	Take many things personally
Tend to see flaws	Quick to compliment others
Motivated by achievement	Motivated by appreciation
Argue or debate issues for fun	Avoid arguments and conflicts

d: (choose one column below)

Make most decisions pretty easily	May have difficulty making decisions
Serious & conventional	Playful & unconventional
Pay attention to time & prompt	Less aware of time & run late
Prefer to finish projects	Prefer to start projects
Work first, play later	Play first, work later
Want things decided	Want to keep my options open
See the need for most rules	Question the need for many rules
Like to make & stick with plans	Like to keep plans flexible
Find comfort in schedules	Want the freedom to be spontaneous

Part 2. Post-Study Questionnaire

Please answer the following questions after playing each different version of Façade.

(1: not at all/strongly disagree, 7: very much/strongly agree)

1. I felt just the right amount of challenge. 1 2 3 4 5 6 7

2. My thoughts/activities ran fluidly and smoothly. 1 2 3 4 5 6 7

3. I didn't notice time passing. 1 2 3 4 5 6 7

4. I had no difficulty concentrating.	1	2	3	4	5	6	7
5. My mind was completely clear.	1	2	3	4	5	6	7
6. I was totally absorbed in what I was doing.	1	2	3	4	5	6	7
7. The right thoughts/movements occurred of their own accord.	1	2	3	4	5	6	7
8. I knew what I had to do each step of the way.	1	2	3	4	5	6	7
9. I felt that I had everything under control.	1	2	3	4	5	6	7
10. I was completely lost in thought.	1	2	3	4	5	6	7

1. When the game ended, I felt like I came back to the "real world" after a journey. (1: Strongly disagree – 7: Strongly agree)	1	2	3	4	5	6	7
2. The game came to me and created a new world for me, and the world suddenly disappeared when the game ended. (1: Strongly disagree – 7: Strongly agree)	1	2	3	4	5	6	7
3. While playing the game, I felt I was in the world the game created. (1: Never – 7: Always)	1	2	3	4	5	6	7
4. While playing the game, my body was in the room, but my mind was inside the world created by the game. (1: Never – 7: Always)	1	2	3	4	5	6	7

5. While playing the game, the game-generated world was more real or present for me compared to the "real world." (1: Never – 7: Always)	1	2	3	4	5	6	7
7. While playing the game, I NEVER forgot that I was in the middle of an experiment. (1: Never – 7: Always)	1	2	3	4	5	6	7
8. The game-generated world seemed to me only "something I saw" rather than "somewhere I visited." (1: Never – 7: Always)	1	2	3	4	5	6	7
9. While playing the game, my mind was in the room, not in the world created by the game. (1: Never – 7: Always)	1	2	3	4	5	6	7

1. How engaged were you in this version of the game? (1: Not at all – 7: Very engaged)	1	2	3	4	5	6	7
2. How engaging was the game overall? (1: Not at all – 7: Very engaging)	1	2	3	4	5	6	7
3. I enjoyed this version of the game. (1:Not at all – 7: Very much)	1	2	3	4	5	6	7

1. When do you feel your overall input (navigation, conversation, interaction with objects, etc.) had the MOST influence on the experience?

a. Beginning of the experience

b. Middle of the experience

c.Other

_____

_____

_____

_____

_____

When do you feel your overall input (navigation, conversation, interaction with objects, etc.) had the LEAST influence?

a. Beginning of the experience

b. Middle of the experience

c.Other

_____

_____

_____

_____

_____

When do you feel you had the most difficulty communicating with Trip and Grace?

a. Beginning of the experience

b. Middle of the experience

c. Other

_____

_____

_____

_____

_____

Breakdowns in communication with Trip and Grace occurred because (choose all appropriate):

a. I cannot type fast enough

b. I cannot think of things to say

c. I did not want to interrupt what they were saying

d. Trip and Grace do not understand me

e. Trip and Grace do not listen to me

f. The situation was tense and awkward for everyone

g. The computer program has errors

h. Other

_____

_____

_____

_____

_____

Which of the following endings have you experienced?

a. Trip left the apartment

b. Grace left the apartment

c. I was thrown out

d. Trip and Grace are going to work on their differences

e. I think Trip and Grace reconciled their differences, but I am not sure

f. Other _____

How much did your interaction influence the story?

No influence 1 2 3 4 5 6 7 Significant influence

Part 3. Post-Study Interview

ENGAGEMENT

Which version was the most engaging? Why?

What character were you playing in each version? What was frustrating/enjoyable/entertaining about the characters you were playing? How did the different interfaces allow you to play out this character?

ENJOYMENT

1. Which version did you enjoy the least? Why?

2. Which version did you enjoy the most? Why?

3. Would you like to play this game again? Why\why not?

AGENCY

Which interface variation was the most challenging to learn and use? Why?

Which interface variation was the easiest to use? Why?

Which interface variation was the most natural to use? Why? Can you give an order?

Which interface gave you the strongest sense of control over the story?

Which interface gave you the least sense of control over the story?

Which interface made you most compelled/motivated to move the story forward?

How did you decide what to do in each different version? What cues did you use to move the story forward?

Do you think you had a strategy? Tell me about your strategy. How easy was it to realize this strategy in each different version?

How much influence did you feel over the story? How did the different interfaces effect your sense of influence within the game world?

How did the characters react to your actions and statements? Did you feel any difference between the different versions?

Did you find the interfaces difficult or easy to use? How so?

What were the advantages and disadvantages of each different version?

Do you have any other thoughts on the different dialogue interfaces you just used?

ADDITIONAL END OF STUDY QUESTIONS

You said that you felt like you [did not] have influence over the ending of the story.

Tell me more about that.

How did you feel about how it ended?

Did it matter to you how it ended?

How did your sense of presence change in those different versions?

Tell me more about that.

On the survey, you marked that [interface x] did not feel natural.

Tell me more about that.

Compare how natural the three experiences felt.

On the survey, you said that you were most engaged/immersed during [version x].

Tell me more about that.

Anything else? Compare the different versions that you just played.

On the survey, you said that [version x] had a more challenging interface:

Why?

How did you cope with the problems you mentioned?

What would make that interface better?

How was the pace of the game? Which one had the best pacing? Which one had the worst? How different did they feel in terms of pacing?

What would make the game more fun for you? What would make the game more engaging to you? What would make the content better to you?

What would your ideal interface be? How would it work?

APPENDIX B. USER STUDIES FOR COMPARING DIFFERENT PACING OPTIONS IN A NLU SYSTEM: PROTOCOL & SURVEY

B.1 Protocol

We recruited participants from an introductory game design class taught at UCSC and the game design undergraduate major. All participants were 18 or older, and all were native English speakers. We conducted within-subject experiments. Participants filled out a survey on their gaming and media consumption habits before the experiment. After playing each version, the participants also filled out a survey aimed at measuring flow, presence, sense of control and enjoyment, and details of their experience with the game. After the play sessions, we conducted semi-structured interviews with the participants focused on more detailed investigations regarding those dimensions.

B.2 Survey

P#: Date: _____
_____

Gender: M / F Occupation: _____
Age: _____

1. Hours Using Computers Per Week: _____

2. Video Game Hours Per Week: _____

3. Preferred Games:

__ Action Adventure
__ Adventure
__ Casual
__ First Person Shooter
__ Massively Multiplayer
__ Puzzle
__ Role-Playing
__ Simulation
__ Sports
__ Other/ Not Sure_____

4. TV / Movie Hours Per Week: _____

5. Favorite Type of Show / Movie:

__ Action
__ Art
__ Comedy
__ Detective
__ Documentaries
__ Drama
__ Love / Romance
__ Mystery
__ Science Fiction
__ Thriller / Horror
__ Western
__ Other/ Not Sure_____

6. I play games on:

__ PC
__ Console
__ Hand held
__ Mobile
__ Others _____

7. How many times have you played Facade?

__ Never
__ Once
__ 2-5
__ 6-9
__ 10 or More

Please answer the following questions before playing the different versions of Facade.

(1: Not at all / Strongly Disagree, 7: Very much / Strongly Agree)							
When I watch a good movie, I can easily put myself in the place of a leading character.	1	2	3	4	5	6	7
When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.	1	2	3	4	5	6	7
After seeing a play or movie, I have felt as though I were one of the characters.	1	2	3	4	5	6	7
I am usually not objective when I watch a movie or play, and I often get completely caught up in it.	1	2	3	4	5	6	7

I really get involved with the feelings of characters in a novel.	1 2 3 4 5 6 7

Part 2. Post-Study Questionnaire

Please answer the following questions after playing each different version of Façade.

(1: not at all/strongly disagree, 7: very much/strongly agree)

1. I felt just the right amount of challenge.	1 2 3 4 5 6 7
2. My thoughts/activities ran fluidly and smoothly.	1 2 3 4 5 6 7
3. I didn't notice time passing.	1 2 3 4 5 6 7
4. I had no difficulty concentrating.	1 2 3 4 5 6 7
5. My mind was completely clear.	1 2 3 4 5 6 7
6. I was totally absorbed in what I was doing.	1 2 3 4 5 6 7
7. The right thoughts/movements occurred of their own accord.	1 2 3 4 5 6 7
8. I knew what I had to do each step of the way.	1 2 3 4 5 6 7
9. I felt that I had everything under control.	1 2 3 4 5 6 7
10. I was completely lost in thought.	1 2 3 4 5 6 7

1. When the game ended, I felt like I came back to the "real world" after a journey. (1: Strongly disagree – 7: Strongly agree)	1	2	3	4	5	6	7
2. The game came to me and created a new world for me, and the world suddenly disappeared when the game ended. (1: Strongly disagree – 7: Strongly agree)	1	2	3	4	5	6	7
3. While playing the game, I felt I was in the world the game created. (1: Never – 7: Always)	1	2	3	4	5	6	7
4. While playing the game, my body was in the room, but my mind was inside the world created by the game. (1: Never – 7: Always)	1	2	3	4	5	6	7
5. While playing the game, the game-generated world was more real or present for me compared to the "real world." (1: Never – 7: Always)	1	2	3	4	5	6	7
7. While playing the game, I NEVER forgot that I was in the middle of an experiment. (1: Never – 7: Always)	1	2	3	4	5	6	7
8. The game-generated world seemed to me only "something I saw" rather than "somewhere I visited." (1: Never – 7: Always)	1	2	3	4	5	6	7
9. While playing the game, my mind was in the room, not in the world created by the game. (1: Never – 7: Always)	1	2	3	4	5	6	7

1. How engaged were you in this version of the game? 1 2 3 4 5 6 7

(1: Not at all – 7: Very engaged)

2. How engaging was the game overall? 1 2 3 4 5 6 7

(1: Not at all – 7: Very engaging)

3. I enjoyed this version of the game. 1 2 3 4 5 6 7

(1:Not at all – 7: Very much)

When do you feel your overall input (navigation, conversation, interaction with objects, etc.) had the MOST influence on the experience?

a. Beginning of the experience

b. Middle of the experience

c. Other

_____

_____

_____

_____

_____

When do you feel your overall input (navigation, conversation, interaction with objects, etc.) had the LEAST influence?

a. Beginning of the experience

b. Middle of the experience

c. Other

_____
_____
_____
_____
_____

When do you feel you had the most difficulty communicating with Trip and Grace?

a. Beginning of the experience

b. Middle of the experience

c. Other

_____
_____
_____
_____
_____

Breakdowns in communication with Trip and Grace occurred because (choose all appropriate):

a. I cannot type fast enough

b. I cannot think of things to say

c. I did not want to interrupt what they were saying

d. Trip and Grace do not understand me

e. Trip and Grace do not listen to me

f. The situation was tense and awkward for everyone

g. The computer program has errors

h. Other

_____

_____

_____

_____

_____

Which of the following endings have you experienced?

a. Trip left the apartment

b. Grace left the apartment

c. I was thrown out

d. Trip and Grace are going to work on their differences

e. I think Trip and Grace reconciled their differences, but I am not sure

f. Other _____

How much did your interaction influence the story?

No influence 1 2 3 4 5 6 7 Significant influence

Part 3. Post-Study Interview

ENGAGEMENT

Which version was the most engaging? Why?

What character were you playing in each version? What was frustrating/enjoyable/entertaining about the characters you were playing? How did the different interfaces allow you to play out this character?

ENJOYMENT

1. Which version did you enjoy the least? Why?

2. Which version did you enjoy the most? Why?

3. Would you like to play this game again? Why\why not?

AGENCY

Which interface variation was the most challenging to learn and use? Why?

Which interface variation was the easiest to use? Why?

Which interface variation was the most natural to use? Why? Can you give an order?

Which interface gave you the strongest sense of control over the story?

Which interface gave you the least sense of control over the story?

Which interface made you most compelled/motivated to move the story forward?

How did you decide what to do in each different version? What cues did you use to move the story forward?

Do you think you had a strategy? Tell me about your strategy. How easy was it to realize this strategy in each different version?

How much influence did you feel over the story? How did the different interfaces effect your sense of influence within the game world?

How did the characters react to your actions and statements? Did you feel any difference between the different versions?

Did you find the interfaces difficult or easy to use? How so?

What were the advantages and disadvantages of each different version?

Do you have any other thoughts on the different dialogue interfaces you just used?

ADDITIONAL END OF STUDY QUESTIONS

You said that you felt like you [did not] have influence over the ending of the story.

Tell me more about that.

How did you feel about how it ended?

Did it matter to you how it ended?

How did your sense of presence change in those different versions?

Tell me more about that.

On the survey, you marked that [interface x] did not feel natural.

Tell me more about that.

Compare how natural the three experiences felt.

On the survey, you said that you were most engaged/immersed during [version x].

Tell me more about that.

Anything else? Compare the different versions that you just played.

On the survey, you said that [version x] had a more challenging interface:

Why?

How did you cope with the problems you mentioned?

What would make that interface better?

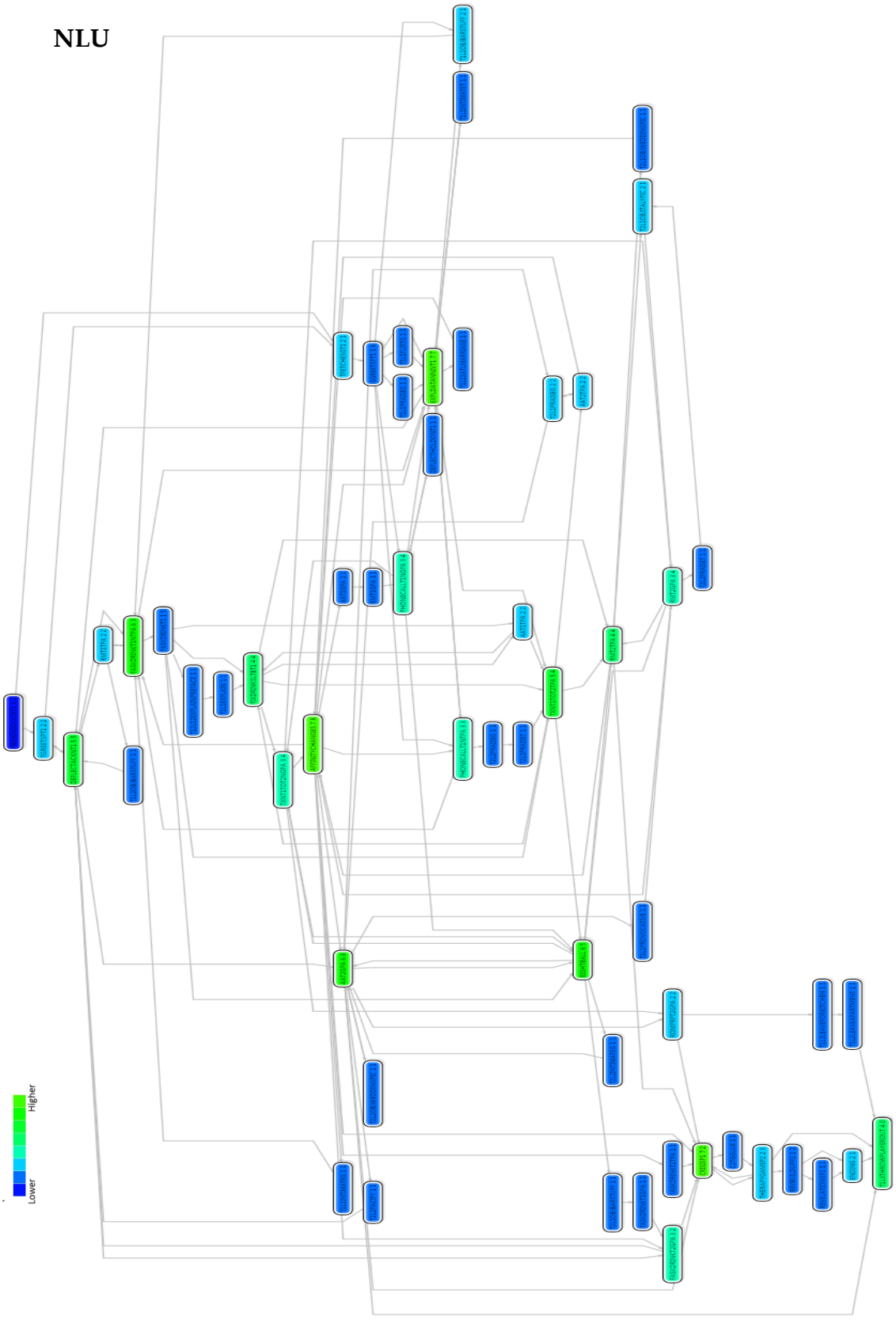
How was the pace of the game? Which one had the best pacing? Which one had the worst? How different did they feel in terms of pacing?

What would make the game more fun for you? What would make the game more engaging to you? What would make the content better to you?

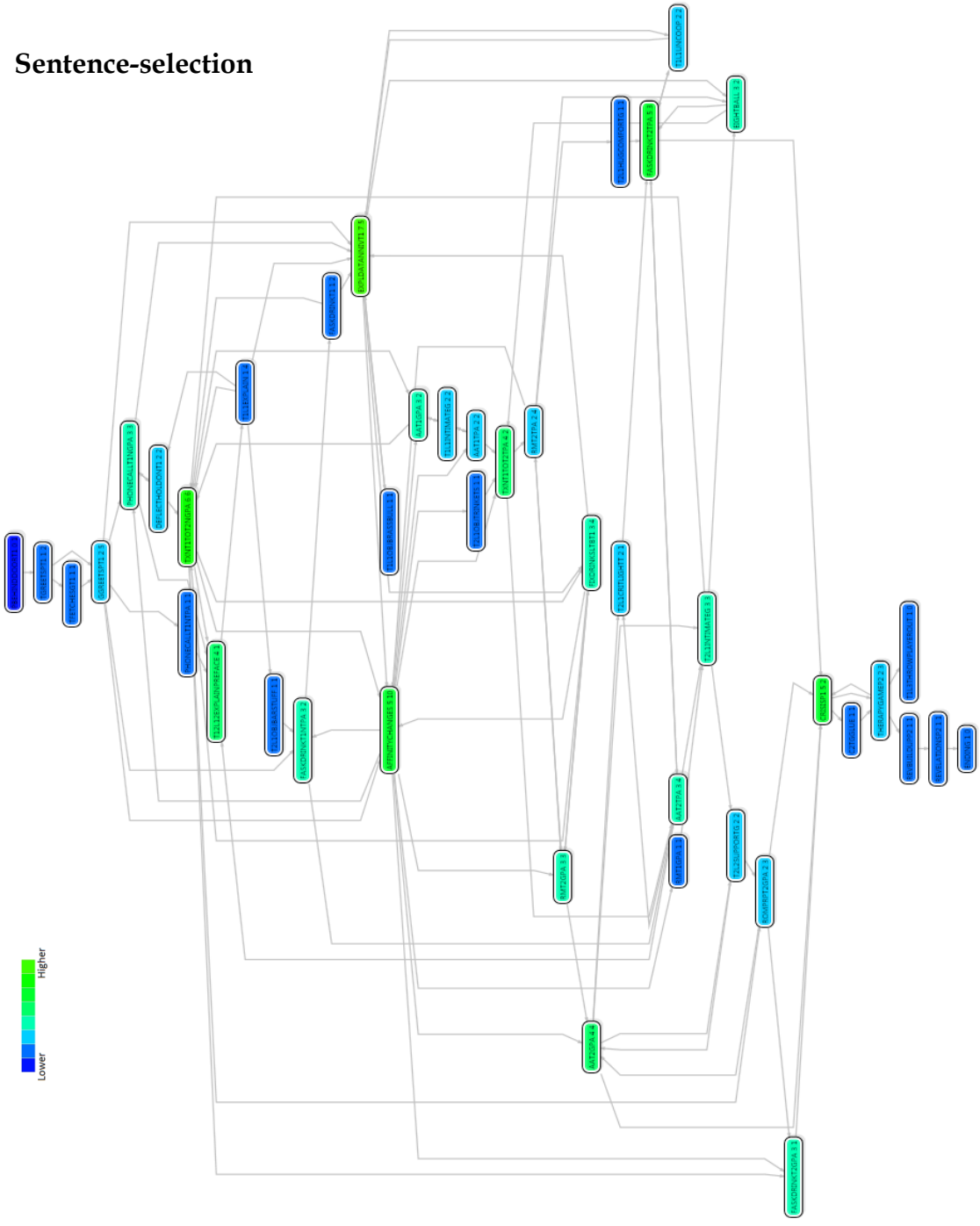
What would your ideal interface be? How would it work?

APPENDIX C. STORY GRAPHS FOR NLU AND SENTENCE-SELECTION

NLU



Sentence-selection



APPENDIX D. STATISTICAL POWER ANALYSIS FOR INSIGNIFICANT RESULTS

Statistical power analysis is a meta-analysis method that can be used to determine the probability of rejecting a null hypothesis when the null hypothesis is indeed false (Cohen 1988). Suppose an experimenter conducts an experiment comparing means across two groups, with the null hypothesis stating that there's no difference in the mean values of the measured variables across these two groups. The power of this test refers to the possibility that if there's indeed a difference in the mean, that difference will be revealed by the experiment. Obviously high statistical power is always a desired goal, with 0.8 being a widely adapted convention as a minimum.

Power analysis can be done a-priori or post-hoc. A-priori power analysis is generally used to determine the required sample size for an experiment to reach a significant power. Power analysis is usually more useful when conducted a-priori, but it requires estimation of important parameters and informed guesses about what effect size is of importance for the particular experiment, which is usually hard to do especially for domains for which there isn't sufficient knowledge to make an educated guess about these parameters. The usefulness of post-hoc power analysis is debated frequently (Thomas 1997), but it could still be useful when conducted in certain ways. In this analysis, I decided to use the observed mean and standard deviation to compute the observed effect size for each item for which we got an insignificant result. Then, assuming this effect size to be a population parameter for the measured item, I calculated the number of samples that would be needed to

reach a power of 0.8. The results for the six items that we couldn't find significant results for are summarized in the table below, with both calculated power and estimated sample size to reach a power of 0.8.

Item	Observed effect size	Observed power (n = 100)	Required sample size for power >= 0.8
1. It was easy to decide what I want to say using this interface.	0.19	0.15	~ 960 (480 per group)
5. The choices I wanted to make were present in the interface.	0.19	0.15	~ 960 (480 per group)
6. The ending made sense to me.	0.09	0.08	~ 3500 (1750 per group)
7. The number of interactions the interface allowed me to have was satisfactory.	0.20	0.17	~ 760 (380 per group)
11. I enjoyed using this interface.	0.11	0.09	~ 2540 (1270 per group)
12. I felt the ending was a direct result of my interactions.	0.12	0.09	~ 2300 (1150 per group)