Persuasive Computer Dialogue Improving Human-Computer Communication

PIERRE Y. ANDREWS

Ph.D. Thesis

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

THE UNIVERSITY of York

Artificial Intelligence Group
Department of Computer Science
United Kingdom

26 September 2008

Abstract

The research reported in this thesis aims at developing techniques for achieving persuasive dialogue between a human user and the computer; in particular, it focuses on the formalisation of argumentation for dialogue planning in combination with a novel framework for dialogue management to improve both the reactivity of the dialogue system and its persuasiveness.

The main hypothesis of this thesis is that a persuasive dialogue requires specific techniques that cannot be implemented in a purely task-oriented system. In particular, *Persuasive Communication* requires to build social bonds, trust and involvement through social cues, as well as tailored reactions to the user's arguments which are difficult to plan a priori. The work presented goes towards finding novel techniques for dialogue management that offer the reactivity and flexibility needed for persuasion.

In the designed dialogue framework, a layered management system that takes advantage of state-of-the-art dialogue management approaches is developed. A planning component tailored to reason about persuasion searches the ideal path in an argumentation model to persuade the user. To give a reactive and natural feel to the dialogue, a reactive component extends this task-oriented layer, using online activation techniques to select dialogue strategies that are difficult to plan a priori.

The use of planning guarantees the achievement of persuasive goals and the consistency of the dialogue whereas the reactive component is able to better adapt to the user reactions and show more flexibility, allowing a better perception of the system by the user.

An evaluation protocol is proposed, providing a simple and independent metric for the persuasiveness of the dialogue system. The evaluation shows that the novel layered dialogue management framework – the EDEN Framework – achieves a measured persuasiveness better than a purely task-oriented system as the novel approach is able to react more smoothly to the user's interaction. In the final evaluation, the Personage (Mairesse & Walker 2007) generator is used to display different personalities to the user during the persuasive dialogue. Experiments with the latter approach show the impact of personality simulation on the user's perception of the interaction and show the influence of the system *displayed personality* on the overall persuasiveness.

Contents

List of Tables

List of Figures

Preface: The Argument Sketch

MAN: Ah! I'd like to have an argument, please.

RECEPTIONIST: Certainly sir. Have you been here before?

MAN: No, I haven't, this is my first time.

. . .

RECEPTIONIST: Mr. DeBakey's free, but he's a little bit conciliatory.

RECEPTIONIST: Ahh yes! Try Mr. Barnard; room 12.

. . .

[*Walk down the corridor*]

MAN: [*Knock*]

ARGUER: Come in.

MAN: Ah! Is this the right room for an argument?

ARGUER: I told you once.

MAN: No you haven't.

ARGUER: Yes I have.

MAN: When?

ARGUER: Just now.

. . .

MAN: You did not!!

ARGUER: Oh, I'm sorry, just one moment. Is this a five minute argument or the full half hour?

MAN: Oh! Just the five minutes.

ARGUER: Ah! thank you. Anyway, I did.

MAN: You most certainly did not.

ARGUER: Look, let's get this thing clear; I quite definitely told you.

MAN: No you did not.

ARGUER: Yes I did.

. . .

MAN: You didn't.

ARGUER: Did.

MAN: Oh look, this isn't an argument.

ARGUER: Yes it is.

MAN: No it isn't. It's just *contradiction*.

ARGUER: No it isn't.

. . .

MAN: Oh, this is futile!

ARGUER: No it isn't.

MAN: I came here for a good argument.

ARGUER: No you didn't; no, you came here for an argument.

MAN: An argument isn't just contradiction.

ARGUER: It can be.

MAN: No it can't. An argument is a connected series of statements intended to establish a proposition.

ARGUER: No it isn't.

MAN: Yes it is! It's not just contradiction.

ARGUER: Look, if I argue with you, I must take up a contrary position.

MAN: Yes, but that's not just saying "No it isn't."

ARGUER: Yes it is!

MAN: No it isn't!

ARGUER: Yes it is!

MAN: Argument is an intellectual process. Contradiction is just the automatic gainsaying of any statement the other person makes.

ARGUER: [*short pause*] No it isn't.

MAN: It is.

ARGUER: Not at all.

MAN: Now look.

ARGUER: (Rings bell) Good Morning.

MAN: What?

ARGUER: That's it. Good morning.

MAN: I was just getting interested.

ARGUER: Sorry, the five minutes is up.

MAN: That was never five minutes!

ARGUER: I'm afraid it was.

MAN: It wasn't. [*Pause*]

ARGUER: I'm sorry, but I'm not allowed to argue anymore.

MAN: What?!

ARGUER: If you want me to go on arguing, you'll have to pay for another five minutes.

MAN: Yes, but that was never five minutes, just now. Oh come on!

ARGUER: [*Hums*]

MAN: Look, this is ridiculous.

ARGUER: I'm sorry, but I'm not allowed to argue unless you've paid!

MAN: Oh, all right. [*pays money*]

ARGUER: Thank you. [*short pause*]

MAN: Well?

ARGUER: Well what?

MAN: That wasn't really five minutes, just now.

ARGUER: I told you, I'm not allowed to argue unless you've paid.

MAN: I just paid!

ARGUER: No you didn't.

MAN: I DID!

ARGUER: No you didn't.

MAN: Look, I don't want to argue about that.

ARGUER: Well, you didn't pay.
MAN: Aha. If I didn't pay, why are you arguing? I Got you!
ARGUER: No you haven't.
...

(From "Monty Python's Previous Record" and "Monty Python's Instant Record Collection". Originally transcribed by Dan Kay[1].)

[1] `http://www.mindspring.com/~mfpatton/sketch.htm`

Acknowledgement

I would like to begin by thanking my parents without whome I would never achieved as much as I have. They have always supported my education and been a great motivation to accomplish this thesis.

This thesis would never have finished without Malihe believing in me and unconditionally caring for my success. It would also never have started without the support of Silvia that walked by my side most of the way. They will always have a special place in my life.

Dr. Suresh Manandhar has also always believed in my research path and strongly advised me to finish this Ph.D. For this and his moral support all along the process of this thesis research, I would like to thank him. I am also grateful to Prof. Marilyn Walker and Dr. Alistair Edwards, the thesis examiners, that should be thanked for the great amount of feedback on the corrections that make this thesis – and associated publications – better.

I would like to acknowledge the very strong help provided by Dr. Paul Cairns for his help in constructing the human centred evaluation and analysing the results of these experiments; Rana Tayara also provided interesting views on the interpretations of personalities. I would also like to thank all of the participants that took part in these experiments and completed the evaluation, providing very interesting results.

I am also grateful to all the people I met around York and that supported me in many ways, with their friendship as well as strong moral and logistical support. I would thus start by thanking Estefania for her strong presence along the hardest part of my thesis. I wouldn't have made it in York without the strong help of Sergio, Annick, Tom, Richard and Burcu as well as all the tangueros, in particular Nada, Juan, Tobias and Radhika. I cannot thank enough all the persons that have been here for me but that cannot fit in this acknowledgement.

Finally, none of this thesis would have been possible without the financial support of Unilever Corporate Research and my industrial contact there, Marco De Boni.

Parts of the present research have been previously presented or published in:

- Andrews, Pierre, Manandhar, Suresh (2009). Measure Of Belief Change as an Evaluation of Persuasion. In *Proceedings of the AISB'09 Persuasive Technology and Digital Behaviour Intervention Symposium*. Edinburgh.

- Andrews, Pierre, Manandhar, Suresh, and De Boni, Marco (2008). Argumentative human computer dialogue for automated persuasion. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 138-147, Columbus, Ohio. Association for Computational Linguistics.

- Andrews, Pierre, Manandhar, Suresh, and De Boni, Marco (2006). Integrating emotions in persuasive dialogue: A multi-layer reasoning framework. In *Proceedings of the 19th International FLAIRS Conference*.

- Andrews, Pierre, De Boni, Marco, and Manandhar, Suresh (2006). Persuasive argumentation in human computer dialogue. In *Proceedings of the AAAI 2006 Spring Symposium on Argumentation for Consumers of Healthcare*, Stanford University, California.

Chapter 1

Introduction

Although dialogue is not new to computational linguistics, it is difficult to find a balance between a computer that is either too task oriented or does not guarantee topic consistency during dialogue. The research presented in this thesis investigates whether, by using the long-studied theories of rhetoric in philosophy and linguistics, in combination to state-of-the-art dialogue management techniques, it is possible to build a more natural *automated* dialogue system to achieve persuasion. The interest of this thesis is to study the application of rhetoric and persuasion theories to build a framework applicable to different dialogue domains. This research extends the work on the persuasive aspects of dialogue, which has already been taken into account in specialised areas such as law (see for example Bench-Capon 2003a;b) and where research has only started recently to investigate persuasion through automated dialogue (Guerini, Stock, & Zancanaro 2004, Mazzotta, de Rosis, & Carofiglio 2007).

The need for a complete planning system in persuasive dialogue has been pointed out in Gilbert, Grasso, Groarke, Gurr, & Gerlofs (2003) which presents a set of complex steps to be followed to generate persuasive dialogue moves. In the framework proposed by Gilbert et al., the dialogue system would need to identify arguments in the user's utterances, evaluate their correctness and build its utterances based on this processing.

The framework proposed by Gilbert et al. (2003) is still highly theoretical and relies on ongoing research in natural argumentation and computational linguistics. However, a number of working frameworks for planning and generating arguments have already been developed in the natural language field (Grasso 2002, Reed & Grasso 2001, Reiter, Robertson, & Osman 2003) using rhetoric theories like the “New Rhetorics” (Perelman & Olbrechts-Tyteca 1958) or the “Rhetorical Structure Theory” (Mann 1999) but have focused on monologue.

Research in natural language dialogues based on argumentation raises the issue of understanding the user’s counter-arguments and reasoning on the reaction to take. Detecting the argument strategy used, deciding whether or not it is a fallacy, and verifying the veracity of each of the premises put forward require a large knowledge base and have – to the best of my knowledge – not been implemented in a working dialogue framework. This thesis proposes a dialogue management model where persuasive strategies and reactive argumentative strategies are separated. The hypothesis of the thesis is that such model will ease the task of persuasive dialogue management by separating the logical aspect of persuasion from the social cues management.

The user needs to feel familiar with the system to be receptive to its arguments; indeed, research in Persuasive Communication (see section 2.2) emphasises the need to take into account the personality of the user to formulate effective arguments, as well as the impact of users’ involvement and motivation on the persuasion performances.

A planned approach to social language modelling has been proposed by Cassell & Bickmore (2002) which uses small-talk to build the user’s trust. Cassell & Bickmore uses an activation network to plan dialogue moves, choosing, according to the user model, between social-talk moves or task achievement moves. This system has the drawback that it mixes small-talk and task planning in the same planning network which can disadvantage the portability of the system (Allen, Ferguson, & Stent 2001b) by adding complexity to the domain authoring.

The hypothesis made in this thesis is that the management of the discourse level features of the dialogue – such as small-talk – do not need to be mixed with the true content planning. Splitting the dialogue management in layers (as proposed by Lemon, Cavedon, & Kelly 2003, Stent 2002 or Zinn, Moore, & Core 2002) allows for an independent management of the user's trust and comfort in the dialogue and simplify the planning of the argumentation.

It is in that perspective that the research performed in this thesis mixes a reactive approach with planning techniques:

1. A *planning* layer, directed by an argumentation model, decides of which argument to defend in the current dialogue segment but is not responsible for the actual dialogue realisation.

2. An *interaction* layer receives constrains from the *planning* layer and uses reactive techniques to construct the actual dialogue within these constrains.

Within such a model, dialogue management is separated between the two distinct components of the dialogue: task achievement and discourse rules management. This makes the planning of the persuasive argumentation more direct as it is not mixed with the dialogue's surface-traits management. In addition, separating the dialogue management between two independent layers eases the development of the social cues and persuasive strategies handling which in turn allows the user to perceive a more natural dialogue and be persuaded more easily (see chapters 4 and 5).

The hypothesis of this thesis is that this hybrid model will allow for an easier implementation of a dialogue management framework for persuasion than strictly planned approaches or strictly reactive techniques. This thesis presents the Eden Dialogue Framework that provides tools for evaluating this hypothesis and exploring diverse parameters of persuasive dialogue.

An objective metric of persuasiveness is introduced to evaluate the dialogue model fitness to the task in a controlled environment. Interactive experiments with users are performed where they have to achieve a ranking task: the users have to rank a set of items, representing their beliefs about a behaviour; they are then faced with a persuasive dialogue that tries to change their beliefs and thus the given ranking. The evolution of the ranking of items during the dialogue provides a controlled metric evaluation of the persuasiveness of the system (see section 4.1).

The Eden Framework reflects the split between the logical level of the dialogue and its discourse level and is composed of three layers of management:

- A *planning layer* creates a plan of the arguments to present during the dialogue. This plan guarantees the achievement of persuasive goals and maintains the consistency of the dialogue. The planning layer performs content selection and planning of the dialogue level structure.

- A *reactive layer* is responsible for the interactive dialogue management in the constraints provided by the plan. For each plan step given by the previous layer, the reactive layer manages a dialogue segment topically related to the argument. This layer is responsible for the naturalness of the dialogue and to manage discourse level interactions that are hard to plan a priori. The reactive layer selects and structure content in sub-dialogues, reacting to the user within the communicative goals provided by the plan.

- A generation layer realises the surface form of each utterance according to the instructions provided by the reactive layer. It is responsible for the content selection, structure planning and lexical realisation of each individual utterances.

Modelling argumentative dialogue raises the problem of the liberty of reactions of the user. In a task-oriented system, where the system and the user share a task – such as booking theatre tickets –, the system can lead the user in a question/answer dialogue where the diversity of content input by the user is limited. However, in argumentative dialogue, the reaction of the users – presenting counter arguments, agreeing, etc. – can take many forms and content. In particular, it is difficult to predict the counter arguments they will provide. The dialogue model must be able to describe a set of reactions to these inputs that can be used to achieve the long term persuasive goals and make the dialogue feel natural.

The dialogue model described in this thesis separates the logical aspect of the dialogue, that can be used to reason about long term consistency and achievement of the persuasive goals, and the reactive aspects of the dialogue. This separation is achieved by delegating the realisation of entire dialogue segments to the reactive component. In this architecture, the dialogue plan – unlike standard planned approaches – does not define the content of specific utterances, but the communicative goals of consistent dialogue segments.

A plan step defines an argument to present to the user, setting constrains on the reactive strategies that the system can use. The actual counter argument of the user, combined with these constrains will activate a reaction that describes one utterance in the dialogue segment. Thus, the dialogue manager can be more reactive to the user without impairing the logical representation of the model and the planning complexity.

In a perfect argumentative dialogue, the system – as described by Gilbert et al. (2003) – will be able to undestand the full *content* of the user counter argument to activate a taillored reaction. The system should be able, through deep parsing of the user's utterance to understand the argument and decide of the best answer. This requires understanding the type of argument the user is using, as well as evaluating the facts presented by the user and their veracity. Such deep understanding of the user's argument supposes the availability of a large knowledge base for fact evaluation, as well as a mechanism to parse the natural language utterances to map with this facts database.

In the current state of the art of natural language understanding and natural argumentation, this task is not possible and the dialogue model must integrate simplifications of the understanding task while keeping the dialogue as natural as possible. In this thesis, the dialogue system, instead of trying to understand the *content* of the user's argument, tries to understand the *intent* of the argument. This understanding is done by classifying each user utterance as either a *disagreement* or an *agreement* (see section 3.9) with the system argument and to perform a shallow extraction of keywords representing the topic of the user's reaction.

The understanding of the user *intent* allows the description of the interaction model in four types of reactions (see section 3.8):

Agreement The user agrees with the system's argument and accepts the new belief presented. The dialogue segment is thus complete and the system can start a new dialogue segment about the next argument selected by the planner.

Partial Agreement The user disagrees with the system's argument and the system has to counter argue and support its argument to achieve the dialogue segment. This implies activating tailored reactions to support the current argument.

Disagreement After extensive support from the system, the user might continue disagreeing with the system's argument. The current dialogue segment failed and the system cannot rely on the beliefs it was trying to introduce to complete its persuasive goals. The plan – that was relying on this belief for its next steps – is thus invalid and the system has to find a new plan to achieve its persuasive goals.

Smoothing Independently from the argumentation process and the user's agreement or disagreement, the dialogue manager must maintain a natural dialogue and uses reactions that generate social cues and chit chat with the user.

To evaluate the persuasiveness of this dialogue model, an experiment is designed that compares the new approach to a strictly planned dialogue management approach. The results to this evaluation show that the reactive dialogue system is able to achieve better persuasion without impairing the user's perception of the system.

In addition to reactive dialogue management, persuasion requires that the system adapts to the user model. Alignment of the system style of interaction has been identified in the literature as a possible factor affecting the persuasiveness of the dialogue. To explore this effect, an experiment is designed that evaluates the effect of personality matching on the persuasiveness. The dialogue system is used with a Natural Language Generator that is able to generate different surface utterance realisations for the same semantic content and simulate different personality traits. An exploratory experimental setup provides results showing that the users' extroversion has an impact on the style of system interaction they prefer.

This thesis is structured to provide an introduction to the problematics of persuasive dialogue, the existing solutions and the requirements that were identified for the design of a novel dialogue model, the persuasive framework and its technical details are then detailed, followed by the evaluations that were performed to test the fitness of the new model to the persuasive task. Chapter 2 presents the state of the art in human-computer dialogue as well as the sociology and philosophy guidelines for persuasion and argumentation. Chapter 3 describes the novel dialogue management approach. Chapter 4 evaluates the system persuasiveness compared to a strictly planned approach and Chapter 5 presents the exploratory experiment on user personality matching.

Chapter 2

Field Survey and Review

2.1 Computational Models

In this section, the state-of-the-art techniques developed in the field of computer science for persuasive communication are reviewed. After an introduction on the existing techniques in human-computer dialogue, the current research in computer science in argumentation and rhetoric is discussed in detail.

2.1.1 Human-Computer Dialogue

Human-computer dialogue is a field of research that has been studied by computer scientists and linguists since the Sixties. It was probably initiated by the *Turing Test* problem, which tries to evaluate Human-Computer Interaction.

The first real computer-dialogue system was probably the Eliza psychotherapist chatbot (Weizenbaum 1966) that uses a simple pattern matching inference system written in Lisp to construct answers from the user's input.

Seven main issues were identified in Lewin, Rupp, Hieronymus, Milward, Larsson, & Berman (2000) to implement a human-like dialogue:

1. The management of *"turn-taking"* is difficult for a system as it is not always clear when it has to reply to the user or wait for more inputs. In

fact, this is linked to the need for fluidity of a dialogue and the need to implement a system that is able to quickly deal with a limited set of data to generate an utterance.

2. *Understanding the user input* is another issue linked to the needs explained above. In fact, most systems prefer to perform a shallow processing of the user's utterance to extract its main features: which modality is used, what is the main topic, for example.

3. At the other end of the processing is the *generation of an utterance*, which implies the construction of replies understandable by humans. The system needs to identify the data to integrate in the utterance and construct a correct grammatical structure to present them. In complex Natural Language Generation (NLG) systems, this is divided in three phases: *text planning* – or *content selection* –, *sentence planning* and *linguistic* – or *surface* – *realisation* (see Reiter & Dale 1997); however, in some simple data retrieval systems, the reply does not have to be in natural language, as it can simply be a timetable or other type of structured information.

4. The *planning* of the dialogue generates some interesting issues:

   (a) The system has to take into account – and understand – the *user's intention* in the dialogue. In complete and complex systems, this implies managing the user's goals, beliefs and knowledge.

   (b) The system, if it wants some independence from the user, has to *manage its own communicative intentions*. This implies deciding of a dialogue strategy that specifies where the dialogue should go next if the user does not take initiative or if the system needs some clarification.

5. Other issues are mainly linked to the management of the *knowledge* required by the dialogue, but note that they are also highly linked to the planning:

(a) The system has to be able to *maintain the context* of the dialogue. This allows to deal with special language construction – like ellipsis – where some of the content is omitted in the user input.

(b) The system must be able to keep trace of the *topics under discussion* to be sure to "talk" about the right topic in the utterances to come.

The current state-of-the-art in dialogue management can be divided along two main types of dialogue. These types relate to the manner in which the initiative is taken and the planned dialogue moves.

Task Oriented

Task oriented dialogues have a task to achieve in one dialogue. The interaction is terminated once the task is finished and the system will perform the same task for each dialogue.

Such systems are mainly used in three different applications:

- *Information retrieval* systems – for example (Carlson 1996) – use the dialogue to construct a database query to find the information in which the user is interested.
- *Enquiry and booking* systems are mainly aimed at answering a particular user request, like planning a train trip (Allen, Ferguson, Miller, Ringger, & Sikorski 2000) or booking a theatre ticket (Hulstijn, Steetskamp, Doest, Burgt, & Nijholt 1996).
- *Tutoring* systems are used to teach new knowledge to the user. These systems often keep the initiative and have to understand what new knowledge has to be explained to the user.

Multiple techniques have been developed to manage these types of dialogue. They are suited to all types of task oriented dialogue and are chosen in regard of the complexity of the dialogue task.

The simplest planning is done by *finite-state machines* where dialogue moves correspond to transitions between a set of states in the machine. In such an architecture, the system keeps the initiative and leaves a limited set of choices to the user that has to make a move that the state machine understands. This type of system can therefore only be applied to tasks that have a restricted and ordered structure that limits the user input variability.

Frame-based systems are more flexible as they are tailored to the structure of the topic of the dialogue. For example, in an information retrieval dialogue (Lisowska, Rajman, & Bui 2005), the system will know the minimum set of fields – or *frames* – to query from the user before it is able to construct a database query. The system therefore creates a dialogue asking the user all the required questions and clarify the ambiguous data to construct the database query.

Dialogues that may require a more complex management system, which depends on more variables for building the dialogue's structure, use real *planning systems*. Thus, the dialogue moves are formally coded as actions with pre and post-conditions that should be verified. A planner is then used to infer a plan from the current knowledge base – that sets the context to compute the pre-conditions – and the goal to accomplish – that sets the possible post-conditions.

The two first types of dialogue management have to be carefully designed and tailored to the particular task, which requires a good knowledge of the different states in which the dialogue is able to go. *Planning systems* are more flexible but they rely on particularly complex knowledge structures that are often difficult to design and to populate by knowledge authors.

In the case of this thesis, a planning system is chosen and extends existing research formalising rules to describe the process of argumentation (see section 2.1.2). In most of these task oriented systems, the hedging, social cues and other dialogue *fillers* are neglected as they are difficult to plan; to avoid adding complexity to the dialogue planning, the framework developed for this thesis introduces a novel approach to dialogue management, using non-task oriented techniques for managing social cues.

Non-Task Oriented

Non-task oriented dialogue management seems to have found little application domains even if it allows the users to interact with the system on almost any subject – i.e. it is open domain – when they do not want to get precise information.

These systems, on the other hand, seem to be better suited to simulate natural human-dialogue and pass the Turing test: systems like AliceBot[1] (Wallace 2004) or Converse (Levy, Catizone, Battacharia, Krotov, & Wilks 1997) have won the Loebner prize[2], which takes place every year and is based on the Turing test rules.

Like Eliza (Weizenbaum 1966), chatbots are mainly based on pattern matching techniques that generate a new answer based on the user input (see section A.0.2 for technical considerations). Chatbots might be suited to the type of social cues generation and relationship building needed in persuasive communication but as they always leave the initiative to the user and are only *reactive*, they do not implement any sort of planning. The fact that they cannot be used in applications where the user has to be led to accomplish the goals shared with the system is a limitation to their application domain.

Some systems (for instance, Eliza the psychotherapist (Weizenbaum 1966) or PARRY the paranoiac bot (Colby 1975)) are still able to simulate language and dialogue constructions that the users assimilate to human discourse. More recently, such *reactive dialogue management* systems are used on web sites to help the user find information on companies. Reactive systems are able to achieve consistent dialogues in domains where the dialogue context can be managed with a small utterances history window; for example, Quarteroni & Manandhar (2008) and Basili, De Cao, Giannone, & Marocco (2007) propose Question Answering (QA) dialogue management systems relying on purely reactive techniques

[1] `http://www.alicebot.org`
[2] `http://www.loebner.net/Prizef/loebner-prize.html`

to deal with the open domain dialogues created by QA sessions. Quarteroni & Manandhar manage the dialogue context through anaphora resolution and topic extraction in the previous utterances, limiting the need for a complex dialogue planning.

Reactive dialogue management systems rely on a large database of patterns to match all the possible inputs from the user. This knowledge base – even if it can be built by automatic learning from a dialogue corpus (Vrajitoru 2003) – like the ones in the task-oriented systems, needs a lot of development and tailoring to the particular subject that should be discussed in the dialogue.

The work presented in this thesis takes an hybrid approach where the dialogue management system is able to perform long term reasoning about the argumentation while staying reactive to the users. Stent (2002) presents an approach, towards such hybrid management, based on conversational acts. Stent uses the TRIPS dialogue management framework (Allen, Byron, Dzikovska, Ferguson, Galescu, & Stent 2001a) to divide the management task in two different sub tasks:

Discourse Reasoning takes care of understanding the user, reasoning about the possible reactions and generating tailored utterances to maintain the discourse context,

Problem Solving reasons about the task to achieve in the dialogue and the general behaviour of the agent. It is responsible for the long term reasoning independently from the discourse level planning.

The framework described in this thesis takes a very similar approach while tailoring the dialogue framework to the persuasive communication problematics and proposing another angle on this separation between the discourse level reasoning and the task level reasoning (see Chapter 3).

2.1.2 Argumentation in Artificial Intelligence

The logical analysis of argumentation – in particular to help in analysing, constructing and visualising legal argumentation – started the computer science community's interest in the topic of argumentation. Some interest has also been given in natural language processing to the automatic analysis of the argument structure (see section 2.3.2). Lately more interest is given to "natural argumentation" field where the argumentation is studied in the perspective of informal logic and Artificial Intelligence even if this research remains mainly theoretical.

Logic and Law

In the field of Artificial Intelligence and Law, *Argumentation Frameworks* (AF; Dung 1995) can be seen as a logical formalisation of the idea of schemes (see section 2.3.2). These frameworks use the classification of the argumentation processes to describe the possible transitions between each scheme.

An *argumentation framework* is used to describe each possible attack to an argument scheme and can be used to choose the best argumentation strategies to counter-argument the interlocutor. The framework forms an oriented graph linking the arguments together and can be used to verify the *acceptability* of an argumentation.

Bench-Capon (2003a) uses this technique to create persuasive argumentation, ensuring it can win the argument by performing graph analysis. To lower some of the complexity of this analysis, Bench-Capon extends the AF model to encode in the graph the *values* of the audience, which also renders the system more sensitive to the context.

A similar approach is presented by Greenwood, Bench-Capon, & Mcburney (2003) where argumentation is seen as a *practical reasoning* task. All the possible attacks on premises and conclusion are coded in logic language and an automatic reasoning can be inferred from them. Greenwood et al. also proposes an application of this system in legal reasoning and persuasion.

These logical approaches are of interest in the development of new planning and verification systems. These, however, do not consider most of the emotional and relational aspects of persuasion – even if (Bench-Capon 2003a) takes the audience values in account. The aim of this thesis is to go past the logic limitations of these systems to be able to include human aspects of persuasion.

The legal argumentation field has also generated research in computer-assisted argumentation within the field of human-computer interaction (Marshall 1989, Verheij 1999), but these systems are dedicated to the assistance of lawyers with interfaces that represent the structure of arguments to help them understand it. In the mentioned research, there is no will to reason on, or generate, arguments but only to develop effective interfaces which is outside the scope of this thesis.

Structure Analysis

The Rhetoric Structure Theory (RST; Mann & Thompson 1988; see section 2.3.2) introduces a formal basis for the analysis of the structure of arguments. Such structure can be represented as a tree (see Figure 2.9) with text segments as nodes and relationships between these segments on the vertices.

Within the field of natural language processing, the automatic construction of this structure can be seen as a case of text parsing, where the rhetorical tree is constructed. In Marcu (2000), the method proposed segments the text using lexical and syntactic rules combined with inference on the discourse markers, which are identified using part-of-speech tagging. For example, if the discourse marker "Although" is followed by a ",", then this part of the text can be considered as a segment. When the segments are identified, the author uses mathematical constraints describing the possible relationships between each segment to construct the possible rhetorical structures.

Soricut & Marcu (2003) also proposes an approach to discourse parsing, using probabilistic rules inferred from syntactic and lexical features. Soricut & Marcu first construct a syntactic parse of the text, extended by lexical features projected from the lexical head needed to further disambiguate the text boundaries given the lexical context. The parse is used to identify the different segments of the text, which are then linked together using a probabilistic model learned from a discourse-annotated corpus. Soricut & Marcu claim that the accuracy of this method can be compared to that of a human annotator; however, the accuracy of the segmentation phase seems to be low.

Corston-Oliver (1998a) presents a functional approach where the discourse parsing is not performed with probabilistic learning but uses a simple heuristic, constructed from observations and trials to rank the possible discourse relationships. Each relation is defined by a set of necessary criteria for it to exist and some cues that are evaluated to score the validity of the relationships in that context. Figures 2.1 and 2.2 illustrate an example of rules that are applied to compute the heuristic score.

---

1. Clause_1 is syntactically subordinate to Clause_2
2. The head of the subject of Clause_2 has the same base form as the head of the title of the section within which the Clause_2 occurs.

(From Corston-Oliver 1998b)

Figure 2.1: Criteria necessary for the **AsymmetricContrast** relationship.

- Clause_1 contains the subordinating conjunction *"whereas"*,
- *Heuristic score*: 30.

(From Corston-Oliver 1998b)

Figure 2.2: Cue to the **AsymmetricContrast** relationship.

This method is similar to the one proposed in Soricut & Marcu (2003) both rely on the syntactic parsing of the text prior to the analysis.

Both the methods could yield more errors in the persuasion process and are reported not to be accurate enough (Soricut & Marcu 2003). A more shallow approach to the identification of an argument's scheme is explored in section B. Nevertheless, the construction of an RST structure could be of interest in the generation of argumentation (for example, see Reed 1998 described later).

Natural Argumentation

Natural Argumentation is a wide field of research. It relates to the non-logical approach to argumentation proposed by the researchers in informal logic and following the work on rhetoric (see section 2.3.1). The dialogue reasoning is not only *"demonstrative"* and logical – like it would be in Mathematics (Aberdein 2005) – but also *"dialectical"* where some of the premises are not proved to be true but are still used to augment the acceptability of the conclusion (Perelman & Olbrechts-Tyteca 1958).

Some interest is also given to this field in computer science. Crosswhite (2000) proposes one of the first roadmaps to apply Artificial Intelligence to rhetoric and argumentation. Crosswhite emphasises the need for the development of argumentative frameworks that could model the different aspects of rhetoric. Crosswhite identified the following main fields of research:

"Techniques of argumentation" that relates to the use and implementation of the system of *schemes* introduced by Perelman & Olbrechts-Tyteca (1958) (see section 2.3.1).

"Rhetorical Figures" that relates to the ability to generate and use the different techniques for *presenting* the data – use of metaphors or periphrasis for example – to be persuasive.

"Audience as standards" deals with the different aspects of modelling of an audience.

Crosswhite also emphasises the technical problems that *dialectic* raises as it is mainly based on the *uncertainty* of the data presented, which a computer has difficulties to model.

Following these recommendations, researchers have proposed different solutions. For instance, Grasso (2002), Guerini, Stock, & Zancanaro (2003), Walton & Reed (2002) propose to tackle the first problem with different formal representations of schemes that could be used to plan argumentation. Their systems rely on knowledge representation of the acceptable fact, beliefs and values, of the system and of the user. However, in these theoretical approaches, Grasso, Walton & Reed, or Guerini et al. do not explain how the system models this knowledge. Neither do they look at the natural language generation of text based on this planning. To promote the development of works that deal with the problem of formalisation of argumentation schemes in Artificial Intelligence, Katzav & Reed (2004) proposes a new set of argument schemes that is grounded on the *intrinsic* properties of an argument.

Current Approaches to Natural Argumentation

Reed (1998) presents a system that can generate a *persuasive monologue*. This implies that the system cannot interact with the user during the generation phase and that the plan is fixed before the generation starts. Nevertheless, it gives an insight on the issue of generating arguments and it emphasises the need for a multi-layer system (see Figure 2.5):

1. The lower level is responsible for the Natural Language Generation (NLG) of the text that is left to another specialised system (Smith, Garigliano, & Morgan 1994). The important point in the approach by Reed is the use of the RST structure to describe the argument generated. However Reed stresses that the structure is not sufficient for generation and that the generator also needs to rely on information – i.e. ordering, language cues and other stylistic parameters – provided by the other layers of the system.

2. The *"Eloquence Generation"* (EG) is responsible for the planning the persuasiveness and the style of the output.

3. The *"Argument Structure"* (AS) layer combines with the latter and decides of the main components – premises, conclusions and scheme – of the argument.

The AS and EG layers are *"interleaved"* by the AbNLP (Fox & Long 1995) planner where planning operators are *"encapsulated"*. An operator to describe the *Modus Ponens* scheme is given in Figure 2.3 and a resulting plan is shown in Figure 2.4.

MP (H, P, X)
Shell: precond: $\exists$ X: X, (X $\rightarrow$ P)
$\overline{\text{BEL}}$(Ag, X)
add: BEL(Ag, P)
Body: goals: t0: PUSH_TOPIC(arg(X,P))
t1: BEL (Ag, X)
t2: IS_SALIENT(Ag, X, arg(X,P))
t3: BEL(Ag, X $\rightarrow$ P))
t4: IS_SALIENT(Ag, X $\rightarrow$ P, arg(X,P))
t5: POP_TOPIC(arg(X,P))

(From Reed 1998)

Figure 2.3: AbNLP operator for the Eloquence Generation layer.

The encapsulated planning allows the complete planning of the Argument Structure layer before its *"refinement"* by the Eloquence Generation layer to complete the planning of the encapsulated operators. However, after this *"refinement"*, some constraints could still be unchecked because of the partial order of the current goals. Therefore, the planning system performs a reordering of the *Argument Structure*'s goals and of the refined *Eloquence Generation*'s plan to construct a complete plan for the generation.

PUSH_TOPIC(arg(b,a))
BEL(h, b)
IS_SALIENT(h, b, arg(b,a))
BEL(h, b $\rightarrow$ a)
IS_SALIENT(h, b $\rightarrow$ a, arg(b,a))
POP_TOPIC(arg(b,a))

MAKE_SALIENT(h, a, _)

PUSH_TOPIC(arg(c,a))
BEL(h, c)
IS_SALIENT(h, c, arg(c,a))
BEL(h, c $\rightarrow$ a)
IS_SALIENT(h, c $\rightarrow$ a, arg(c,a))
POP_TOPIC(arg(c,a))

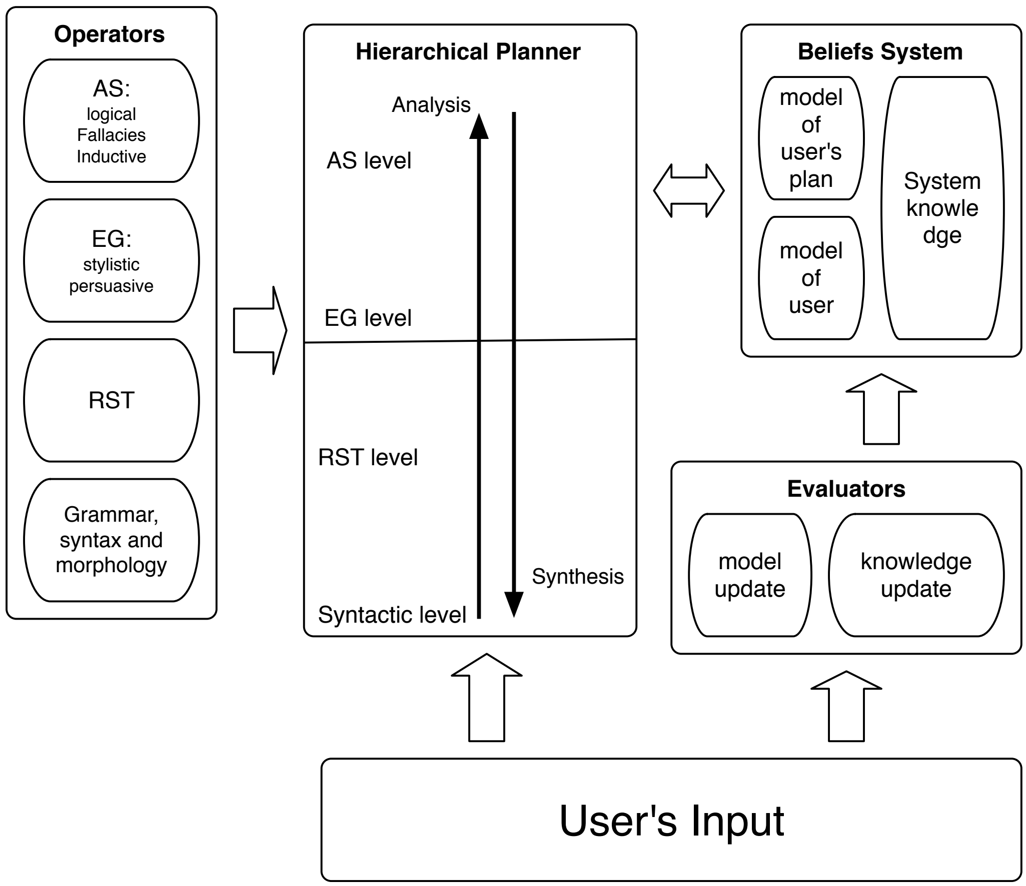
PUSH_TOPIC(arg(d,a))
BEL(h, d $\rightarrow$ a)
IS_SALIENT(h, d $\rightarrow$ a, arg(c,a))
BEL(h, d)
IS_SALIENT(h, d, arg(d,a))
POP_TOPIC(arg(d,a))

(From Reed 1998)

Figure 2.4: AbNLP ordered plan after the Eloquence Generation and Argument Structure layers.

Reed already tackles most of the issues that were raised regarding the planning of a persuasive discourse and uses the formalisation of argumentation schemes to generate a first plan that is refined by rules derived from stylistic and persuasive considerations into a monologue.

Reed's system is in fact not developed in the perspective of an interactive human-computer dialogue. Nevertheless, an approach to persuasive dialogue is discussed in the recently published state-of-the-art in natural argumentation (Norman & Reed 2003).



(From Reed 1998)

Figure 2.5: Reed's Architecture for Argument Generation.

Gilbert et al. presents a hypothetic "Persuasion Machine" and all the steps that it should follow to create a *persuasive dialogue*. Gilbert et al. describes a system similar to the one proposed in Grasso, Cawsey, & Jones (2000), the difference is that the latter does not give as much importance to understanding the user input. According to Gilbert et al. persuasive systems have to achieve a number of tasks to generate an answer to the user:

1. Analyse the user's utterance:

    (a) The system first has to *analyse the structure of the user's argument*, identifying which scheme is used, what are the premises and the conclusion. As seen before, this is a task mainly based on syntactic and lexical cues (see section 2.1.2).

    (b) It then has to evaluate each part of the argument, verify the facts used and the validity of the conclusion. This is a complex part of the process as it is difficult to check the validity of a fact:

    - How to be sure that the knowledge base of the system will know all the possible facts?
    - How to be sure that a fact that cannot be verified is false or just absent from the system knowledge base?
    - How to be sure that a falsehood is not true in the user's value system?

    Moreover, some premises can be omitted – i.e. in enthymeme arguments, premises are implied by the context but not explicitly stated in the dialogue – therefore, the validity of the conclusion cannot be inferred from all the supporting facts.

2. It then has to *update the user model* according to the beliefs and values observed in the user utterance. This implies the construction of a user model that is able to store the user's beliefs, values as well as the social norms influencing the user.

3. When this is done, the system can start constructing the next utterance; it has to *identify the next move*. This implies finding the possible move implied by the current plan and the last user's input; the system needs to be able to select the weak points to attack in the user argument but also to comply with the persuasive strategy it tries to accomplish.

4. The *selection of the next move* has to choose the right move to use amongst the possibilities compatible with the plan. To achieve this, the system must understand the repercussions of an argument as attacking a weak point could offend the user when striking a value related subject. The system must also be sure that it is be able, in the future, to support the argument it has put forward.

5. *Preparing the next move* implies the selection and of the presentation of the right premises to create the right scheme. The system has to be careful to present information that has not already been attacked by the user and present it in a persuasive way,

6. The *generation* of the new utterance involves the creation of human readable text.

A different approach is proposed by Carenini & Moore (2000a) to generate *evaluative* arguments. This type of argumentation uses the preferences of the users to create the rhetoric instead of relying on purely logical reasoning. Carenini & Moore propose to use additive multiattribute value function (AMVF) to reason about the users' preferences and generate a tailored argument. The authors use a value tree representing the users' preferences on entities and their attributes and apply user tailored AMVF to find the attribute to present according to their utility to the argument and to the user's preference values. Carenini & Moore thus present a non logical approach to argumentation reasoning that can help solve preferences and value issues in argumentation. Walker, Whittaker, Stent, Maloor, Moore, Johnston, & Vasireddy (2004) proposes a similar approach in tailoring the generation of evaluative arguments to the user's preferences. Walker et al. use multiattribute theory to filter the information presented to the users and select only the set of options – when doing recommendation and comparison – that makes sense to the user. Carenini & Moore (2001) shows that tailoring evaluative arguments to the users improves the persuasiveness of these arguments while Walker et al. presupposes that this improvement is due to the reduction of the information overload provided by the tailoring of evaluative arguments.

Corpora of Argumentation

Approach also aims to construct corpora of argumentation discourse.

Grasso (2003) is especially dedicated to the study of an annotation scheme for persuasive *health promotion dialogues* and proposes to divide a dialogue in three layers:

1. the *"meta-goal level"* describes the major goal of the dialogue. The persuasive goal that is to be eventually achieved,

2. the *"rhetorical level"* describes the argumentative goal of a segment of the dialogue,

3. the *"move level"* describes the different moves used in the utterances of the dialogue.

However, no corpus is currently available with this annotation.

Katzav, Reed, & Rowe (2004) provides a corpus of texts containing arguments, which is used in this thesis to attempt to classify arguments (see section B). However, it is not completely tailored to this thesis domain as it is generally composed of all types of discourses – mainly newspaper excerpts – where arguments can take very different forms than those of an interactive dialogue.

Conclusions

Argumentation is an active trend in computer science and more and more work is emerging on the subject that provides interesting inspirations for my research.

However, even if some of the current approaches (Bench-Capon 2003a, Gilbert et al. 2003, Reed 1998) try to go beyond the purely logic field, they seem to be forgetting the *emotional* aspect of argumentation. This could be explained by the fact that most of these approaches try to counter-argument the user; the user makes an argument and the computer has to analyse it and construct a counter-argument.

In the framework developed in this thesis, the task of persuasive dialogue generation and planning is led by the computer. Because it has the *initiative*, the dialogue manager can use argumentation strategies that leave the user with limited dialogue moves; the analysis of the user's answers is simpler and the achievement of a persuasive goal is easier to plan (see chapter 3).

2.2 Persuasive Communication

Persuasive Communication is a subject of research that spans over multiple fields, amongst which are social sciences and philosophy.

The philosophical approach to persuasive communication is mainly focused on the art of argumentation and is discussed in details in section 2.3.

This section describes persuasive communication as seen by sociologists and introduced in Stiff & Mongeau (2002). It will discuss the different aspects of persuasion and the models that have already been developed in social sciences.

2.2.1 Behaviour and Attitude

Attitude Change

Persuasion has long been defined as the act of changing people's *attitude* through communication. A possible definition of attitude is given by Rokeach (1968; p.

112) as:

> "A relatively enduring organisation of beliefs around an object or situation predisposing one to respond in some preferential manner."
>
> (from Stiff & Mongeau 2002; p. 12)

Persuasive Communication is therefore a set of messages that tries to change several beliefs ("organisation") that the auditor has cultivated over a certain period of time ("enduring").

The *attitude* represents one's bias of a situation. This judgement is based on either *"descriptive"* beliefs – that are verified by facts – or *"prescriptive"* ones – that are formed around the receiver's values, morals and ethics.

An attitude is therefore difficult to change as the set of beliefs forming such values, morals, and ethics are supported either by irrefutable facts or by positions that cannot be attacked in one's values.

Attitude versus Behaviour

Miller (1980) proposed a different definition of *Persuasive Communication*, in contrast with the previous readings:

> "Any message that is intended to shape, reinforce or change the responses of another or others."
>
> (from Stiff & Mongeau 2002; p. 4)

There are two important nuances in the above definition compared to the previous one; Miller does not refer anymore to *attitude*, but to the "responses of others" that does not only focus on attitude but focuses also on the other outcomes of persuasion: perceptions, emotions, beliefs and behaviours. Another difference is that *Persuasive Communication* is not only changing, but also "shaping" and "reinforcing" a response:

Response Shaping refers to the act of modifying one's behaviour by being aware of the positive and negative results of this behaviour for other persons. This can be regarded as social learning.

Response Reinforcement is most often accomplished by advertising. The persuasion process reinforces one's beliefs and values by providing new favourable evidences. For example, an advert tries to reinforce one's attraction to a product more often than it tries to create a response to a new product.

Beliefs, emotions and perceptions are part of one's *attitude*. However, the resulting *behaviour* does not always completely correlate with the attitude. For example, the attitude:

> "People dislike filling-in tax forms."

and the behaviour:

> "People file their tax return."

are both linked but not correlated. In fact, people's *behaviours* are composed by their attitude to an object or situation as well as the *"subjective norms"* that are associated to the social context (see section 2.2.2).

The weight of the agreement between an attitude and the resulting behaviour is determined by some known factors. First, the behaviour is more likely to be parallel with the linked attitude if the results are of importance in one's values. However, people with developed construct systems – developed representations of their experiences – can show a lower similarity between their behaviour and their attitude.

Influencing Behaviour

The theory of *cognitive dissonance* (see Festinger 1957) shows that it is sometimes possible to influence people's behaviour.

It uses the principle that people want to keep a certain *"cognitive consistency"*, i.e. they do not like to contradict their past behaviours and their social image.

Therefore, when *inconsistency* arises, people feel constrained and in such a state of mind, they are more motivated to restore the normality and resolve the inconsistency.

The motivation level is correlated with the *dissonance* level at various degrees amongst people. Indeed, Brehm & Cohen (1962) showed that the motivation was higher when the object or the situation leading to the dissonance was of importance to the subject. Furthermore, Aronson (1968) found that people permit fewer inconsistencies if they have higher self-concepts.

To diminish the cognitive dissonance, people are ready to change their behaviour or attitude and, sometimes unconsciously, they change the *cognitive elements* linked to the behaviour or the attitude in question. By doing so, they can, for example, integrate new information – new beliefs – or distort, even deny, existing ones.

Conclusions

Persuasive Communication is a field of research well studied in Sociology, which results in multiple studies of interest (see Stiff & Mongeau 2002) for the formalisation of persuasion for computer dialogue.

To summarise, these research points of view, the main goal in persuasive communication is the change of user *behaviour*. This is achieved through the shaping of the user's beliefs and perceptions on a subject.

This implies modelling such characteristics of the user as well as developing efficient strategies along the lines discussed above. A persuasive system has to be able to monitor the user's state of mind, its current beliefs and the ones that can – or have to be – changed in order to achieve the persuasion.

2.2.2 Models

A number of models have been developed by sociologists to plan the change of behaviour as well as evaluating its evolution.

Behaviour Evolution

Ajzen & Fishbein (1980) present the *Theory of Reasoned Action* that is designed to predict one's *intention* I_B to perform a particular *behaviour* B as a function $f()$ of one's attitude toward this behaviour A_B and of the subjective norms the behaviour is exposed to the subjective norm SN_B. Equation (2.1) – where W_1 and W_2 represent the personal importance attributed to each component – represents this influence:

The *attitude* (2.2) where b_i is the "*belief value*" and e_i is the "*evaluation*" of that belief,

The *subjective norms* (2.3) where b'_i is the "*normative belief value*". i.e the reference belief of the group the receiver considers himself in – and m_i the "*motivation*" to follow the group beliefs.

$$I_B = f(W_1 \times A_B + W_2 \times SN_B) \tag{2.1}$$

$$A_B = \sum b_i \times e_i \tag{2.2}$$

$$SN_B = \sum b'_i \times m_i \tag{2.3}$$

Elaboration Likelihood Model

The *Elaboration Likelihood Model* tries to predict the different ways someone can elaborate on a received persuasive message.

As shown in Figure 2.6, the model is quite simple; it states that elaboration on a message can be measured along a simple continuum spanning between the two extreme types of processing of a persuasive message:

- On one extreme are the messages that generate an active processing from the receivers. They get involved in a high cognition effort that makes them aware of the issue-relevant points in the message. This is the *"central processing"* where the real content of the message affects the receivers' attitude.

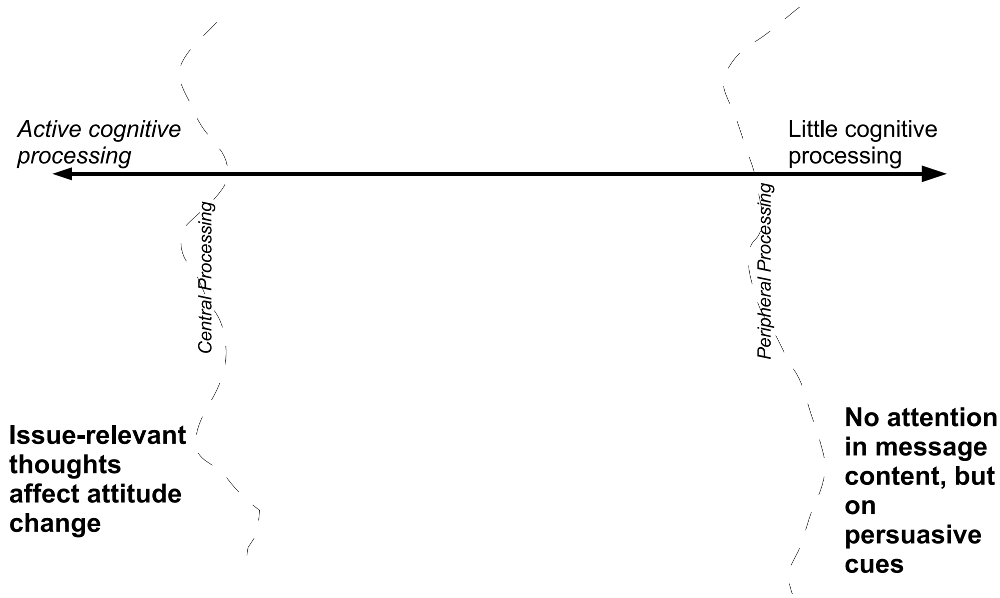


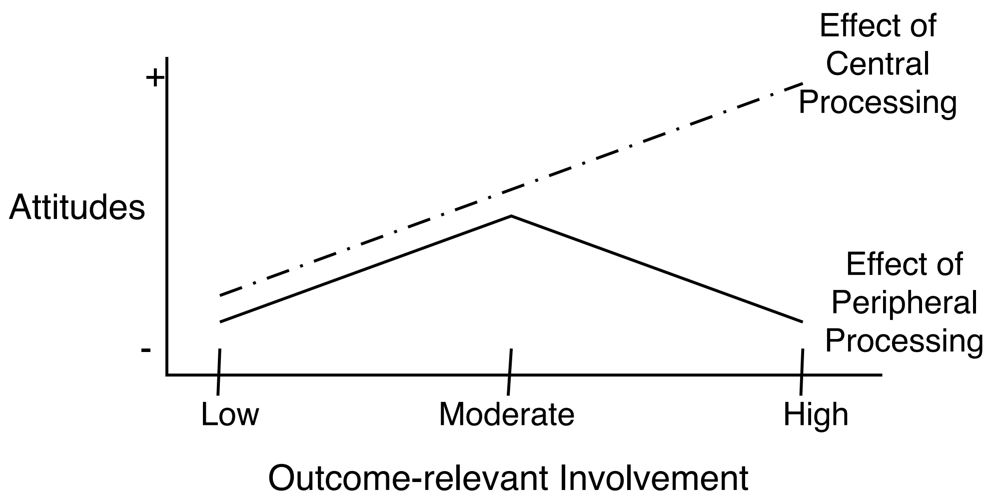
Figure 2.6: Elaboration Likelihood Continuum.

- On the other extreme of the continuum are the messages that do not imply cognitive efforts. The processing of the message is only done at the *"peripheral processing"* shallow level because there is no direct attention given to the message content. The main factor affecting the message is therefore its form and the persuasive cues it contains.

A few factors can influence the elaboration likelihood and help to predict the type of messages needed. *Central processing* is performed if the receiver is:

Motivated to process the message. This motivation, as mentioned before, is mainly linked to the "output-relevant" involvement of the receiver (see Figure 2.7) – if the person is interested in the result of the current discourse – but also by the social norms influencing the receiver.

Able to process the message. The receiver has to be capable of elaborating on the message content to achieve central processing. Therefore, messages that are difficult to process by the receiver produce less cognitive elaboration.



(From Stiff & Mongeau 2002; p. 225)

Figure 2.7: Cognition Effects on Attitude.

Moreover, people try to limit their cognitive effort as much as possible. Therefore, they tend to use *heuristic processing* instead of *systematic processing* of the message. People use simple rules resulting from their experience and their social education to elaborate on the message as efficiently as possible.

Compliance Gaining Strategies

Marwell & Schmitt (1967) as well as Wiseman & Schenck-Hamlin (1981) have developed limited sets (see Tables 2.2 and 2.1 respectively) of strategies to win the target agreement to a request; these strategies can be used in many situations but their relevance is dependent on the context and the actors of the discourse.

In fact, Sillars (1980) explains that the choice of strategies should be based on the importance of the target's compliance as well as social interaction rules.

Strategy	Definition
Ingratiation	Actor's offer of goods, sentiments or services precede the request for compliance
Promise	Actor promises goods, sentiments or services in exchange for compliance.
Debt	Actor recalls obligations owed him or her as a way of inducing the target to comply.
Esteem	Target's compliance will result in automatic increase of self-worth.
Allurement	Target's reward arises from persons or conditions other than the actor.
Aversive Stimulation	Actor continuously punishes target, making concession contingent on compliance.
Threat	Actor's proposed actions will have negative consequences for the target if he or she does not comply.
Guilt	The target's failure to comply will result in automatic decrease of self-worth.
Warning	The target's noncompliance could lead to circumstances in which other people become embarrassed, offended or hurt.
Altruism	Actor requests the target to engage in behaviour to benefit the actor.
Direct request	The actor simply asks the target to comply.
Explanation	Offer reasons for asking compliance.
Hinting	Actor represents the situational context in such a way that the target is led to conclude the desired action or response.
Deceit	Actor requests compliance by intentionally misrepresenting the characteristics of the desired response.

Table 2.1: Wiseman & Schenck-Hamlin compliance gaining strategies.

Strategy	Definition
Promise	If you comply, I will reward you.
Threat	If you do not comply, I will punish you.
Expertise (pos)	If you comply, you will be rewarded because of the "nature of things".
Expertise (neg)	If you do not comply, you will be punished because of the "nature of things".
Linking	Actor is friendly and helpful to get target in "good frame of mind" before making request.
Pregiving	Actor rewards target before requesting compliance.
Aversive Stimulation	Actor continuously punishes target, making cessation contingent on compliance.
Debt	You owe me compliance because of past favours.
Moral Appeal	You are immoral if you do not comply.
Self-Feeling (pos)	You will feel better about yourself if you comply.
Self-Feeling (neg)	You will feel worse about yourself if you do not comply.
Altercasting (pos)	A person with "good" qualities would comply.
Altercasting (neg)	Only a person with "bad" qualities would not comply.
Altruism	I need your compliance very badly, so do it for me.
Esteem (pos)	People you value will think better of you if you comply.
Esteem (neg)	People you value will think worse of you if you do not comply.

Table 2.2: Marwell & Schmitt 1967 compliance gaining strategies.

The effectiveness of a strategy is also linked to the cost and reward on the source/ target *relationship*.

In addition, Hunter & Boster (1978) with the "*Ethical Threshold Model*" emphasises the importance of the social appropriateness of compliance. People with a high *ethical threshold* – dependent on the context set by the intimacy of the interlocutors and the benefit of compliance – to comply more easily with *prosocial* strategies. For example, cynical persons (i.e. Machiavellians) have a lower ethical threshold as they see others as non-reliable and deceitful.

Social Learning Model

The *Social Learning Model* (see Stiff & Mongeau 2002) describes how people develop behaviours and attitudes by discovering their positive and negative outcomes in social interactions. The learning takes place in six phases:

1. Acknowledgement of the new behaviour or attitude.
2. Recognition of one's similarity with the observation
3. Identification of the outcomes.
4. Remembrance of the behaviour or the attitude.
5. Reproduction of the behaviour or the attitude.
6. Reinforcement of the model.

Stages of Change Model

Prochaska & Diclemente (1992) introduced the *Transtheoretical Stages Of Change Model* that models the change of the user's behaviour in five stages. An individual follows a path from the first stage – where a behaviour is not yet formed/modified – to the last stage – where the new behaviour is followed:

1. in the *"Precontemplation stage"*, the individual sees no objection to the current behaviour,

2. the *“Contemplation stage”* is when the individual recognises that there is a problem with the current behaviour,

3. in the *“Preparation stage”*, the individual is preparing to make some change in the “immediate future”,

4. the *“Action stage”* is the stage when the individual takes action to change the behaviour,

5. the *“Maintenance stage”* is the last stage when the individual has changed the behaviour and continues reinforcing it.

Prochaska & Diclemente propose different actions to influence the individual to progress from one stage to the other but remind that these stages are neither *“stable”* nor *“linear”*, so the individual could move back to a prior stage when encountering a problem with the new behaviour.

Like the *Social Learning Model*, this model is of interest to formalise the user’s state of mind in computer science (Grasso et al. 2000). Both models provide distinct *states* to describe the current behaviour of an individual, as well as *actions* to move from one state to the other. This can be used to model the user’s attitude towards a behaviour in a state-transition machine, which is designed to plan the future actions that the dialogue system can follows. This *Stages of Change Model* is not directly included in this thesis work as it is applicable to long-term persuasive tasks, which is out of the scope of the current research.

Health Behaviour Model

The *Health Behaviour Model* (see Stiff & Mongeau 2002) describes the main factors of persuasion in *health communication*:

- The receiver evaluates health behaviours in parallel with the evaluation of their personal health – done either by themselves or by an external source of advice.

- The persuasion efficacy is dictated by the perceived factors of threat to the person, severity of the health conditions tackled by the behaviour and its efficacy; the latter being influenced by the objective and subjective benefits and barriers perceived by the receiver.

Conclusions

The strategies described in these models are interesting in the research on human-computer dialogue as they give real insights on the intricacies of effective persuasion.

The concrete strategies from the last four models could be a starting point to design a dialogue planning system and select the right dialogue moves to achieve the persuasive goals. Actually, Grasso et al. (2000) uses the *Stages of Change Model* to model the current state of the persuasion in the dialogue.

In addition, the *theory of reasoned action* could serve as the root to the necessary user-modelling component of the system. Indeed, when planning arguments and deciding when the persuasion is accomplished, being able to evaluate the behaviour and belief change of the user is a challenge.

The *Elaboration Likelihood Model* stresses the need for rational planning of the dialogue extended by a reactive component. Indeed, Figure 2.6 emphasises the main scopes of this thesis:

- when the system needs to send important information to the user, it should use dialogue modalities that trigger the *central processing* of the data,
- when it wants to construct a relationship with the user, the system should trigger the receiver's *peripheral processing* by using social and persuasive cues.

These observations stress the need for a hybrid planning system that is able to deal with the achievement of the persuasive goal which presents the main data, but also with parallel goals of building a relationship.

2.3 Argumentation and Rhetoric

Aristotle (trans. 1998) started the study of the art of persuasive discourse a few centuries ago. Other schools of ancient philosophers – like the sophists – also taught *rhetoric* to help the Greek aristocratic class to argument and win prestige by the art of discourse.

From then, rhetoric was often misconceived as the art of formatting discourse to render it attractive and persuasive and was neglected by some. However, Aristotle's original intent with rhetoric was to influence emotions with language alone.

Today, an *argument* is often regarded only as the strong part of a discourse, where people use force – and contradiction – to try to impose their point of view on the audience. It is important to understand that argumentation is not just the task of using force in discourse. It is the act of presenting facts to an audience in a persuasive way.

Rhetoric and *argumentation*, in the Ancients' studies, go further than the use of force. These complex theories of the discourse came back in focus with the popular treatise of "The New Rhetoric" (Perelman & Olbrechts-Tyteca 1958).

2.3.1 Definitions

In the "New Rhetoric", Perelman & Olbrechts-Tyteca present the important notions of argumentation and this section summarises the basis of this theory.

> "The goal of every argumentation is the creation, or the augmentation of one's adhesion to the thesis presented to their approval."
> (Perelman & Olbrechts-Tyteca 1958, Vol. 1 p. 59, translation)

Arguing is not only a question of using a fancy "envelope" to present the facts; it is also – perhaps more importantly – the selection of the notions that support facts and the construction of strategies to organise these facts.

Toulmin (2003) presented an argument as *"premises"* – or *"data"* – that lead to a *"conclusion"* through a *"warrant"* and sometimes balanced by a *"refutation"* – or *"rebuttal"*. Figure 2.8 presents in a visual manner an argument as understood by this definition.

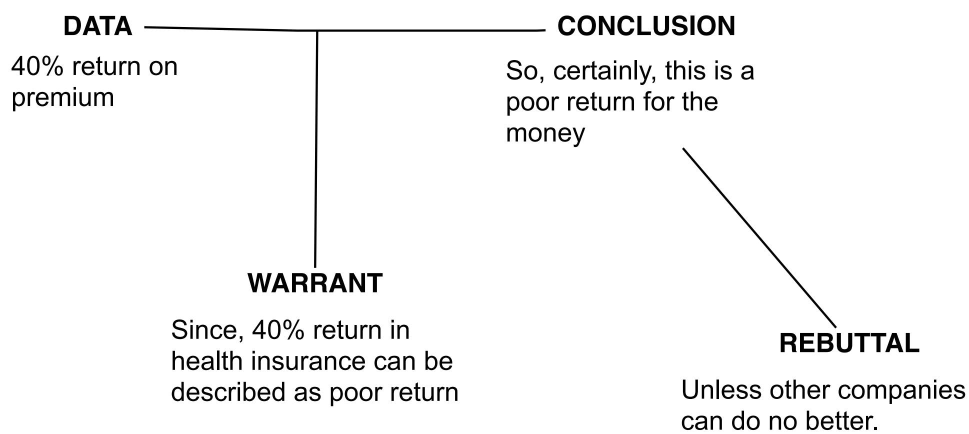


Figure 2.8: Toulmin's Argument Representation.

Facts, Values and Audiences

It is impossible to be sure of the *factuality* of an object in any circumstance and for any audience. However, both the audience and the speaker – or the writer – accept facts as either:

Observations, which are accepted by the interlocutors from real-world observations and that form the majority of the premises.

Suppositions, which are possible or probable facts that need to be discussed and accepted by the interlocutors before the argumentation.

In addition to these elements of the premises, the philosopher also gives importance to the *"values"* of the audience. Unlike facts, there is no universal agreement on *values*, but only small groups share values and agree on the related facts. Values always form part of the argumentation and cannot be rejected without leaving the domain of the discussion for that of force, which is not anymore persuasive communication but is coercion.

Selection of the Data

One part of argumentation for the rhetorician is, as mentioned before, the selection of the data to be presented.

> "It is not enough that something exists to have the sentiment of its presence. [...] Therefore, one of the preoccupations of the speaker will be to render, by the sole power of the verb, the presence of what is effectively absent, and he or she considers as important for the argumentation, or to give more valour, by augmenting their presence, to some of the elements offered to the audience's conscience."
> (Perelman & Olbrechts-Tyteca 1958, Vol. 1 p. 156, translation)

One element to take into account in this selection is the *"plasticity of the notions"*. Unlike Mathematics and its formal context where every notion is univocally defined, in the general discourse a notion cannot be tailored to one specific mental image and therefore produces ambiguities in the audience's interpretations.

In an argumentation, the orator takes advantage of this plasticity, and present the defended notions as flexible, and in opposition, the opponent thesis is presented as fixed.

> "flexibility of principles to new circumstances enables to support the impression that we keep alive the same notion."
> (Perelman & Olbrechts-Tyteca 1958, Vol. 1 p. 185, translation)

In an argumentation, the presented premises gain from the interpretations they generate, making the argumentation different from a simple demonstration. An effective argumentation plays with this freedom of interpretation to put the audience in presence of facts important for the orator and to leave the others aside.

Presentation of the Data

Even if it does not form the only part of an argumentation, the orator still needs to consider the presentation of the data. Since an argumentative discourse – more so a dialogue – is limited in the time, the construction of the argument has to take the amount of time into account and some notions have to be neglected to spend time on the ones that have to be reinforced in the audience's mind.

While *selection of the data* chooses which notions should be given importance in the discourse, *presentation of the data* decides of how the presence of the notions is created and which ones should stay implicit.

The presentation can use different strategies to format the data correctly. For example, the orator can use the different modalities of discourse:

- the *assertive* modality is the most common one as it is convenient for any argumentation,

- in the *injunctive* modality, the orator most probably uses – the rarely persuasive – imperative – "Go to the station!" for example. Indeed this modality is linked to force and persuasiveness will depend on the dominance of the orator, which, if absent, could make the use of injuctive sound like a plea,

- the *interrogative* modality can be used – and considered by the audience – as a trick. A question can easily be used to create an implicit accord on an object. For example, "what were you doing on the 11th in the station?" supposes the accord on the fact that the interlocutor was at the station this day.

- to express some notions, the use of the *optative* is recommended as wishing introduces an approbation regarding an object. For example "May he succeed" shows that the speaker is clearly in favour of someone's success.

Perelman & Olbrechts-Tyteca introduce other discourse strategies – like the *synonymy*[3] or the *periphrasis*[4] for example – but it would be too long to discuss all these modalities; to go further, Perelman & Olbrechts-Tyteca (1958) or Sandell (1976) are good reading – the latter proposes a complete study on linguistic style and persuasion.

Argumentation Schemes

Philosophers have also tried to take a more formal approach to the question of presentation of data by creating limited taxonomies of all the different strategies that are available for argumentation.

The classification of argumentative techniques was initiated in ancient Greece with the so-called *Topics*. Aristotle developed ordered lists of the different types of premise that could be used in each argumentation occasion – political discourses and judicial speech for example. Along the same line, Perelman & Olbrechts-Tyteca (1958) introduces the term of *argumentation scheme* to qualify the different strategies that can be used to link premises with a conclusion in an argument. Other criteria – such as the type of conclusion of an argument or the way the argument type should be evaluated (Garssen 2001) – can characterise these classifications.

The argumentation schemes are not based on highly pragmatic considerations and Perelman & Olbrechts-Tyteca insist on the fact that in their classification, it is possible that arguments fall in more than one scheme. Perelman & Olbrechts-Tyteca divide the arguments between the *"quasi-logical arguments"*, the *"arguments based on the structure of reality"* and the *"arguments that establish the structure of reality"*.

"Quasi-logical arguments" use the logic of the audience to link premises to the conclusion – by using the reciprocity and transitivity rules for example.

[3] where an idea is repeated using different words with identical meanings

[4] or *circumlocution*

The arguments *"based on the structure of reality"* are not based on rationality like the previous ones. However, they still use judgements on which there is accord to promote new judgements. According to Perelman & Olbrechts-Tyteca, this promotion does not generate questions about its validity, as long as the "solidarity" between the existing judgements and the ones that have to be accepted is created with sufficient confidence.

The arguments that *"establish the structure of reality"* are different from the previous ones as they link accepted *special cases* to create new rules or "*empirical realities*". For example, they create generalisation from examples.

Hastings (1963) proposes another detailed classification of argument schemes. Hastings bases the analysis on the Toulmin representation of an argument (see Figure 2.8) and provides a comprehensive collection of examples and their Toulmin analysis, thus classifying arguments by their reasoning process. An important difference with previous attempts for a classification is that Hastings verified that other persons were able to use consistently this classification in order to decide which scheme was used in an argument.

Conclusions

Rhetoric is not only the art of presenting facts to render them persuasive, nor is it a question of pure logic in the planning of arguments – as some of the theories discussed in the next sections present it; rhetoric is also the sensible selection of the premises and of strategies to link these premises to the conclusion. Indeed, Aristotle and Perelman & Olbrechts-Tyteca, amongst others, demonstrated that rhetoric is also the competence of an orator to understand and adapt to the emotions of the audience. The orator has to be ready to cope with the audience's values and emotions to render, in its conscience, the conclusions that are of importance to the argumentation. Perelman & Olbrechts-Tyteca insist that:

> "The form chosen for the data presentation is not only useful for its argumentative effects in the discourse; it can also offer expedients related to the communion with the audience."
>
> (Vol. 1 p. 220, translation)

2.3.2 Argument Structure

Continuing these considerations on argumentation schemes and the limited set of argumentation processes, some researchers have developed pragmatic approaches to describe the internal structure of an argument.

Arguments can be considered as more complex structures than the Toulmin representation which is used to separate the parts of an argument to understand the argumentation process. Whereas the Toulmin representation identifies the premises, warrants and the conclusion of an argument, the theories presented here are developed to study more precisely the relationship between the premises and the warrants and how they lead to the conclusion. However, these theories effectively base this study on the initial identification of these components, thus relying on the ability to segment the argument.

Rhetorical Structure Theory

The *Rhetorical Structure Theory* (RST) was introduced by Mann & Thompson (1988)[5]. This theory is based on the division of a text in pairs of segments that are linked together by specific relationships.

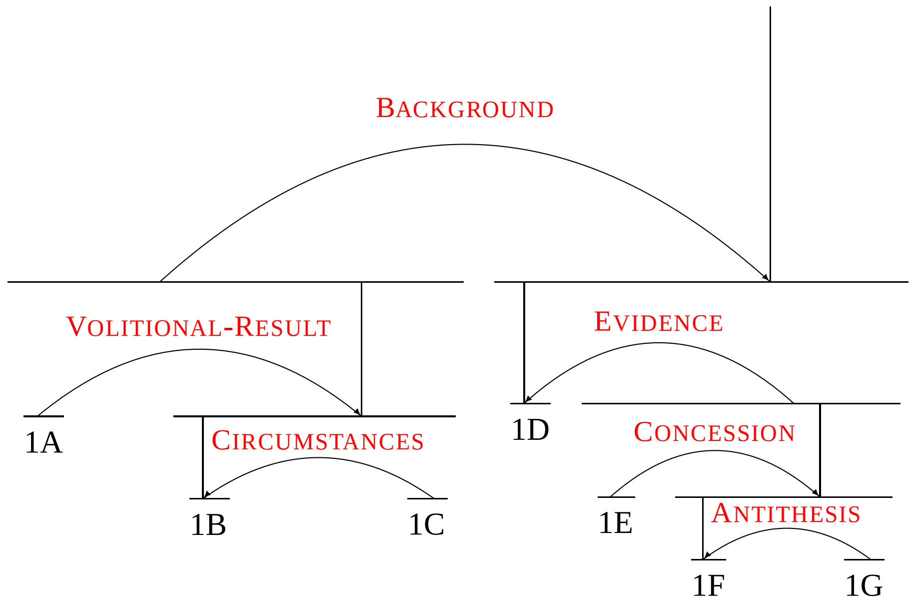
Each pair is composed of a *"nucleus"* and a *"satellite"*. The *nucleus* is the *affirmative* part of the two segments and the *satellite* is the *support* for this affirmation. For example, a *satellite* can be a demonstration, an elaboration or a preparation.

Mann & Thompson identify seventeen different *satellite/nucleus* relationships and a limited set of *multinuclear* relationships that are not making such

[5] see (Mann 1999) for an introduction

hierarchical distinction between the segments types they apply to, and represent, for example, lists or sequences of affirmations.

Figure 2.9 illustrates how RST relationships can be used to construct an analysis of an argument and encompass the complex relationships between each of its components.



[Farmington police had to help control traffic recently][1A] [when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.][1B] [The hotel's help-wanted announcement – for 300 openings – was a rare opportunity for many unemployed.][1C] [The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.][1D] [Every rule has exceptions,][1E] [but the tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a pay check illustrates a lack of jobs,][1F] [not laziness.][1G]

(The Hartford Courant [from `http://www.sfu.ca/rst/`])

Figure 2.9: RST Analysis of a text.

This representation has been criticised as it cannot represent parallel relationships and is sometimes limited by its simple tree representation of the structure of arguments. Some reviews of the state-of-the-art in the argument structure theories (for example Henkemans 2001) do not even acknowledge RST. However, this simple, pragmatic approach made RST popular in the field of computer science (see section 2.1.2).

Graphing

The analysis of the structure of arguments is generally taught in a manner that uses a less formalised view of the relationships. Most of the textbooks in argument analysis consider three different types of structure: *"serial reasoning"* , *"linked reasoning"* and *"convergent reasoning"*.

The approaches to analyse these structures have led to some disagreements in the field of argument analysis; some textbooks use a *structural* analysis – which only gives importance to the structure of the reasoning in the argumentation – whereas others prefer the *functional* analysis.

Walton (1996) takes a functional approach and gives importance to the functions of the structure in the reasoning. Walton produces a number of examples to explain the distinction between *linked* premises – that need each other to support the conclusion – and *convergent* premises – that support the conclusion independently.

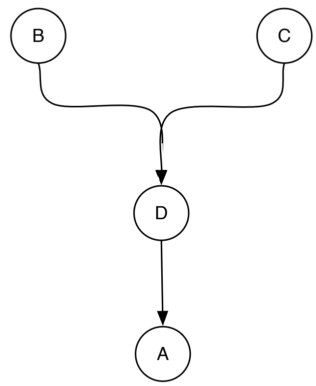
Walton also introduces a new graphing method to analyse the argument structure (see Figure 2.10) as well as tests to decide which type of reasoning is used in the argument:

Falsity/No Support Test If one premise is false, the conclusion is not given any support.

Suspension/Insufficient Proof Test If one premise is suspended (not proved, not known to be true), the conclusion is not given enough support to prove it.

Falsity/Insufficient Proof Test If one premise is false, the conclusion is not given enough support to prove it.

> [I think that active euthanasia, in the form of helping people to die, is something that will come to be accepted in the future.]^A [For when people become old or debilitated by illness, they may lack the means or strength to end their own lives.]^B [Such individuals may try many times, unsuccessfully to end their own lives, causing themselves and others great suffering.]^C [Therefore, the need to have assistance in ending terminal pain is becoming more evident.]^D



(Walton 1996, p. 93)

Figure 2.10: A graph analysis representing the premises supporting each conclusion.

Suspension/No Support Test If one premise is suspended (not proved, not known to be true), the conclusion is not given enough support.

Degree of Support Test Premises are dependent when together they make the overall strength of the argument much greater than they would consider separately.

Conclusions

Gilbert et al. (2003) states the need for the analysis of an argument structure in an argumentation dialogue system (see section 2.1.2); being able to determine the components of an argument, as well as the scheme used and the role of each part is a key in the construction of a counter-argument. In fact, according to Gilbert et al., a computerised system should be able to find the weak points of an argument to create an effective counter-argument.

Moreover, the structure analysis of an argument could contribute to other parts of the automatic argumentation process. For instance, it should give insight in determining if a dialogue utterance is an argument (see section B) and help in the generation of a well-formed argument (Reed 1998; see section 2.1.2).

2.4 Human Aspects

2.4.1 Characteristics of Persuasive Communication

Stiff & Mongeau (2002) identifies the different parts of a Persuasive Communication: the *source* produces an argumentative *message* for the *receiver*. Stiff & Mongeau have set restricted characteristics for the persuasion to be effective.

Source Persuasion is more efficient if the orator is *credible* and *trustworthy* to the audience.

However, some other aspects are less intuitive and difficult to recreate in human-computer dialogue. For instance, the *perceived similarity* between the orator and its audience is important when the issue discussed is relevant to this similarity. It has also been shown that, with a lower influence, *physical attractiveness* was an advantage in persuasion. However, it is difficult to recreate this last one within a textual dialogue system.

Message As mentioned before (see section 2.2.2), the processing of evidence in the message needs the receiver to be motivated but also *able* to process the message. The orator can use different strategies to ensure the interest in the message. For example, *rational appeals* can be used to use the implicit knowledge of the audience and apply logic rules. The speaker could also use *emotional appeals* that influence most of us – like, for example, the *fear appeal* that is often used in adverts on road safety.

Receiver In the *Social Judgment Theory*, the receivers are defined by their *latitude of acceptance*. If a message is accepted by the receivers in this latitude, then their attitude changes according to the discrepancy between the message and the receivers' positions. However, if this discrepancy is too high, then the message falls in the *latitude of rejection* and the attitude is not be affected. In which case, the source could lose credibility.

In the field of persuasive technologies, Fogg (2003) points out the same requirements and recognises that computers have some advantage over humans in persuasion. Amongst others, computers seem anonymous and ubiquitous which could be of interest in many persuasion processes where the user has to make concessions on personal matters.

Fogg also notes that people give credibility and trust more easily to computers as they are often viewed as *incorporating expertise*. However, this credibility is *presumed* and users assess it from *surface* characteristics of the system, such as the design or the layout on the screen.

2.4.2 The Media Equation

Reeves & Nass (1996) introduces the concept of *"The Media Equation"*. Following their study, Reeves & Nass conclude that the social responses of the human in a human-media interaction are almost similar to normal human-to-human interaction. Nass & Moon (2000) complete this observation in the field of human-computer interaction.

Nass & Moon demonstrate that it is possible that humans use social cues in their interaction with the computer – even if they are aware of interacting with a computer and that it cannot really feel emotions. However, they insist on the fact that these are *"premature cognitive commitments"* in what they call *mindless* interactions. This type of interaction with the computer results from learned social behaviours that are applied without the user having to think. However, it is important to note that these *"mindless" reactions* are only triggered by a minimum of *social cues* generated by the computer (like, for example, politeness).

Even if this concept is widely accepted and cited by reference textbooks (Fogg 2003; for example), some concerns have been raised by other researchers.

Shechtman & Horowitz (2003) studied the perception of human-computer interaction when users thought they were using the computer to chat with a real person and when they thought they were only interacting with a machine.

Two study groups were achieving the same task with the same program; however, one thought that the program was controlled by another human – which was false. This showed that people were keener to show group-related behaviours and engage in a relationship with human interlocutors.

Bhatt, Argamon, & Evens (2004) undertook the same type of studies and were able to show that people are showing less *hedged* responses with computer-believed[6] systems than with humans. In addition, the users are less polite and give less acknowledgement and feedback to computers. In a Wizard of Oz experiment, Brennan & Ohaeri (1994) shows that human interlocutors do not evaluate telegraphic language generations as less intelligent than anthropomorphic generations and that, in addition, their answers to simpler generations were easier to interpret by a computer system.

2.4.3 Conclusions

In general, the previously cited studies show that, in human-computer interactions, users are less inclined to keep the conversation going or build a relationship with the interlocutor. This was perhaps already recognised in Nass & Moon (2000) that stated that *mindless social response* occurs with a computer if one of these conditions is verified:

1. individuals erroneously believe that computers warrant human treatment,

[6]Systems that are directed by human in reality but are believed to be a computer by the test users.

2. individuals orient their responses to some human “behind” the computer,
3. individuals determine that the experimenter wants participants to exhibit social responses.

This question is of great importance in the present research as the persuasion process needs to build a relationship (see sections 2.4.1 and 2.2) with the user. This is why the dialogue interface designed in this thesis tries to use social cues and show empathy by improving the dialogue manager reactiveness.

Cassell & Bickmore (2002) applies these considerations in a dialogue system that simulates an estate agent. The system uses social cues in the dialogue to gain the user confidence, so as to be able to ask him personal questions. Cassell & Bickmore construct the system around a user model to monitor the relationship and did not use a conventional planning system but an *activation network* to be able to have mixed goals – relationship building and task achieving – in the dialogue. As recommended in Allen et al. (2001b), one of the hypothesis of this thesis is that separating persuasive goal planning and social cue management eases the authoring of the system and improves its persuasiveness.

Chapter 3

Persuasive Dialogue Management

3.1 Context for a Novel Framework

Persuasion through dialogue is a novel field of human-computer interaction. Bench-Capon, Atkinson, & Chorley (2005), among others, provides theoretical work on the formalisation and reasoning for logical argumentation with no application to dialogue management. Reiter et al. (2003), Reed (1998) and Carenini & Moore (2000b) also apply persuasive communication principles to natural language generation, but focus on monologue.

It appears from the study of the state-of-the-art of persuasive communication and argumentation that a different dialogue framework is needed to achieve persuasiveness. For the dialogue to be persuasive, the dialogue management system needs to guarantee the involvement of the user in the conversation as persuasiveness is strongly influenced by the perception of the user, which implies that trust, credibility and a sense of relationship have to be maintained by the system.

To achieve these requirements, this thesis sets the hypothesis that a novel method is needed to manage social cues and render the dialogue more natural to improve the persuasiveness.

A persuasive dialogue framework needs to achieve consistency and guarantee the completion of persuasive goals, this requires the ability to reason a priori about the dialogue path and what knowledge the system will introduce during the dialogue to argue with the user. The argumentation strategies have to be chosen to support the goals and fit the user's knowledge. The three-tier planner for tutoring dialogue discussed in Zinn et al. (2002) provides a dialogue management technique close to the requirements of persuasion planning: a top-tier generates a dialogue plan, the middle-tier generates refinements to the plan and the bottom-tier generates utterances. Mazzotta et al. (2007) also proposes a planning framework for user-adapted persuasion where the plan operators are mapped to natural language (or Embodied Conversational Agents) generation. The novel framework design proposed here also adopts such planning approach to manage the dialogue long-term goals and ensure the consistency of the argumentation.

The knowledge required by such planning systems becomes, however, difficult to manage if the framework needs to include a mechanism to react to the user's counter-arguments as well as social cues management. In fact, these reactive parts of the dialogue are inherently difficult to plan a priori. This observation leads to the design of a novel dialogue management framework for persuasion. The intent of this framework can be compared to the work presented by Stent (2002) where the author presents a dialogue management system based on the TRIPS architecture (Allen et al. 2001a) that proposes to separate the discourse reasoning from the problem solving.

In preliminary researches, different strategies have been studied to achieve persuasive dialogue. In particular Gilbert et al. (2003) proposes a road-map for the development of an argumentative machine, where the system needs a deep understanding of the user's argument content, which involves understanding the type – or scheme (see section 2.1.2) – of the argument to choose a tailored counter argumentation. One of the preliminary researches performed for this thesis tried to label text snippets with their argumentation scheme by using an automated classifier (see Appendix B). The classifier was trained on an annotated corpus, but the small size of the training sample as well as the complexity of the features that define the argumentation scheme showed that such deep understanding of the argument was not yet possible.

The persuasive communication guidelines as well as previous research in human computer interaction (Cassell & Bickmore 2002) show that there is a need for generating social cues and chit chat within the dialogue to improve its naturalness and persuasiveness. In another preliminary research, the hypothesis that chatbot systems could manage better these discourse level features was formed and a study of a chatbot knowledge-base structure was performed (see Appendix A). Both of these preliminary researches influenced the choices made during the design of the dialogue model presented in this chapter.

A novel framework design is proposed to provide the flexibility needed to react to the user and use social cues while guaranteeing the achievement of persuasive goals. The novel framework is referred in this document as the EDEN Framework.

When developing this framework, one main hypothesis was set:

> *Managing direct reactions to the user's utterances outside of the dialogue plan improves persuasiveness.*

This hypothesis was founded on the reasons described in the previous chapters as well as observations made during the iterative development of this framework:

- Persuasion requires the system to show empathy to the user. It needs to provide utterances for chit-chat and for empathising which cannot be planned a priori.

- When defining the planning component of the system, it appeared highly impractical for domain authoring to include all possible argumentation related reactions at the planning level (see section 3.4.1).

- Predicting the user's beliefs is impossible, in particular the counter-arguments used during the dialogue cannot be predicted without extended knowledge of the user. Relying on the planner for managing this part of the dialogue requires complex online planning or constant replanning.

The design of the dialogue management framework is also constrained by the experimental needs of computer science research. Validating the hypothesis of the thesis requires:

- Iterative development and testing of the framework.
- Ease of testing the framework on different domains.

These requirements led to the design of a modular framework where each independent layer could be replaced easily and where the domain authoring was separated from the dialogue management logic.

3.2 Case Studies

Part of the problem in evaluating persuasive dialogue is using an effective evaluation framework. Moon (1998) uses the Desert Survival Scenario (Lafferty & Eady 1974) to evaluate the difference in persuasion and trust in interaction when the impression of physical distance changes with the interlocutor.

The Desert Survival Scenario is a negotiation scenario used in team training. The team is put in a scenario where members are stranded in the desert after a plane crash. The participants have to negotiate a ranking of the most eligible items (knife, compass, map, ...) that they should keep for their survival.

For the evaluation of the dialogue system, a similar scenario is presented to the participants (see section 4.2). The user has to choose a preferred initial ranking of items and then goes through a discussion with the dialogue system that tries to persuade the user to change the ranking. At the end of the dialogue, the user has the opportunity to change the choice to a final ranking.

The desert scenario design has the drawback that none of the participants are experts in this domain nor do they have any particular preferences about which item to choose. A second experiment was thus designed to take into account user's preferences and study how to account for them in the dialogue framework.

The scenario used is inspired by a restaurant recommendation system, but instead of fully fulfilling the participants' preferences, the system tries to persuade them to change their mind. The task of the users is similar to the desert scenario: first rank a few restaurants according to their own preferences and then the dialogue management tries to change this ranking according to its own "imaginary" preferences – as if the user was chatting with another user – as well as factual information about the restaurants.

These scenarios and the experimental designs where they are used are detailed in the chapters 4 and 5. The following sections will describe the architecture of the dialogue system using examples from the Desert Survival Scenario domain and the Restaurant domain.

3.3 Dialogue Framework Overview

The EDEN Framework is designed as a modular system composed of four main components (see Figure 3.1):

The Knowledge Model is common in dialogue systems and is split here between:

- **The User Model** that stores the system's knowledge about the user.
- **The Domain Model** that describes the knowledge of the domain and how utterances should be generated for this domain.

The Reasoning Component is divided in two independent layers providing the tools to achieve the reactivity and continuity required by this work hypothesis:

- **The Long Term Reasoning** layer is responsible for keeping the dialogue on track and achieve the persuasive goals.
- **The Reactive Strategies** layer performs short-term reasoning and is directly responsible for the reactions to the user.

The Generation Module realises the natural language form of the content selected by the reactive layer.

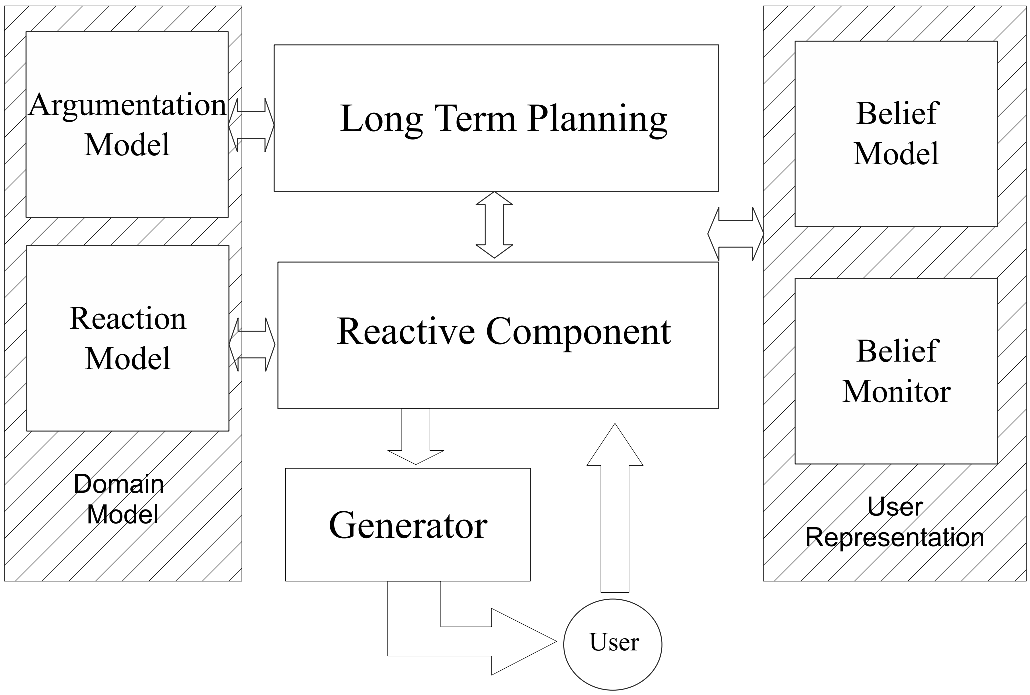


Figure 3.1: Mixed Planning/Reactive Framework

This split reflects the consideration of Allen et al. (2001b) to separate discourse level reasoning from the long term dialogue planning. Stent (2002) proposed a different approach to this problematic, extending on Allen et al.. In the EDEN Framework, a different approach is explored that is tailored to argumentative and persuasion considerations.

This approach to dialogue management is somehow different from standard layered approaches such as Zinn et al. (2002) as it introduces a new layer between the dialogue reasoning module and the generation layers[1]. The EDEN

[1]content planning, sentence planning and surface realisation.

Framework uses a planning component to select the dialogue content but delegates the exact utterance content selection to the reactive layer. The reactive layer adds flexibility to the plan and helps keeping a simple planning approach by keeping the discourse level reasoning out of the long term knowledge structure.

In Natural Language Generation (NLG) the standard approach splits the generation in three tasks: *text planning* – or *content selection* –, *sentence planning* and *linguistic* – or *surface* – *realisation* (see Reiter & Dale 1997). When approaching dialogue management, similar tasks must be performed, the system must decide of what to present to the user, then decide how to structure this content and finally realise it as natural language generations. However, a dialogue management framework also needs to manage the interaction with the users; thus the content selection and structure planning have to be able to adapt to the user's input.

Figure 3.2 shows the responsibilities of each layer of the framework for the selection and structuring tasks. The bottom generation layer is responsible for the utterance generation tasks of sentence planning and surface realisation. However, the content selection is shared between the reactive component and the generator as the reactive component decides of the structure of a dialogue segment and of the semantic content of each utterance but the actual content might be slightly modified at the generation phase. In a similar manner, the selection of content for the segments is shared between the planning component – that sets arguments to be discussed according to the dialogue structure – but also depends of the reactions of the user to that argument.

The approach taken in the EDEN dialogue management framework provides a simplified dialogue model for managing argumentation. Gilbert et al. (2003) proposes a roadmap for natural argumentation management requiring the understanding of the whole user argument, however, our preliminary research (see Appendix B) shows that this is currently impossible. Instead of trying to understand the full *content* of the user's inputs as proposed by Gilbert et al., the EDEN

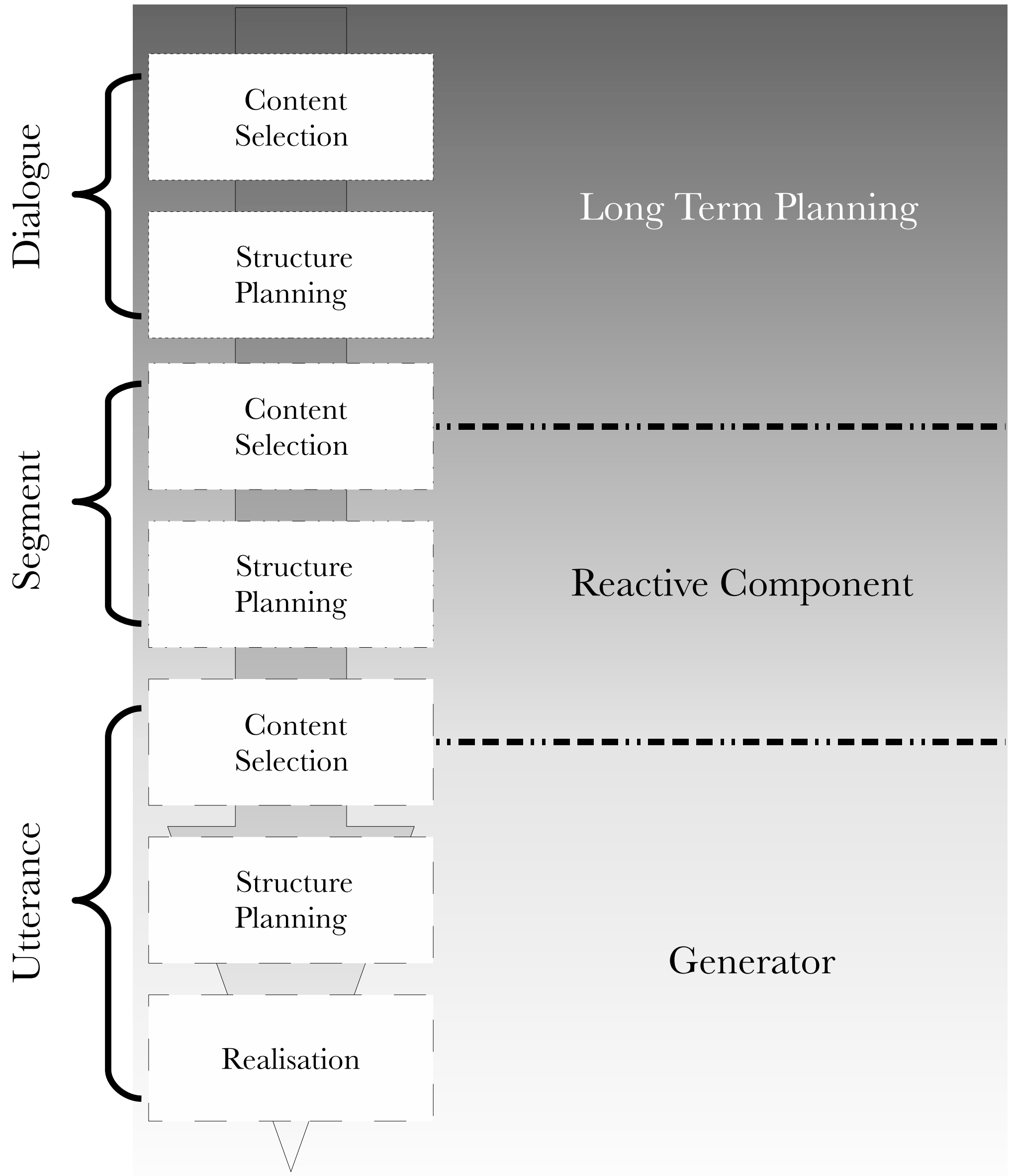


Figure 3.2: Division of the Content Selection, Structure Planning and Realisation tasks between the EDEN Framework layers.

Framework simplifies the dialogue model, while keeping a natural interaction, by trying to understand the *intent* of the user input (see section 3.9).

The rationale for this framework and details on how it manages dialogue are given in the following sections. Experimental validation of the fitness of such a framework to the persuasive task is provided in chapters 4 and 5.

3.4 Knowledge Model

Perelman & Olbrechts-Tyteca (1958) splits argumentation in two phases: the *selection* and the *presentation* of the data (see section 2.3.1). This split is reflected in the knowledge model that is divided in two modules:

the Argumentation Model holds the arguments that are used for persuasion. It is a raw representation of the data to be presented to the user, formalised for performing automated reasoning and *select* appropriate data (see section 3.4.1).

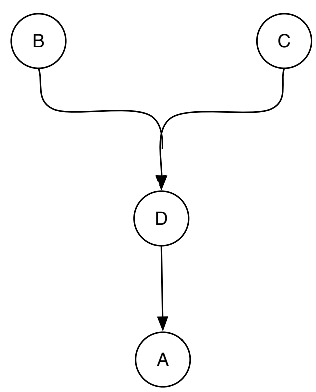
the Reaction Model is responsible for the *presentation* of the data and is parallel to the argumentation model as it represents the details of each argument and how to present them to the user (see section 3.7).

Figure 3.1 shows that this split of the knowledge model corresponds to the split of the dialogue management layers. The *Long Term Planning* component is responsible for the macro level *selection* of the arguments to present to the user to achieve the persuasive goals (see section 3.6) whereas the *Reactive Component* takes care of the micro level *presentation* to the user and the parts of the persuasive strategies that are independent from the data selected (see section 3.7 and 3.8).

3.4.1 Structure of Argumentation

Formal argumentation and its use in artificial intelligence have previously led to formalisations of the argumentation process (see section 2.3.2). Walton (1996) presents a functional representation of arguments that is well suited for computer formalisation. The argument's structure is graphed as relations between the premises. Premises are *linked*, *convergent* or *serial* depending on whether or not premises require each other to support a conclusion (see Figure 3.3 for an example). Section 2.3.2 explains how such structure is used to functionally evaluate an argument validity by considering the premises independently and their relationship.

> [I think that active euthanasia, in the form of helping people to die, is something that will come to be accepted in the future.]^A [For when people become old or debilitated by illness, they may lack the means or strength to end their own lives.]^B [Such individuals may try many times, unsuccessfully to end their own lives, causing themselves and others great suffering.]^C [Therefore, the need to have assistance in ending terminal pain is becoming more evident.]^D



(Walton 1996, p. 93)

Figure 3.3: Linked and Serial Premises Forming an Argument.

Dung (1995) and Bench-Capon et al. (2005) formalise this structure to logically define an argument as:

- the set of arguments attacking it
- the arguments it attacks

Dung then proposes a system that reasons about the validity of an argument over another argument through logical reasoning. Guerini et al. (2004) also formalises the persuasion process according to relationships between arguments but reason on the *conflict* and *support* relationships instead of using the *attack* relationship that Dung proposed.

For the task set by Dung, reasoning on attacks is easier as the validity of an argumentation is evaluated by asserting which attacks hold. As in the system proposed by Guerini et al. (2004), the EDEN Framework's task is to plan new arguments; indeed the dialogue is goal-oriented to achieve persuasive goals. In this case, two approaches can be used to reason on the dialogue process:

- Finding the argument that the system can use to *attack* the users' own arguments and win over their argumentation process,

- Finding the argument that the system can use to *support* its persuasive goals.

The *attack* approach is described by Gilbert et al. (2003) and implies that the dialogue uses mixed initiative: the user or the computer can present the initial argument and start the argumentation process; both can also introduce a new line of argumentation during the dialogue. The preliminary research presented in section B shows that it is computationally difficult to follow such a mixed initiative dialogue; indeed understanding the user's argument to then choose an appropriate attack is not yet possible.

To use *support arguments* consistently, the system must keep the initiative as much as possible to avoid entering in a complex argument/counter-argument cycle with the user.

In addition, relying solely on *attack* relationships to manage a dialogue becomes complex as reasoning on the long-term goals of the dialogue is rendered impossible. Indeed, to be able to plan a dialogue on a set of attack links, the planner needs to have a detailed knowledge of the user's beliefs. For example, as illustrated by Figure 3.4(b), if the system decides to attack the user's belief

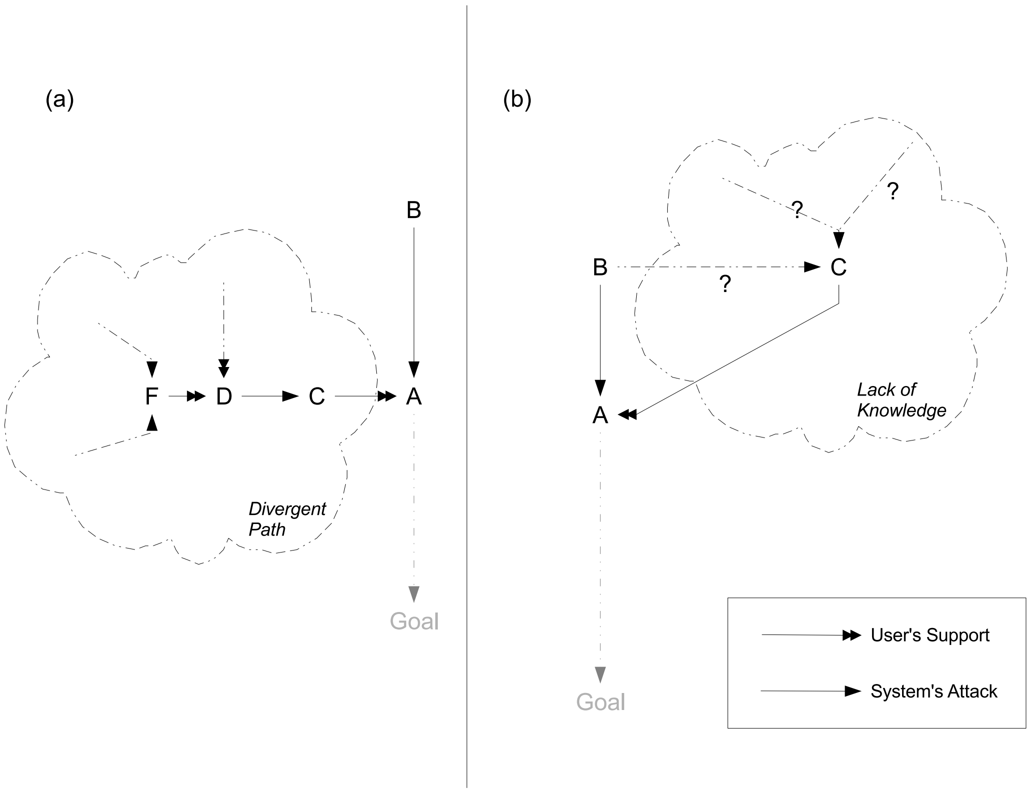


Figure 3.4: Drawbacks of an Attack-only Knowledge Base. (a) If the user defends a system's attack on the belief A with the belief C and starts arguing about C, the path of the dialogue diverges from the path to the system's goal. (b) If the system does not have sufficient knowledge of the domain, the user can introduce new beliefs on which the system cannot argue.

A with a known attack link from a belief B, the user can use a novel belief C, unknown of the system, to defend A. In this case, if the system only relies on attack links, it needs to know an attack on C or drop the goal of attacking A, since, even if the system presents more attacks on A, it cannot know if they are stronger in the user's values than the support link form C used by the user. In this example, two drawbacks of relying on attacks are illustrated:

1. The system needs to know that the user believes in A before attacking this

belief.

2. The user can always introduce a belief C unknown by the system.

Even if the system has an attack on C, the user could start arguing about C – forgetting about A – and lead the system outside of the argumentation path of the main system's goal (see Figure 3.4a).

If the knowledge model includes *support* links, it can introduce new beliefs to the users without having to know if they already believe it or not.

- If the system introduces a belief that the users already hold, then they agree and the dialogue continues.
- If the system introduces a belief and the user attacks one of the support links, the system can react by:
  - Introducing new supports for the attacked belief that the user might agree with.
  - Introducing a belief attacking the user's reaction to support the goal belief.

Attacks can still be used in combination with support links but, this time, as a reaction to the user's arguments.

Attacks are easier to manage as reactions to the user's counter-arguments as planning them a priori has drawbacks. In the design of the EDEN Framework, the *attack* relationship can be seen as having a *reactive* function in the dialogue. Meanwhile, the *support* relationship has both a *reactive* and a *predictive* function as support relationships can be used both to answer directly to the user's reaction and to plan the dialogue goals a priori.

Going back to the graphing method proposed by Walton (1996) (illustrated in Figure 3.3), the support links can be traced up from the conclusion A to the required arguments D, B and C. The dialogue management system needs to do the same: a dialogue is defined by one or multiple persuasive goals (conclusions) and the dialogue management has to find the required arguments (premises) to present to *support* the persuasive goals.

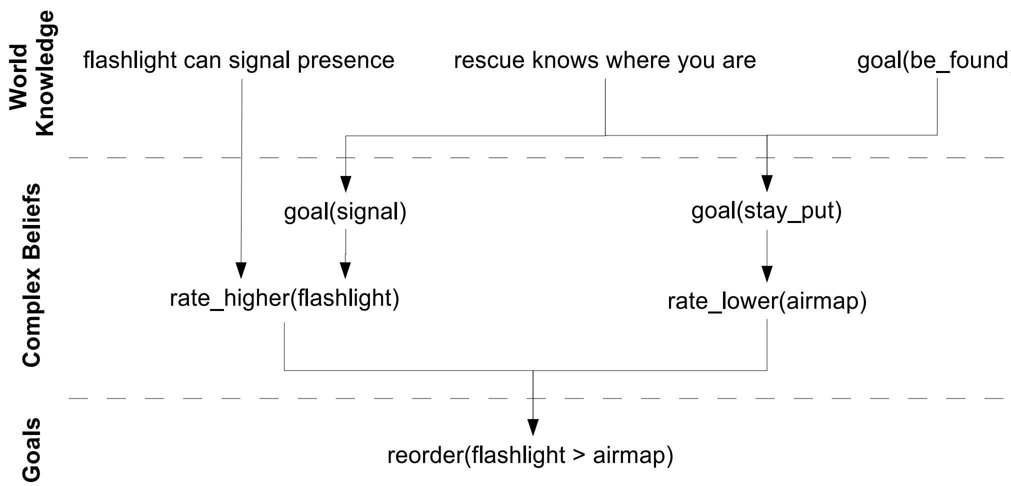


Figure 3.5: Sample Argumentation Hierarchy. The goal beliefs are supported by a hierarchy of beliefs that should be introduced to the user during the argumentation. At the top of the hierarchy are *world knowledge* beliefs that do not need strong defence with the user.

The argumentation model represents the argumentation process as a set of arguments each of them presenting a new belief (see Figure 3.5):

Nodes in the hierarchy formalise beliefs that can be used during the argumentation.

Links between the beliefs are *support* relationships representing a specific argument where the system can introduce a novel belief by using ancestor beliefs that the user already holds.

The dialogue manager tries to persuade the user by presenting new beliefs, either representing facts (e.g. `flashlight can signal presence`) or desired behaviours (e.g. `reorder(flashlight > airmap)`). The goal belief is the goal of the persuasion but cannot be introduced directly to the user; for the goal argument to have strength, the required beliefs must be presented first in the argumentation.

The first layer of the EDEN Framework is designed to extract a dialogue path from the hierarchy of beliefs. For example, considering the hierarchy given in Figure 3.5, the dialogue goal is for the user to have the belief that “reordering the flashlight over the air map” is preferable ($reorder(flashlight > air\_map)$). Because the dialogue manager does not know what the user already believes, the system must first introduce the linked beliefs higher up in the hierarchy to introduce this conclusion.

The *Long Term Planning* component’s function is therefore to find a satisfactory path in a graph. Carofiglio & de Rosis (2003) uses a *dynamic belief network* to model the consequences of an argument by using utility theory to compute the variation of emotions, combining the probability of achieving a goal and its utility. In this thesis a non probabilistic model is used, where beliefs are linked logically and inference is used to find the ideal path of beliefs to present. The approach chosen for the long term reasoning of the EDEN Framework is similar to the *missing axiom theory* presented by Smith (1992) where the dialogue manager uses theorem proving to find missing axioms (missing beliefs in the user model) to create dialogue segments that will allow to infer these axioms and prove a theorem (representing the system’s goal). In the EDEN Framework, the knowledge model can be seen as the set of missing axioms needed to prove the goal beliefs; however, the reasoning technique is different: Smith uses *interruptible theorem proving* and modifies the final theorem to include sub-dialogues; instead, here, the planner decides of a set of missing axioms to elicit from the user at the beginning of the dialogue; the system then knows it will need to clarify each of the beliefs represented by each axiom through an argument to know if the user accepts this new belief. The creation of sub-dialogues is thus delegated to the reactive component.

Three types of beliefs are described in the hierarchy (see Table 3.6 for examples):

Simple Facts or *world knowledge*, are beliefs representing a fact assumed to be accepted by all, which do not need a strong defence during the dialogue.

Facts are more complex beliefs at the top of the hierarchy. There is no encoded support in the hierarchy for each fact, but it might not be accepted directly by the user and requires defence.

Complex Facts need to be supported by other beliefs in the hierarchy before being presented to the user. A support relationship `support(`L`,`B`)` is composed on the left hand side of a list L of prior beliefs that are linked to support the right hand side conclusion B. Each belief of L taken individually cannot support B, however, in combination with the other members of L, these beliefs are strong enough to support the conclusion. A single $support$ relationship represents what Walton (1996) calls *linked premises* while two distinct $support$ relationships with the same conclusion encode *convergent premises*.

A complex fact can be encoded by a generic formalisation and the EDEN Framework applies logical inference to construct the hierarchy for this argument. For example

```
support([can(Y, X), goal(Y)], rate_higher(X)).
```

translates as: if an item X can achieve Y and Y is your goal, then rate the item X higher.

3.5 User's Preferences

In addition to planning the persuasion based on arguments describing the world with factual information (*descriptive* beliefs), the argumentation can also be

Simple Facts	
`simple_fact(is(lethal, item(knife)))`	The knife is lethal.
`simple_fact(can(signal, item(flashlight)))`	The flashlight can be used to signal presence to the rescue team.
Facts	
`fact(rescue_know)`	The rescue team knows the flight plan and will have a rough idea of your position.
`fact(goal(be_found))`	It is your goal to be found.
Complex Facts	
`support([goal(be_found)], goal(signal))`	As your goal is to be found by the rescuers, you should signal your presence to them.
`support([rescue_know, goal(be_found)], goal(stay_put))`	The rescuers know your position from the flight plan AND your goal is to be found, so you should not move away from the plane wreckage.

Figure 3.6: Examples of Beliefs in the Argumentation Hierarchy for the Desert Survival Scenario.

based on the user's preferences (*prescriptive* beliefs). Tailoring the dialogue to what the user believes allows applying the EDEN Framework in less factual domains where an argument can win, not because it makes logical sense, but because the user prefers the conclusion and premises over others.

Bench-Capon et al. (2005) presents a valued argumentation framework where arguments are associated with a personal value in which they hold. For example, an argument defending "abortion" could hold for someone with "pro-choice" values but would not for someone with the opposite values. Users' preferences are considered in this thesis as a type of values that do not span over a large hierarchy of arguments, and in the EDEN Framework, preferences are seen as beliefs held by the user as a simple fact can be held. Preferences are at the top level of the argumentation hierarchy. They do not need to be supported by higher knowledge and will not create complex argumentation phases during the dialogue.

	Caffe Cielo	Bobby Van's Steakhouse
Food Quality	decent	very good
Service	decent	good
Cost	moderate	expensive
Decor	mediocre	decent

Table 3.1: Example of Restaurant Attributes.

In the example of the Restaurant domain case study (see chapter 5), restaurants are defined by a set of five valued attributes: food quality, service, decor, cost, and food type. In the case of the two restaurants listed in Table 3.1, Caffe Cielo is better if looking only at the *cost* attribute but the steakhouse is better for the *service* and *food quality*. The choice of the user will thus depend on personal preferences when looking at restaurants. If the *food quality* is valued higher than the *cost*, then the steakhouse is considered better. However, if the user values *cost* over the other attributes, Caffe Cielo seems better.

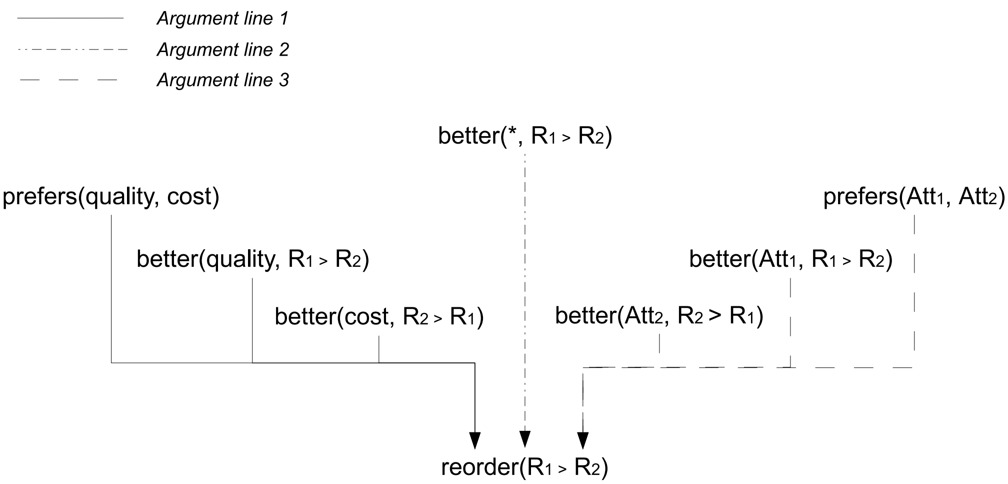


Figure 3.7: Argumentation Hierarchy with Preferences in the Restaurant Domain. The *argument line 2* shows a logical support between the fact that a restaurant is better than another and the goal of reordering these restaurants. The *argument line 1* shows a *prescriptive* argument where the user preference is used to solve the tie between the restaurant attributes. *Argument line 3* is the actual encoding of this preferential support in the domain model where variables are used for generalising the previous argument line to any arguments pair.

If preferences are encoded in the argumentation model as beliefs, the planner can reason in the same way as it does with the factual model and no modification of the planning model is required. Figure 3.7 is a sample of the argumentation model extended with preferences for the restaurant domain. The argumentation line 2 is a logical argument, the fact that all attributes of R_1 are better than the one of R_2 supports that R_1 should be ranked higher than R_2. The argumentation line 1 needs to use the user preference of *quality* over *cost* to support the reorder of the restaurants. The argumentation line 3 shows how this argument is actually generically encoded in the actual argumentation model.

This view of the preferences is a simplified model that can be integrated in the logical support hierarchy. However, it might not be the best to describe all the subtleties of preferences. User's preferences are usually not all of the same *potential* and this cannot be encoded by the simple *support* relation used in the EDEN Framework belief model. Carenini & Moore (2000a) proposes to encode arguments about likes and dislikes – known as *evaluative arguments* – through an *additive multiattribute value function* (AMVF) that encodes valued relations between attributes and entities to be discussed in the arguments. These valued relations can be used to encode and reason about the user's preferences.

3.6 Planning Argumentation

The planning component is responsible for the content selection and high level structure planning of the dialogue. The planner uses the Argumentation Hierarchy (see section 3.4.1) and the knowledge of the user's beliefs and preferences to select the arguments to present during the dialogue.

A graphplan (Blum & Furst 1995) algorithm is used to construct a planning graph from the belief relationships given by the argumentation hierarchy and to find the shortest available path in the hierarchy. The planner performs a search (see Figure 3.8) for a minimum set of ordered operations to go from the current belief state (what it knows about the user) to the goal belief state (what it wants the user to believe eventually). Each operation is an instance of a dialogue operator.

A dialogue operator specifies a dialogue segment (see section 3.8) and is defined in the planner by:

- The set of beliefs it depends on – i.e. what the user should already believe before using this operation.
- The new belief it introduces to the user.
- The existing beliefs it "removes" from the user[2].

Three operators are available in the framework to use the argumentation hierarchy nodes:

[2]These feature of the model is not used in the example domains.

`ground(`F`)` asserts a *simple fact* in the user's beliefs

> **Required Beliefs** none
>
> **Introduced Belief** F

`use_world(`F`)` introduces a *fact* to the user

> **Required Beliefs** none
>
> **Introduced Belief** F

`support(`$[B, C, \ldots]$`,` D`)` supports a new belief D with the existing beliefs B, C, $\ldots$.

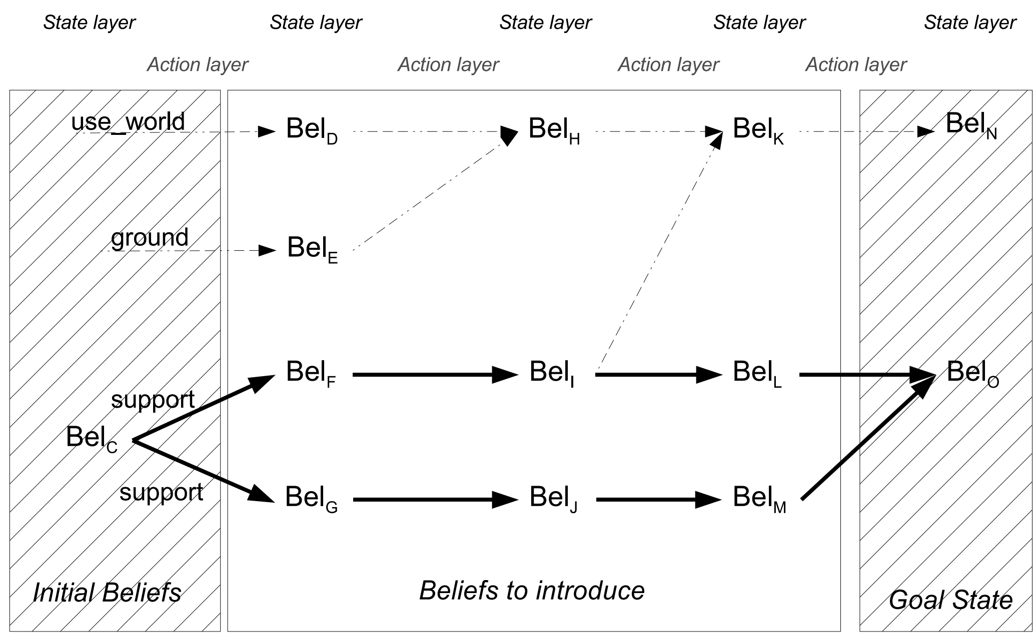
> **Required Beliefs** $[B, C, \ldots]$
>
> **Introduced Belief** D

From their planning definition, the `ground` and `use_world` operators are identical for the planner, they both logically do the same operation of inserting a new belief without support; however the expected reactions from the user during the dialogue is different and their realisation in dialogue segments differ. Figure 3.9 provides an example of a plan for the desert domain.

Strategies and Planning

The argumentation hierarchy is kept sparse for easing the domain authoring and the automated planning and therefore, no complex relationship between beliefs is encoded (see Appendix F for the Desert Survival Scenario). The planner relies solely on *support* links and is thus unable to order the operations in complex argumentation strategies.

For example, in the plan provided in Figure 3.9, the planner cannot determine from the operator's definitions how to group the operations. The plan provided by the graphplan is in two dimensions, as at each step of the plan, a set of operations can be applied. Each set is a collection of operations that have no conflicting prior requirements and thus can be applied in any order. The first phase



In this example, the planner is given the *goal belief* BEL_O. This belief defines the last *state layer* of the graph and the planner then iteratively adds layers to the graph in a backward search for operators that can be used with the current *state layer*. The chosen operators construct an *action layer* that links the current state to a candidate prior state.

The planner constructs a graph of interleaved state and action layers until it finds a state layer with all the beliefs available in the user's *initial beliefs*.

Note that the top part of the graph is an example of a path for another belief goal. Because the `use_world` and `ground` operators have no prior belief requirements, these operators can be used by the planner at any time and always match the *initial belief* state of the user.

Figure 3.8: Structure of a Planning Graph.

Initial Beliefs: [] **Goal Beliefs**: `reorder(flashlight, raincoat)`

Operator	Required Beliefs	Introduced Belief
	Step 1	
`ground`		`can(helpatnight, flashlight)`
`ground`		`not(are(breathable, raincoat))`
	Step 2	
`support`	`not(are(breathable, raincoat))`	`rate_lower(raincoat < flashlight)`
`support`	`can(helpatnight, flashlight)`	`rate_higher(flashlight > raincoat)`
	Step 3	
`support`	`rate_higher(flashlight > raincoat)` `rate_lower(raincoat < flashlight)`	`reorder(flashlight > raincoat)`

Figure 3.9: Example of a Plan in the Desert Survival Scenario.

ground	*SYSTEM:* We might have to stay stranded in the desert during the night. In the middle of the desert, there is no light and a flashlight could be useful. *USER:* right, I see.
ground	*SYSTEM:* The raincoat is probably made of a fabric that is not breathable as it's waterproof. In the desert, the heat will build up under this kind of fabric and we will sweat more. *USER:* Oh, yeah. I didn't think of that.
support	*SYSTEM:* As the raincoat is not breathable, it seems to me less useful than the flashlight, what about rating it lower? *USER:* …
support	*SYSTEM:* the flashlight is actually more useful at night, I would rate it highter. *USER:* …
support	*SYSTEM:* so, let's reorder the flashlight and the raincoat. *USER:* …

Figure 3.10: Dialogue without Flattening

of the planning is thus responsible for *selecting content* in the argumentation hierarchy and roughly structure it according to the precedence constrains provided by the hierarchy. This can be compared to the *content planning* phase in Natural Language Generation (NLG) but stays at a higher level as the structuring of the sub-dialogues for each argument is delegated to the reactive component as it depends on the user's reactions.

For the dialogue to sound natural, the sequence of steps must be topically consistent; for example, in the plan given in Figure 3.9, if the two operations of *Step 1* are performed consecutively, followed by the first operation of *Step 2* arguing about the *raincoat*, the first grounding about the flashlight seems out of place to the user as it is only needed for the second operator of the second step (see Figure 3.10). A *flattening* algorithm is used to reorder each step in a consistent manner according to argumentative strategies. For each dialogue goal, the *flattening* algorithm constructs a sub-plan out of the main plan, grouping topically related operations in the same sub-plan. In a second step, because some operations can be shared between dialogue goals, the *flattening* algorithm orders the sub-plans so that a required belief is introduced before all of the plan steps relying on it.

The second phase of the planning is thus responsible for further *planning the structure* of the argumentation to construct a dialogue session that sounds consistent and natural. This can be compared to the *content structure* and *sentence planning* phases in NLG, but is in this case related to *dialogue segments* and not actual sentences. The final realisation in natural language being left to the lower layers to dynamically adapt to the user's interaction.

When the plan is constructed, each operation is linked with the main goal(s) it defends during the dialogue. The operations are also labelled with the beliefs they require and the belief they introduce according to the operator they use. A bottom-up search is performed in the set of steps to split the raw plan in sub-plans (see Figure 3.11):

1. A first goal belief is selected and a sub-plan is constructed and annotated

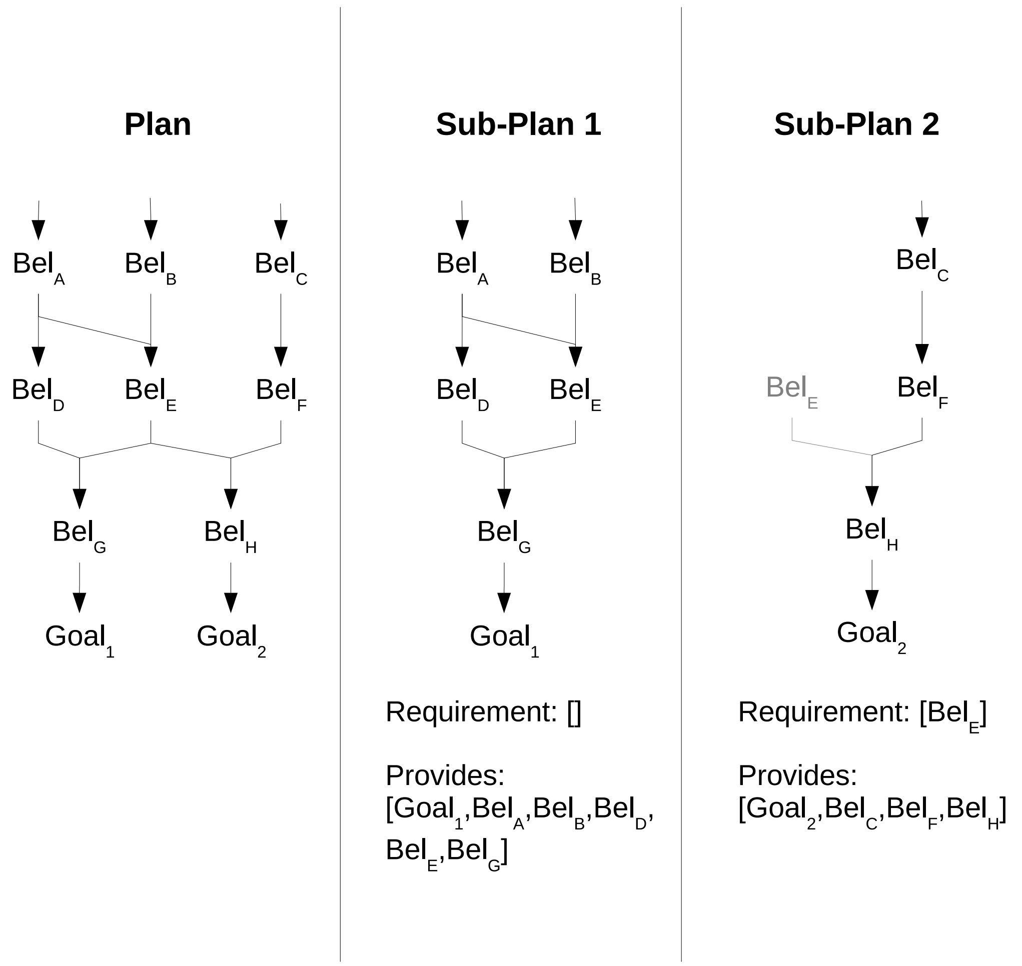


Figure 3.11: Sub-Plan Creation Example. The plan extracted from the argumentation hierarchy by the planner is two dimensional and has to be split in sub-plans relating to each goal. In this example, each Bel_X represents a plan operation introducing a belief and the arrows between each are the relationships of support between beliefs; the top level beliefs are introduced from world knowledge and have no support in the hierarchy (see section 3.4.1).
The separate argumentation processes of each persuasive goal can share some premises and sub-plans are annotated with the premises they introduce and the one they require – e.g. Sub-Plan 2 requires Bel_E provided by Sub-Plan 1. The resulting sub-plans are then ordered so that all the beliefs required by a sub-plan are introduced by a prior plan. In this example, the ordering will be Sub-Plan 1 $>$ Sub-Plan 2.

with this goal as requirement.

2. The plan is then searched upward, iteratively adding all the operations that provide the required beliefs,

   - when an operation is added to the sub-plan, the latter is annotated with the beliefs provided by this operation and the required beliefs annotation for the sub-plan is updated to include the requirements of this operation.

3. At the end of the search the first sub-plan provides a set of beliefs but has no requirement as all the required operations have been selected in the first pass.

4. A second goal is then selected to construct another sub-plan and a similar process takes place.

5. The first sub-plan might have already blocked operations providing some of the beliefs required by the second sub-plan, and the end of the search finishes with a sub-plan with unfulfilled requirements.

6. An "incomplete" sub-plan is created in the same way for all of the persuasive goals.

7. Each sub-plan is then ordered so that a sub-plan requiring a belief is placed after the sub-plan providing this belief. Each sub-plan P_i is defined by the set of beliefs it provides $Prov_i$ and the set of beliefs it requires Req_i, the partial ordering of two sub-plans is defined by Equation (3.1). If more than one partial ordering is possible, one of them is selected randomly.

$$P_1 > P_2 \rightarrow Prov_1 \cap Req_2 \neq \varnothing \tag{3.1}$$

The resulting plan is an ordered set of two-dimensional sub-plans regrouping topically related operations.

The final plan for the dialogue should be one-dimensional and each sub-plan should be flattened in a sequential plan where the order of the premises introduction appears natural. A partial ordering is applied inside each sub-plan to order the operations so that premises are introduced close to the argument that requires them. Each operation Bel_{op} is defined by the set of beliefs it requires Req_{op} and the belief it provides (i.e. introduces to the user) $Prov_{op}$. The partial ordering of two operations is defined by Equation (3.2).

$$Bel_1 > Bel_2 \rightarrow Prov_1 \in Req_2 \tag{3.2}$$

In the example of Figure 3.11, the Sub-Plan 1 has two possible orderings:

$$Bel_A > Bel_D > Bel_B > Bel_E > Goal_1$$
$$Bel_B > Bel_E > Bel_A > Bel_D > Goal_1$$

while the Sub-Plan 2 has one ordering:

$$Bel_C > Bel_F > Goal_2$$

One of the possible order for the Sub-Plan 1 is randomly selected and, with the given ordering of Sub-Plans, the sequential plan is:

$$Bel_B > Bel_E > Bel_A > Bel_D > Goal_1 > Bel_C > Bel_F > Goal_2$$

In addition, by applying heuristics to each sub-plan inner order, strategies can be implemented for the introduction of each premise. In particular, a simplification heuristic is used to prune the `ground` operations and leave the introduction of the *simple facts* to the first `support` relying on them. In the previous example, if Bel_B is introduced by a grounding operator, the plan can be simplified as:

$$Bel_{B+E} > Bel_A > Bel_D > Goal_1 > Bel_C > Bel_F > Goal_2$$

where Bel_{B+E} is a single dialogue segment introducing the beliefs B and E together.

Figure 3.12 illustrates the final realisation of the plan steps given in Figure 3.9. The *grounding* operators have been attached to the support they were relevant too, and as they do not require a different argumentation, they are merged with the *support* steps to create one single dialogue segment.

The final plan provides the topical path for the dialogue, presenting the dialogue segments that should be used to generate utterances and to react to the user counter-arguments.

3.7 Reaction Model

The Planning Component selects arguments and provides a structure for the segments of the dialogue. This content ordering will not be changed in the rest of the dialogue unless the selected plan fails. The Reactive Component is responsible for planning and structuring the content of each dialogue segment according to the user's reactions and the communicative goals provided in the plan structure.

The operations provided by the planning component represent dialogue segments topically focused on one argument to introduce a specific belief. The second layer of the framework, the *reactive component*, is responsible for generating as many utterances as necessary to persuade the user about this new belief.

The main difference of the EDEN Framework with state-of-the-art planned – or state oriented – approaches (for example Mazzotta et al. 2007) is that each plan step is not directly mapped to one utterance generation. The selected operations are used as constraints to the possible reactions used for the defence of arguments during the dialogue. Indeed, the planner decides on the main dialogue path but cannot decide on how to react to the user's counter-arguments. The Reactive Component manages the reactions to the user utterances and can use strategies constrained by the plan but that could not have been planned without prior knowledge of the user's beliefs and values.

The planner is responsible for the high level dialogue structuring and content selection by creating consecutive communicative goals. The realisation of these goals in natural language utterances is delegated to the lower layers of the EDEN Framework:

1. The *reactive layer* has the control over the structure of the dialogue segments. The user can react in many ways to the argument presented in each segment, the reactive component uses the constrains provided by the communicative goals to react within the argument by activating context specific reactions. The reactive content is thus able to select content for each specific utterance tailored to the argument and to the user's reaction.

2. The *generator* is then responsible for the natural language realisation of the content selected by the reactive component.

This final generation of natural language utterances can be performed through Natural Language Generation (NLG) techniques as is discussed in section 3.7.3. However, to simplify the understanding of the reactive layer activation model, the following section simplifies the generation problem by discussing a simpler template engine.

3.7.1 Activation Tree

The reactive component contains a domain specific database of argumentation strategies and directives for the generation of the raw beliefs provided in the argumentation hierarchy. This database is a specialised search tree presented in Appendix A.0.2 and Figure 3.13. Each plan step can be mapped to a sub-tree representing the available reactions for a specific argument (see Figure 3.13). Each leaf of the tree is a possible reactive strategy, specifying generation directives and updates to the belief model that should be used in the dialogue context defined by its branch in the tree.

Figure 3.14 provides an example of the reactions available for the operation `use_world(goal(survive))`. These three reactions can be used to match the user's utterances during the dialogue segment corresponding to this plan operator: *Reaction 1* is not bound by a context and is available as the initiative-taking generation. It is the first utterance to be used when the system comes from a previously finished step as it can match any user's utterances and context. *Reactions 2-3* are defences that support *Reaction 1*. They are used to react to the user if no agreement is detected from the last utterances.

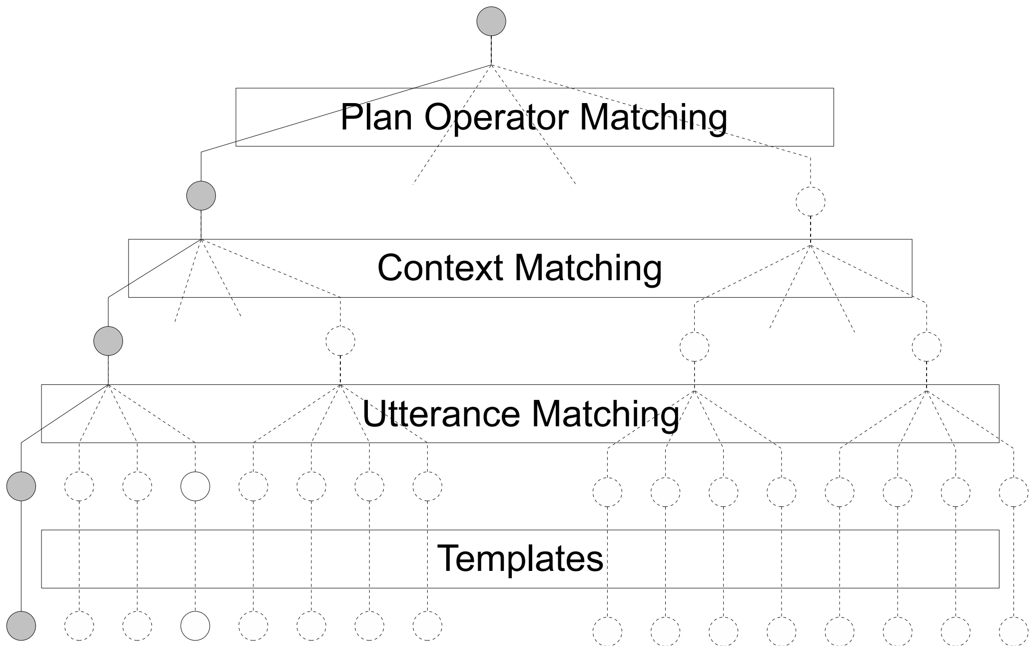
Figure 3.15 is an example of a dialogue segment using these reactions. A disagreement is detected to the *Reaction 1* and the reactive component uses *Reaction 2* as a contextualised answer.

support
SYSTEM: The raincoat is probably made of a fabric that is not breathable as it's waterproof. In the desert, the heat will build up under this kind of fabric and we will sweat more. What about rating it lower?
USER: ...

support
SYSTEM: We might have to stay stranded in the desert during the night. In the middle of the desert, there is no light and a flashlight could be useful. I would rate it highter.
USER: ...

support
SYSTEM: so, let's reorder the flashlight and the raincoat.
USER: ...

Figure 3.12: Dialogue sample after Flattening. The algorithm reorder topically relevant utterances and then merge content that might not need extra argumentation (i.e. grounding operators).



The *Reactions* (see examples in Figures 3.14, 3.16, 3.17, 3.18) are parsed into a search tree similar to the one in Figure A.2 where a top-down search is performed to find the best utterance to use for the current plan step and context. The pruning of possible reactions for the current dialogue stage is thus similar to a pattern matching system.
Each node of the tree represents one input token or a wildcard matching as many input tokens as possible. The tree is divided in four layers: a branch of the *Plan Operator* layer defines a matching pattern used to match dialogue plan steps. When a plan operator is matched, a sub-tree in the *Context Matching* layer is selected. This tree defines a set of patterns matching the last utterance of the system, setting a context for the reactions. Matching a specific plan operator and dialogue context selects a sub-tree from the *Utterance Matching* layer. The latter defines reactions to particular inputs by the user and is linked to leaves representing generation templates.

Figure 3.13: Global View of the Matching Tree.

Plan operator: `use_world(goal(survive))`

Reaction 1 :

Utterance Matching *

Template Surviving is our priority, do you want to hear about my desert survival insights?

Reaction 2 :

Utterance Matching * insights

Context Matching * survival insights

Template I mean, I had a few ideas ... common knowledge I suppose.

Reaction 3 :

Utterance Matching *

Context Matching * survival insights

Template Well, we are in this together. Let me tell you what I think of desert survival, ok?

Figure 3.14: Sample Reactions in the Desert Survival Scenario for `use_world(goal(survive))`.

SYSTEM: Surviving is our priority, do you want to hear about my desert survival insights? [*Reaction 1*]
USER: what kind of survival insights??
SYSTEM: I mean, I had a few ideas ... common knowledge I suppose. [*Reaction 2*]

Figure 3.15: Dialogue Example with Reactions from Figure 3.14.

3.7.2 Genericity of the Activation Engine

Plan Operator: `support([`VAR_1`],ratelower(`VAR_2`,airmap))`

Reaction 4 :

Utterance Matching: *

Template well, I think VAR_2 is less useful than the air map, that we could use to start a fire. What about rating it lower?

Reaction 5 :

Utterance Matching *

Context Matching * that we could use to start a fire

Template I am not sure how we would ignite the thing, but we might find a way. We have to be inventive.

Figure 3.16: Generic Matching in the Desert Survival Scenario.

In addition to matching the user input and the system latest utterances, the search tree can also match plan operators in a generic manner. Figure 3.16 is an example of part of the matching tree for the desert domain. This tree can be used for any communicative goal of the form: `support([...], ratelower(..., airmap))` which supports the belief that the "airmap" item should be ranked above some other item, with the argument that "the map can start a fire" which does not need explicit comparison with the second item and thus can be applied to any item of the domain.

- *Reaction 4* is a generation for a main argument supporting that belief. VAR_2 is a variable matching the second item name in the current plan operator. This variable is replaced in the *template* when constructing the utterance by the value matched from the user's input.

- *Reaction 5* is a defence for the *Reaction 4*. If the user has a negative reaction to the argument about "starting a fire", the system can defend its idea with another argument.

Initiative, Supports and Attacks

Reaction 1 and *Reaction 4* are good examples of the reactive component's initiative taking design. Perelman & Olbrechts-Tyteca (1958) details a set of discourse modalities for presenting the data (see section 2.3.1), in the case of the EDEN framework, most of the generations are formatted in the *interrogative* modality. This creates the impression that the system takes the initiative and the user tends to answer with utterances formulated as agreement or disagreement, sometimes completed by an elaboration on the disagreement.

Reactions 1 and *4* are *support*s for the current belief. As explained earlier, the hierarchy of beliefs encoded in the argumentation model is composed of *support* links defining known arguments leading to the persuasive goals. At the beginning of the dialogue the system has no prior knowledge of the user's beliefs and there is no way for the planner to foresee the use of *attack* links.

Plan Operator: `support([*], ratehigher(*, compass))`

Reaction 6 :

Utterance Matching *

Template Why would we need a compass? If the daily temperature is 35°C, walking would require twice the quantity of water. Really, the temperature in the desert is surely higher than that, we don't have much water, I think we really shouldn't move.

Reaction 7 :

Utterance Matching * find a road *

Template if we move, there will be no chance that the rescue team finds us. I just don't see the point in rating the compass that high.

Figure 3.17: Reactive Attacks in the Desert Survival Scenario.

In the EDEN Framework, the *attack* components of the argument structure are managed directly in the reactive layer of the system. In the current implementation of this layer, this means that the attacks are encoded in contextualised counter-reactions linked to a main support. Figure 3.17 is an example of such a reactive attack in the desert domain. *Reaction 6* and *7* are attacks on the user's belief that it is "useful to move"; both present counter-arguments directed against a particular belief presented by the user and are not generic supports to the main argument as is *Reaction 5*.

Knowledge Authoring

Authoring the dialogue domain and all the possible generation and reactions can be a tedious task. Writing tailored arguments to every possible disagreement with such matching system is almost impossible. It is not only a problem of domain coverage and knowing all the possible beliefs of the user, but also of matching all the natural language variations in which the beliefs can be expressed as the pattern matching system uses a shallow understanding of the content, and for a larger coverage, the matching system would require deeper parsing of the inputs.

Instead, a *catch-all* approach has been chosen to argument with as much reactions as possible. For example, it is worth considering how broad are the dialogue contexts in which the counter-reactions *Reaction 2*, *Reaction 3* and *Reaction 5* can be used. The *utterance matching* pattern of each of these reactions can match almost anything the user says following the initial system's argument. Even if there is no limitation on how many reactions the database can contain, predicting and encoding all the possible user's counter-arguments is a difficult task for the author of the database; therefore using wide matches can be useful.

In the model of the desert survival scenario, over 180 *Reactions* are encoded to support 86 beliefs in the argumentation model which discuss the ranking of five items. This domain could have been described in greater detail, both in the argumentation model (containing the main support arguments that can be used) and in the generation model (containing generations for supports and for reactive attacks); however, the results obtained by the evaluation of the framework show that even with limited knowledge available for this domain, the EDEN Framework is able to persuade users to change their ranking (see section 4.2). The drawbacks of this limited knowledge on the interaction are also detailed in section 4.2.

A wider knowledge of the domain will improve the persuasiveness of the dialogue, but such knowledge requires a considerable amount of authoring work and a deep knowledge of the domain. The argumentation hierarchy and main *support* reactions for the Desert Survival Scenario are built using the information provided by Johnson (2003). This bootstrapped dialogue domain is used as a base for a Wizard of Oz experiment (Dahlback, Jonsson, & Ahrenberg 1993): ten users are asked to interact with the dialogue system, however they do not know that an expert user (the Wizard) is controlling the reactions of the system, either choosing an existing reaction from the generation model, or entering a new one. The ten dialogue instances were then analysed and used to input extended knowledge for the reactive component.

3.7.3 Generation Model

As the aim of the current research was to develop a novel dialogue management system for persuasion, no resources were available to develop a generation component. The task of realising the surface forms for the argumentative content was thus primarily based on a canned text system as illustrated by the previous examples.

The dialogue management framework is designed in modules that allow the extension and replacement of each component to evaluate different algorithms for each layer. In particular, the leaves of the matching tree can refer to any type of realisation method.

The reactive component, through the matching tree, activates content compatible with the communicative goals given by the planner. This content is selected and structured in dialogue segments according to the user's reactions. The generation layer is then responsible for producting natural language utterances from the instructions created by the reactive layer. The EDEN Framework is able to either use template based generation – where the final content selection and structuring are performed by the human author when creating the domain – or it can delegate the utterance content selection and structure planning to a full Natural Language Generation (NLG) system. The EDEN Framework then provides a semantic description of the utterance's communicative goal – beliefs, argument type – to be generated and the NLG component is responsible for the final utterance content selection, syntactic structuring and lexical realisation.

The examples of generation given previously illustrated canned text developed for the Desert Survival Scenario. The system would directly use the utterances linked in the leaves of the tree and fill in defined frames with matched content from the user's previous utterance and the current plan step. In the restaurant domain however, an utterance generator tailored to this domain was publicly available and the experiment could evaluate the EDEN framework beyond canned texts.

Mairesse & Walker (2007) introduces a tool for content generation parameterised with personality traits called Personage. The utterance generator takes a content descriptor and a set of personality parameters and is able to generate different surface forms for the same pragmatic content by creating generation plans for the RealPro realiser (Lavoie & Rambow 1997).

Integrating such a generation framework in the reactive component does not require a redesign of the dialogue management framework; the template leaves of the reactive tree are extended to call the of the Personage framework's generation commands.

Personage provides three pragmatic content descriptors for the Restaurant domain:

`recommend(`R**`,`**$AttList$**`)`** takes a restaurant R and a list of attributes (food quality, service, . . .) $AttList$ and generates a positive recommendation for this restaurant.

`disapprove(`R**`,`**$AttList$**`)`** generates a negative advice about a restaurant R according to the specified attributes $AttList$.

`compare(`R_1**`,`**R_2**`,`** $AttList$**`)`** recommends one restaurant R_1 over another restaurant R_2 according to the specified attributes $AttList$.

In the setup used to experiment the impact of personality on persuasion, using the Personage generator (see section 5), the personality parameters are set at the beginning of the dialogue and are global to all utterances. Thus, there is no need to specify them in the reactive tree and the templates only describe in which context to use the generation instructions and how to extract the parameters from the current plan step.

Figure 3.18 is an example of a matching tree for a set of support dialogue steps in the restaurant domain. The template of each *Reaction* is a mix of canned text and instructions for the generator. For example, the generation of *Reaction 9* could be:

> "Hum, **I believe Caffe Cielo isn't great mate, you know, alright? Actually, the atmosphere is darn bad.**"

where the last part (in **bold**) of the utterance is generated by Personage according to the global personality parameters.

In the example, the plan step is a generic support step to rank one restaurant ($Rest_1$) over another restaurant ($Rest_2$). In fact, the use of a generation engine such as Personage allows specifying fewer reactions in the domain model as the generator is able to create a variety of surface forms for the utterances from the same instruction. Hence, the matching tree is used for its core function: it constrains particular support utterances to their specific plan steps and defines the possible counter-arguments and attacks to be used in that context. The matching tree allows the selection of tailored utterance content relevant to the dialogue segment as well as the structuring of this content according to the user's interaction.

Plan Operator `support([recommend(`$Rest_1$`,` $Rest_2$`,` $AttList_1$`,` $AttList_2$`)], reorder(`$Rest_1$`,` $Rest_2$`))`

Reaction 8 :

Utterance Matching *

Template Ok, isn't $Rest_1$ better than $Rest_2$? `recommend(`$Rest_1$`,`$AttList_1$`)`.

Reaction 9 :

Utterance Matching *

Context Matching ok, isn't * better than *

Template hum, `disapprove(`$Rest_2$`,`$AttList_2$`)`.

Reaction 10 :

Utterance Matching which * you recommend *

Template `compare(`$Rest_1$`,`$Rest_2$`,` $AttList_2$`)`.

Figure 3.18: Matching Tree Sample with Generator Instructions. The variables $Rest_1$, $Rest_2$, $AttList_1$ and $AttList_2$ correspond to frames that match the current plan step and are then used to fill in the parameters of the generation command.

3.8 Layered Dialogue Management

The *reactive component* and the *planning component* defined in the previous sections are extensions of the state-of-the-art in dialogue management. The *planning layer* is a specialised dialogue planner tailored to persuasive dialogues and *the reactive layer* is a mechanism that will activate particular dialogue strategies according to the plan operators and to the user's input.

The novel approach of the EDEN framework – in addition to the specific formalisation introduced in the individual layers – is the interaction, during the dialogue, of these two layers.

Gilbert et al. (2003) describes a full argumentation system where the com-

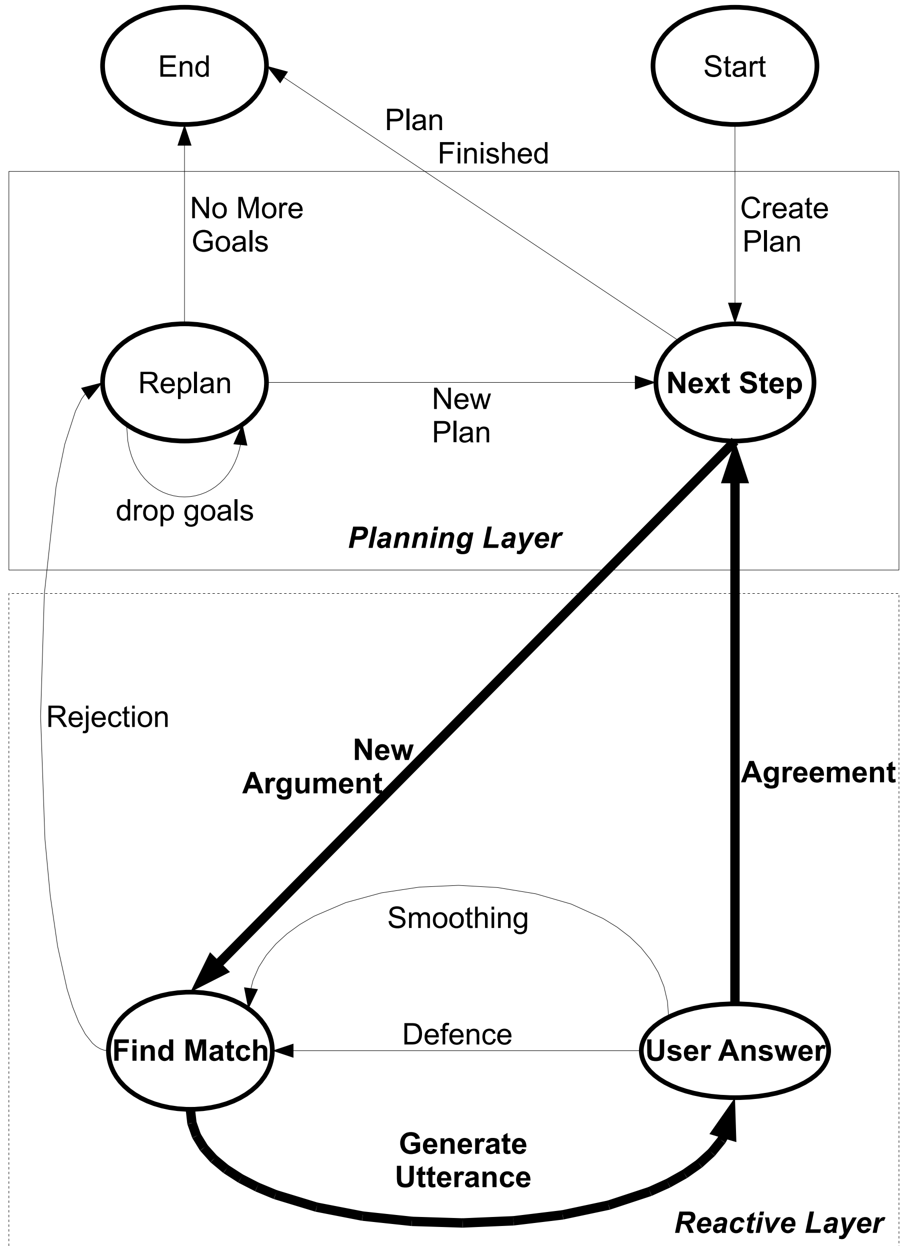


Figure 3.19: Dialogue Management States Machine. The dialogue starts by creating a plan (`Create Plan`) to achieve the persuasive goals specified for the conversation, it then presents the first argument (`New Argument+Generate Utterance`) to the user that can either agree (`Agreement`) with it or reject the new belief (`Defence`). In the case of disagreement, the system can either choose to defend the current argument and generate more utterances within this dialogue step, or it can replan to find a better dialogue path (`Rejection`). If the user agrees to the argument, the system moves to the next dialogue segment and continues the dialogue loop (in *bold*). If the system could replan (`New Plan`), it can continue in the dialogue loop, if it cannot, it needs to `Drop Goals` and replan. If the system drops all the existing goals, there is no more reason to continue the dialogue and it ends (`No More Goals`).

puter has to understand the content of each user's argument in detail to perform the argumentation. In the perfect *Persuasion Machine*, the system will understand the scheme an argument is using, the facts present in the argument and the veracity of each fact. Segmenting the user's argument in facts and testing their veracity to decide which fact to attack is a problem that has not been solved yet. This would require a deep understanding of the semantic of the user's utterance as well as a large database of facts. This database would allow the system to decide which fact can be attacked and how; however, as discussed in section 3.4.1, there is no guarantee that the fact used by the user are in the database. It is also difficult to reason about these facts without a full understanding of the user's values. In addition, a preliminary research for the thesis showed that deciding of the argumentation scheme used in a text is very difficult without deep processing of such text (see Appendix B).

In the EDEN framework, a pragmatic approach is taken to simplify the understanding of the user's utterance while keeping a working model for argumentation. Instead of understanding the full *content* of the user's arguments, the system tries to understand the *intent* of the utterances. In the approach described in this thesis, the argumentation process is divided in distinct dialogue phases. This simplification of the model allows to only perform a shallow processing of the utterances as described earlier and in the section 3.9. The dialogue model is divided in the following phases:

Argument Agreement is the phase of the dialogue where the system has presented an argument and the user accepts it without objections.

Partial Argument Agreement defines the dialogue segments where the users do not agree directly with the system's argument, but given more support they accept the new belief.

Argument Rejection is a dialogue segment where the system's argument failed; the user rejects the main argument even after new support is given.

Smoothing Dialogue takes place non deterministically in the dialogue when the system finds social cues or empathising utterances it can use.

The interaction of the two layers through these phases is summarised in the Figure 3.19 and is detailed in the following sections.

3.8.1 Agreement

SYSTEM: don't you agree that we should stay close to the plane wreckage? so the rescue can find us. [*from plan step:* `support([goal(be_found), rescueknowwhereyouare], goal(stayput))`]
USER: I guess you are right. [*agreement*]
...

Figure 3.20: *Agreement* Dialogue Sample in the Desert Survival Scenario.

The previous sections explain how the EDEN Framework selects a dialogue segment and matches a first generation for this plan step. Once provided with a belief to present to the user, the reactive component chooses an initial initiative realisation from the matching tree and starts to argue about this belief.

Figure 3.20 is a sample dialogue segment presenting the belief that "the user should stay close to the wreckage" (`goal(stayput)`) supported by the beliefs that "the rescue team knows where you are" (`rescueknowwhereyouare`) and that the goal is to "be found" (`goal(be_found)`). The system chooses to format the argument with an interrogative utterance and the user replies by an agreement.

From the user's utterance, the system has no possibilities to know if the belief `goal(stayput)` is internalised but, as the user has agreed, the system assumes that the belief has been accepted.

This is the simplest dialogue phase that can appear in the persuasive dialogue, the plan step is now achieved and the system can move to the next dialogue step.

Before starting the next dialogue segment, the EDEN Framework extends the user's belief model with the new belief `goal(stayput)`. For each agreed dialogue step, the system will *learn* more about the user's beliefs and build up the user model as the dialogue goes on.

3.8.2 Partial Agreement

SYSTEM: I think the flashlight could be useful as it could help us at night, let's rank it higher. [*from plan step:* `support([can(helpatnight, flashlight)], rate_higher(item(flashlight)))`]

USER: How is that? We are not going to move during the night. [*disagreement*]

SYSTEM: well, if we want to collect water, it will be best to do things at night and not under the burning sun. [*defence in the same plan step*]

USER: I see. It could be useful then. [*agreement*]

...

Figure 3.21: A *Partial Agreement* Dialogue Sample.

The user might not always accept the new belief without arguing. This is where the contextual reactions encoded in the matching tree are used. The system tries to defend the selected belief with the available defences and attacks.

Figure 3.21 is a sample of such phase. The user is disagreeing with the first argument that the "flashlight will be useful at night"; the user holds the belief that "there is no need to move at night" and the system chooses to defend its argument with an *attack* to this belief.

To maintain consistency and give an impression of understanding to the user, the dialogue framework must be able to defend its arguments and not fail every dialogue segment at the first rebuttal from the user. This is where the addition of the reactive component improves the persuasiveness; the EDEN Framework searches in the matching tree for an applicable defence in the context defined by the plan step and the user's disagreeing utterance.

If the defence argument is strong enough, the user will finally agree to the system's main argument. In addition, as in this example, presenting more arguments to the user is an opportunity for the system to learn more about the user's beliefs.

In the example of Figure 3.21, the system presents two new beliefs in the second utterance to support its argumentation: `goal(collect_water)` and `can(burn_skin, sun)`. By explicitly agreeing to the first argument, the user acknowledges to believe in `can(helpatnight, flashlight)` and `rate_higher(item(flashlight))`. By doing so, the user also implicitly hints an agreement with the other beliefs introduced in the second utterance of the segment.

The *Template* section of each defence *Reaction* in the domain knowledge is linked to a set of new beliefs they introduce to the user. When the system detects an agreement ending a dialogue segment, it inserts these new beliefs in the user belief model in addition to the main plan step's belief. In the current example, this means that the system has *learned* – in two dialogue utterances – that the user believes in:

- `can(helpatnight, flashlight)`
- `rate_higher(item(flashlight))`
- `can(burn_skin, sun)`
- `goal(collect_water)`

3.8.3 Rejection

The opposite of a direct *Agreement* phase is a *Rejection* phase where the user disagrees with the system's argument and does not accept the new belief.

In this case, and if possible, the system will react and defend its argument as explained in the *partial phase* section. However, the available reactions for a particular dialogue segment are limited by the size of the domain knowledge and the system might not be able to support the belief anymore. In addition, insisting on a belief irritates the user and is perceived as coercion.

The system must then abandon the belief that was chosen by the planner. Because the rest of the dialogue plan relies on the user's holding this belief, the plan is invalidated and cannot be used anymore. The planning component needs to find a new plan achieving the persuasive goals through a new path in the argumentation hierarchy.

As with the other phases, the output of a rejection is the extension of the user model. The disagreement teaches the system that the user did not internalise the defended belief. In addition, the EDEN Framework has also *learned* more about the user's beliefs during the previous successful phases of the dialogue; it can then find a different plan relying on the extended user model.

Finding a new plan has the drawback that it could jeopardise the consistency of the dialogue by jumping to a new line of argumentation. It is therefore preferable that the generation model includes enough supports to defend each argument to achieve as much *partial agreement* as possible.

Another drawback of replanning is that the argumentation hierarchy might not be large enough to contain a second path from the current user's beliefs to the persuasive goals. In that case, the system cannot continue the persuasion on that specific goal and must ease the persuasive goals by dropping the failed goal.

A dialogue contains more than one persuasive goal and the EDEN Framework maintains, for each plan step, a list of the goals the step is needed for. If the dialogue segment fails, the framework will try to find the minimum number of these goals that have to be dropped to find a feasible plan with the learned user's beliefs.

Dropping goals impairs the persuasiveness of the dialogue but is the only solution to avoid the computer being perceived as coercive. Having a large argumentation hierarchy will therefore improve the persuasiveness by avoiding the need to drop goals too often.

3.8.4 Dialogue Smoothing

In addition to presenting well ordered arguments and following rhetorical strategies, another part of persuasion is making the user feel comfortable in the dialogue. The persuader should build trust and create a *social bond* with the interlocutor (See section 2.2). Currently, bonds similar to human-human interaction are difficult to create with a computer but dialogue management systems should still create dialogues as natural and smooth as possible.

The three phases presented earlier are tools to create an argumentation structure in the dialogue, but tailored reactions to each argumentation step are not sufficient to make the dialogue appear as natural. The system needs to introduce social cues and backchannel utterances in the interaction so it does not appear too computerised.

Dialogue systems using reactive techniques activated by the dialogue context, such as Cassell & Bickmore (2002) *activation network* or simplistic chatbot systems, such as the AliceBot, are currently the best at showing these dialogue backchannels and social bonding cues as these are hard to encode as a functional parts of a plan. They need to be unpredictable and match the user's reaction and personality to seem natural. As discussed earlier in the thesis, Cassell & Bickmore (2002) mixes social cue strategies and domain relevant planning that provides reactiveness but impairs the authoring of dialogue domains and thus the portability of the system. A different approach is preferred for the EDEN Framework to ease the authoring of domains and allow authors to concentrate on the argumentation strategies design.

The EDEN framework interfaces with the AliceBot reactions database to find dialogue smoothing interactions when appropriate. The AliceBot project provides the sets[3] of reactions for the chatbot that won the Loebner prize but has

[3] see `http://www.alicebot.org/downloads/sets.html` for available AIML

no planning mechanism, making it solely reactive. The EDEN Framework uses 8,724 of these reactions to manage chit-chat and generate social cues, using the knowledge encoded in the AliceBot to react to the user in addition to the actual plan given by the higher layer.

The AliceBot database is merged with the domain matching tree without any specific dialogue segment constraint. When the reactive component matches a reaction in the AliceBot part of the search tree, the reaction is presented to the user and the system waits for a reaction from the user. If no reaction comes after a threshold delay, then the system continues with the normal argument management steps – i.e. either one of the three previously mentioned phases, depending on the dialogue state.

Using the reactive activation of dialogue strategies to manage the social cues allows us to easily separate the argumentation planning and reactions from the discourse level content management. However, it remains a reactive technique that cannot plan for complex relationship building strategies; still, the smoothing phase of the dialogue has access to reactions that have shown to be effective in simulating natural human conversation by helping the AliceBot win the Loebner prize, which tries to repeat the *Turing test*.

SYSTEM: We don't have much water, we need to be rescued as soon as possible. [*from plan step:* `user_world( goal(be_found)))`]
USER: right [*agreement*]
SYSTEM: I am glad we agree.[*backchannel*]
SYSTEM: There is a good chance that the rescue team already knows our whereabouts. We should be optimistic and plan accordingly, don't you think?[*from plan step:*`use_world(rescue_knows_where_you_are))`]

Figure 3.22: *Smoothing* Dialogue Sample in the Desert Survival Scenario.

sets.

Figure 3.22 is an example of a simple dialogue smoothing after an agreement from the user. Instead of jumping to the generation of the next dialogue segment, the system finds a backchannel reaction “I am glad we agree” to show empathy with the user.

In addition to the domain knowledge, the system can access any of the 8,724 reactions available in the AliceBot database to interact with the user outside of the main argumentation segments.

3.9 Agreement/Disagreement Classification

The dialogue management system relies on the four phases in section 3.8 – Agreement, Partial Agreement, Rejection and Smoothing – to simulate a natural argumentation dialogue. This simplification of the model in four phases of the dialogue allows the implementation of a working dialogue management system that can argue with the users without needing the inclusion of, yet to be researched, deep argument understanding techniques.

To support this split of the interaction in these four phases, the system must be able to detect if the user is agreeing with the system or rejecting the argument. The reactive component needs to detect if the dialogue is entering a partial agreement phase – with the user disagreeing to the new argument – or if the current argumentation segment is finished – with the user agreeing to the new belief. The dialogue management system is developed from the perspective of a system easily ported to different domains and, therefore, a domain independent and robust agreement/disagreement detection is used.

3.9.1 Previous Work

Hillard, Ostendorf, & Shriberg (2003) proposed a first step towards a statistical method for agreement/disagreement classification by developing a supervised learning classification based on an annotated selection of meetings from the ICSI Meeting corpus (Janin, Baron, Edwards, Ellis, Gelbart, Morgan, Peskin, Pfau, Shriberg, Shriberg, Stolcke, & Wooters 2003). The ICSI corpus is a collection of transcripts of meetings, which contains prosodic annotation in addition to the content of the dialogues. Hillard et al. selected 1,800 segments of speech, called *spurts* that correspond to segments of the dialogue with no pauses in the speech. These spurts were manually labelled with one of four possible labels:

BackChannel are short spurts that, having the form of agreement – e.g. yeah, ok, yep – could also be “encouragement for the speaker”,

Positive is used for spurts that are clear agreement,

Negative are used for disagreement spurts,

Others are long spurts that cannot be classified as either agreement or disagreement.

The classifier proposed by Hillard et al. is based on a decision tree and uses *lexical* features such as the number of words, the number of positive/negative words and the class of the first word of the sentence combined with *prosodic features*. The class of the words was inferred from their frequency in each class of labelled spurts.

Galley, Mckeown, Hirschberg, & Shriberg (2004) extends on Hillard et al. approach by introducing novel features and using a *Bayesian network* for spurt classification. The previous work uses only *local* features of a spurt to decide its class, in Galley et al., the feature set is extended with *contextual* features. Galley et al. use *adjacency pairs* to encode the interaction between speakers and relationships between consecutive spurts. The *adjacency pairs* provide extended information on where the spurt is used and give more clues to the classifier on the class of the spurt.

To detect the *adjacency pairs* relationship between spurts, Galley et al. use a statistical ranking algorithm based on maximum entropy. This algorithm can learn, with 90.2% accuracy, how to find the previous element of a pair given the last spurt of the pair. The Bayesian classifier is then trained/tested on an extended annotated corpus of 8,135 spurts using the *contextual* features combined with a set of *durational* features (length of the spurt, length of silences in the spurt, . . .) and with *lexical* features (number of words, positive polarity adjectives, . . .).

Comparative results of each approach are detailed in Table 3.3 and 3.2.

3.9.2 A Text Based Approach

The EDEN Framework is developed primarily as a textual dialogue framework (see Figure 4.2), using an interface similar to instant messaging. In the current implementation, it cannot rely on *prosodic* or *durational* features like the state-of-the-art classification presented previously. However, there is no relevant annotated corpus for text-based dialogue and the works of Hillard et al. (2003) and Galley et al. (2004) on the ICSI corpus are, to our knowledge, the only available research on agreement/disagreement classification.

The EDEN Framework uses a classifier based on the Agreement/Disagreement class annotation of the ICSI corpus provided by Galley et al. and Hillard et al. for training and evaluation. An algorithm based only on textual features achieving practical results comparable to the state-of-the-art algorithms was developed. In fact, Hillard et al. found that the *prosody* features did not improve the accuracy of this classifier.

Because of a lack of resources, the EDEN Framework classifier could not rely on automatically extracting Adjacency Pairs. A multi-class Support Vector Machine classifier (SVM; Vapnik 2000) is trained on the following shallow *local features* of the spurts:

- the length of the spurt (in characters),
- the first word of the spurt,
- the bi-grams of the spurts (i.e. all consecutive pairs of stemmed words in the spurt),
- part of speech tags,

- number and type of punctuations in the spurt.

In addition, a lexicon of *agreement* words and of *other* words is created by counting the number of each word in each class. If the frequency of a word in a particular class is significantly higher than in the other class, the word is assigned this class in the lexicon. In addition to the previously mentioned features, the classifier also uses the number of *agreement* words and *other* words in the spurt as learning features.

The multi-class classifier is composed of 3 binary SVM classifiers in cascade in the order BackChannel → Agreement → Disagreement vs. Others (BADO) (see Figure 3.24). The rationale for the cascade classifier is that the difference between the BackChannel class and the other classes is easy to make based on the *length of the spurt* feature (See Figure 3.23). After this first classification, there is less noise added by the BackChannel spurts in the learning. In a similar way, the *Agreement vs. Disagreement+Others* classification is mainly dependent of the *first word* feature as agreement/disagreement spurts usually start with a limited set of words (e.g. yes, agree, no, well, . . .).

3.9.3 Results

Two experimental setups were used to evaluate the accuracy of the classifier:

Setup 1 reproduces Hillard et al. (2003) training/testing split across meetings. One meeting transcript is held out for testing and the classifier is trained on the rest of the meetings as a three-ways classifier – *Agreement* and *BackChannel* classes being merged. However, this split is biased as one particular meeting might contain specific topics and features that are not found in the rest of the meetings.

Setup 2 performs a five-fold cross-validation with the four-way classifier (comparable to the experimental setup used in Galley et al. 2004). The 8,135 spurts are split randomly in five samples; each sample is consecutively

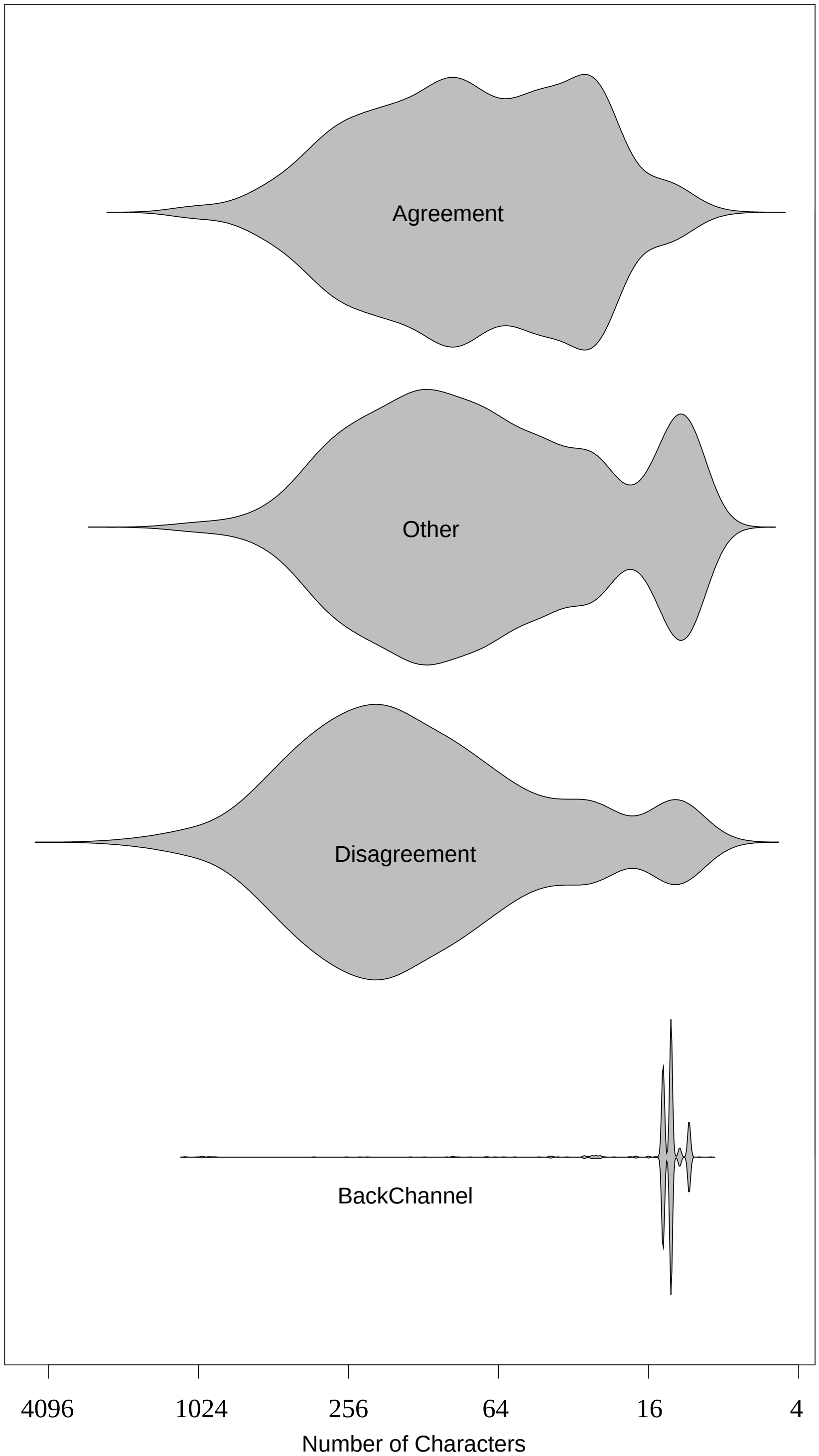


Figure 3.23: Scaled density of the repartition of spurts according to their length feature for each agreement class (log_2 scale). The distribution of spurts in the BackChannel class is skewed towards short spurts whereas the other classes spread along all the possible length.

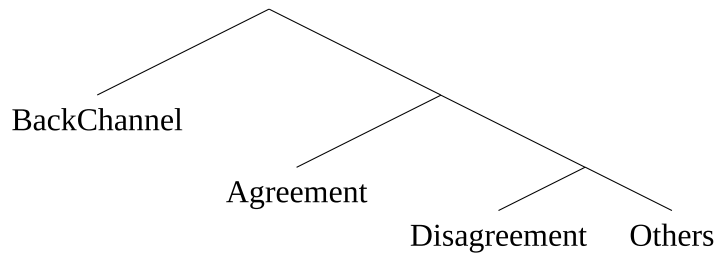


Figure 3.24: Binary SVM Classifiers in Cascade. Three binary SVM classifiers are used consecutively to label the spurts. The top classifier decides of the *BackChannel* class, if the spurt is not of this class, the second classifier is applied to decide of the *Agreement* class. If the spurt is not an *Agreement*, the last classifier is applied to choose between the *Disagreement* and *Others* classes.

used individually as testing sample against a classifier trained on the rest of the samples.

					Galley et al. Features	
	BODA	**BADO**	One vs. All	Hillard et al.	Local	Global
Error Rate	13.53%	**13.48%**	17.78%	18%	14.38%	13.08%
Std. Dev.	1.30%	**0.96%**	0.87%	NA	NA	NA

Table 3.2: *Setup 1* classifiers comparison. The BackChannel → Others → Disagreement vs. Agreement cascade Support Vector Machine (SVM BODA) and BackChannel → Agreement → Disagreement vs. Others cascade SVM (BADO) achieve better results than the One vs. All SVM classifier. The cascade classifiers' accuracies are comparable to the state-of-the-art techniques, in particular the accuracy is better than the classifier using only spurts features (*Local Features*) by Galley et al. and close to the classifier using adjacency pairs (*Global Features*).

Setup 1 was tested on two different classifier cascades: BackChannel → Others → Agreement vs. Disagreement (BODA) and BackChannel → Agreement →

		Galley et al.	
	BADO	Global Features	Local Features
Error Rate	**13.47%**	15.93%	16.89%
Error Std. Dev.	**0.57%**	NA	NA

Table 3.3: Error rate of the classifiers for *Setup 2*. The accuracy of the *BackChannel* → *Agreement* → *Disagreement* vs. *Others* (BADO) SVM classifier is better than the state-of-the-art classifiers even if it only uses spurt's features but could benefit of the addition of adjacency pairs and other global features.

Disagreement vs. Others (BADO) as well as a One-Versus-All classifier. The error rates of all the classifiers are comparable (see Table 3.2). The accuracy of the second cascade classifier turns out to be the best and the one-vs.-all classifier's accuracy the worse.

Setup 2 results are reported with comparison to the state-of-the-art techniques accuracies in Table 3.3. The SVM classifier outperforms the performances reported in Hillard et al. (2003) as well as the one from Galley et al. (2004) classifiers. In addition, the BADO cascade appears to achieve more robust accuracies across the two corpus splits. An accuracy of 86.36% is enough to achieve the task of initial filtering of the dialogue utterances in the EDEN Framework as the matching tree is then used to decide of the reactions to take.

The *Agreement* and *Disagreement* classes are decreasing the accuracy of the classifier (see Table 3.4) with an accuracy of 39% for the *Disagreement* class while the *BackChannel* class has an accuracy of 98%. This is due to the small number of examples available in the corpus for the *Disagreement* and *Agreement* classes. The *BackChannel* class, relying on the strong *length of spurt* feature can be predicted easily while the classification of the *Others* class, with 1,103 examples in the corpus, can be trained with good accuracy.

During the dialogue, the classifier is applied on each of the user's utterances, trying to determine if the user is agreeing or disagreeing with the system. Ac-

	BackChannel	Others	Agreement	Disagreement
Precision	**0.99**	0.90	0.67	<u>0.38</u>
Recall	**0.98**	0.91	0.62	<u>0.40</u>
F_1	**0.98**	0.91	0.64	<u>0.39</u>
Error Rate	**2.2%**	9.1%	37.8%	<u>59.8%</u>
Error Std. Dev.	1.5%	1.4%	9.9%	4.5%
Repartition in Corpus	22.6%	**61.7%**	9.4%	<u>6.3%</u>

Table 3.4: Precision and Recall for Individual Classes in the BackChannel → Agreement → Disagreement vs. Others Cascade Classifier. The Table 3.5 shows that the *Agreement* and *Disagreement* classes have poor classification results as they are sometimes classified as *Others*. This is due to the lack of examples in the corpus, with only 9.4% of spurts being instances of the *Agreement* class and 6.3% being instances of the *Disagreement* class. The *Others* class with 67% of the examples in the corpus achieves a 91% accuracy while the *BackChannel* class has a good classification accuracy thanks to a strong discernable feature. The best results are in **bold** and the worst <u>underlined</u>.

	Predicted Classes			
Real Classes	BackChannel	Others	Agreement	Disagreement
BackChannel	**97.8%**	0.7%	<u>1.2%</u>	0.3%
Others	0.5%	**90.8%**	3.4%	<u>5.3%</u>
Agreement	4.7%	<u>26.1%</u>	**62.1%**	7.1%
Disagreement	0%	<u>51.8%</u>	8.0%	**40.2%**

Table 3.5: Confusion Matrix for the *BackChannel* → *Agreement* → *Disagreement* vs. *Others* Cascade Classifier. This table presents the repartition of the annotated spurts by predicted classes. Each line represents a known class of spurt – from the corpus annotation – and the percentage of these spurts that were classified in another class. The spurts that were correctly predicted by the SVM classifier are in **bold**; the worst confusion of each line is <u>underlined</u>.

cording to the definition of the corpus classes in Hillard et al. (2003), the *Others* class is considered as a *Disagreement* by the system since it corresponds to long spurts that may contain argumentation and do not require a reaction. The *BackChannel* class is taken as an *Agreement* and also characterises an *Agreement phase* in the dialogue as Hillard et al. defined this class for spurts used as "encouragement for the speaker". Even if the accuracy of the *Agreement* and *Disagreement* classes are individually low, 51.8% of the misclassified *Disagreements* are labelled as *Others* (see Table 3.5) which does not impair the use of this classifier to detect the dialogue phase. In addition there is little confusion between the *Agreement* and *Disagreement* classes themselves.

As one of the interlocutors in the dialogue is controlled by the system, adding *Adjacency Pairs* features and other contextual features – such as the type of answer expected by an utterance – might be easy to implement and could improve the accuracy of the classifier for this particular application in future work.

3.10 User's Beliefs Modelling

The EDEN Framework division in layers relies on the domain model for planning and reacting in the dialogue; however, there is a need for the system to keep a representation of the users. Indeed, when planning for dialogues, the system needs to know what the users already believe, their personality, and their preferences.

3.10.1 Beliefs Monitors

When the plan fails (see section 3.8.3), the dialogue needs to stay consistent during the replanning, which requires that the system knows which beliefs were accepted by the user and which still have to be defended.

Section 3.8 explains how the EDEN Framework makes assumptions about the beliefs internalised by the user based on the reactions in the dialogue. This is not always possible as the user can do extra reasoning based on the new beliefs and infer new beliefs not expected by the system.

For example, in the Desert Survival Scenario, users tend to rank the *map* and the *compass* higher than the *flashlight*, thinking that they can “walk back to safety”. When the system tries to convince the user that the *flashlight* should be ranked higher than both the *map* and the *compass*; the system starts, for example, by arguing that the *compass* is not useful because “it is dangerous to walk in the desert” and then continue by saying that, for the same reason, the *map* should be ranked lower. The system cannot assume that, because the users reranked the *compass*, they extend the reasoning to the *map* without arguing. However, the user probably has done the rest of the reasoning, in which case the argumentation step discussing the *map* ranking becomes redundant.

This is a simplistic example that could be planned for to merge the *compass* and *map* argument if the domain knowledge is designed properly. However, this kind of problem can arise in more complex situations and the system cannot plan on how changing one belief will influence the whole domain in the user’s mental representation. These changes cannot be detected directly through dialogue either and the EDEN Framework therefore integrates another protocol to *learn* about these changes in the user’s beliefs. The *Beliefs Monitors* are external monitors that can notify the EDEN Framework that the user’s beliefs have changed.

For example, in the simple task given for the Desert Survival Scenario, the users have access to a module where they can reorder the items as they change their beliefs during the whole dialogue (see Figure 4.2). An external monitor is keeping track of this list and can notify the system of what the user actually internalised. In the previous example, if the user changes the order from compass-map-flashlight to flashlight-compass-map on the ranking module, swapping three items at once, then the system is notified that two persuasive goals were actually achieved.

When a goal is achieved as a side effect to a dialogue segment and detected by a *beliefs monitor*; the EDEN Framework can simplify the actual plan to disable the steps linked to this goal that are rendered redundant and would be perceived as repetitions by the user. Replanning is not needed for this simplification, for each goal the system keeps a list of all the steps supporting the goal, and, when a goal is achieved, and if these steps do not support other active goals, the linked steps will be deactivated in the plan.

A *beliefs monitor* can detect more than an achieved goal. In fact, the users can change their beliefs in a direction opposite to the system's goals, creating new persuasive goals to be achieved. In this case, the *beliefs monitor* helps to detect the need for replanning even if the user is not explicitly rejecting arguments in the dialogue.

3.10.2 Learning the User's Beliefs

The EDEN Framework is able to reason on the beliefs to introduce based on the user's preferences and a domain knowledge base to support this reasoning, the system needs to be able to store a model of the "user's mind". The *User Beliefs Model* is a representation of what the system knows about the user's knowledge at a given step of the dialogue.

The EDEN Framework models the user's knowledge as a set of beliefs assumed to be held by the user. The system starts the dialogue without any prior knowledge of the user and has to plan to assert all the world knowledge and factual beliefs it needs to support the introduction of the goal beliefs.

At each step of the dialogue, the EDEN Framework *learns* more about the beliefs held by the user by analysing the output of a dialogue segment or through external beliefs monitors. The set of beliefs in the user model grows until the end of the dialogue when it holds the goal beliefs (see the Figure 3.25).

The system can also start a dialogue with existing knowledge of the user. In fact, if the system is used in a domain that requires more than one dialogue, it will have created a belief model of the user in the previous dialogues and will be able to tailor the consecutive dialogues to the user knowledge.

The restaurant domain is another example as the system needs to know the

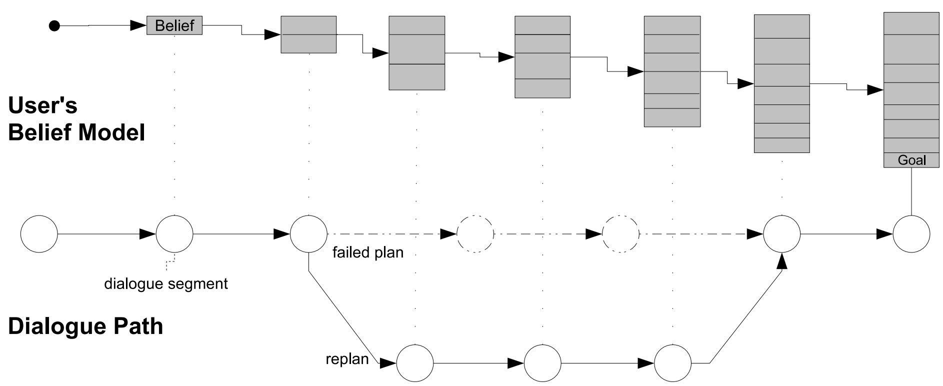


Figure 3.25: Learning User's Belief. As the dialogue goes on, even when a plan fails and the system has to choose another dialogue path, the system *learns* more about the user's beliefs.

user's preferences to plan the dialogue. These preferences are asserted before the dialogue and bootstrap the user's model with top level beliefs that can then be used to find a more persuasive plan for this particular user (see sections 3.5 and 5.2).

3.11 Framework Benefits

The EDEN Framework introduces a novel dialogue management technique to produce persuasive dialogue. The framework design focuses on producing natural dialogues where the user feels comfortable and the system can easily implement argumentation strategies.

The argumentation is formalised in a hierarchy that is easy to understand and to author by humans and can be used for automated reasoning without computational drawbacks. In addition, the novel dialogue management technique, by leaving the management of reactions to user's utterances outside of the plan, allows for a more flexible dialogue where the argument/counter-argument cycle can be controlled.

The design of the framework also includes tools to detect and learn the change in the user's beliefs to achieve tailored arguments. By adapting the selected arguments to the user's needs, the system can produce shorter dialogues while keeping focused on the persuasive goals.

Chapter 4

Persuasiveness of the Framework

4.1 Evaluating Persuasion

4.1.1 Output of Persuasion

Current practical research on computer persuasion concentrates on Embodied Conversational Agents (ECA). For example, Bickmore & Picard (2005) uses an ECA to persuade university students to walk more and improve their physical condition. Bickmore & Picard create an evaluation that follows the students over a long period of time and uses a pedometer to assess the evolution of the subjects behaviour and hence the persuasiveness of the system. In this thesis, such long-term evaluation was impractical, in particular, it is difficult to guarantee the independence of the behaviour change from external elements not controlled in the experiment; therefore, an atomic evaluation protocol is used, where the dialogue is the only element influencing the change of behaviour.

In the design of the experiments discussed in this thesis, the sociology models of behaviour change (see section 2.2.2 for details) are exploited to solve the problem of observing the user's long-term behaviour change. According to some of these theories, one's intention towards a behaviour is function of the *Attitude* towards this behaviour and the *Subjective Norms* linked to this behaviour. The latter is conditioned by the social pressure existing about such behaviour and the motivation of the user to abide to such pressure. The *Attitude* is conditioned by the user *beliefs* about this behaviour and beliefs' importance to the user. Evaluating a behaviour change requires a long-term observation of the behaviour; in contrast, an evaluation protocol that evaluates the change in the beliefs underlying the behaviour is preferable to avoid the need for long-term behaviour observation.

Evaluating the persuasion by testing the acquired *Attitude* of the user towards a behaviour makes the evaluation imperfect as it forgets the importance of the *subjective norms*. For the evaluation to be unbiased, the domains of evaluation were chosen to be as independent of *subjective norms* as possible. In addition, the user achieves the evaluation task through the anonymity of Internet, avoiding explicit social pressure.

4.1.2 Measuring Belief Change

Section 3.2 introduces the Desert Survival Scenario which is used in the first evaluation. The persuasiveness metric explained in this context is also used for the second evaluation in the restaurant domain (discussed in the section 5).

In the Desert Survival Scenario, participants first give their preferred initial ranking R_i of items (knife, compass, map, $\ldots$) and then engage in a dialogue with the system in which it attempts to change the participants' ranking to a different ranking R_s through persuasive dialogue; at the end of the dialogue, the participants can change their item choice to a final ranking R_f.

The persuasiveness of the dialogue is measured as the evolution of the distance between the user rankings (R_i, R_f) and the system's goal ranking (R_s). If the system is persuasive, it changes the user's beliefs about the items ranking towards a ranking similar to R_s. The change of beliefs is reflected by the evolution of the distance between rankings.

The *"Kendall τ permutation metric"* (Kendall & Gibbons 1990) is used to compute the pairwise disagreement between two rankings; measuring the number of swaps between adjacent items to get from one ranking to the other ranking. This measure has been chosen over other existing rank distance measures(Kendall & Gibbons 1990) as it provides a metric distance and, more importantly, is easy to interpret in the context of a persuasiveness measure. Kendall τ permutation metric is defined in Equation (4.1)[1] where P_{airs} is the set of all possible pairs of items of R_1 and R_2.

$$K_\tau(R_1, R_2) = \sum_{\{i,j\} \in P_{airs}} \bar{K}_{i,j}(R_1, R_2) \tag{4.1}$$

$$\bar{K}_{i,j}(R_1, R_2) = \begin{cases} 0 & \text{if the pair of items } i \text{ and } j \text{ are in the same} \\ & \text{order in the two rankings,} \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

Equation (4.3) defines the evolution of the Kendall τ permutation metric during the dialogue and provides a metric evaluation of the dialogue's persuasiveness.

$$\begin{aligned} P_{ersuasiveness} &= \Delta(K_\tau(R_i, R_s), K_\tau(R_f, R_s)) \\ &= K_\tau(R_i, R_s) - K_\tau(R_f, R_s) \end{aligned} \tag{4.3}$$

For example, if the user's initial ranking of the items is $R_i = map > flashlight > compass$ and the system goal ranking is $R_s = compass > flashlight > map$. The Kendall τ permutation metric is calculated with the table of pairs:

[1] from Fagin, Kumar, & Sivakumar (2003)

R_i	R_s	$\bar{K}$
map $>$ compass	compass $>$ map	1
map $>$ flashlight	flashlight $>$ map	1
flashlight $>$ compass	compass $>$ flashlight	1
	$K_\tau(R_i, R_s)$	3

If the final user ranking is R_f=flashlight $>$ compass $>$ map, the table of pairs is:

R_f	R_s	$\bar{K}$
compass $>$ map	compass $>$ map	0
flashlight $>$ map	flashlight $>$ map	0
flashlight $>$ **compass**	**compass** $>$ **flashlight**	**1**
	$K_\tau(R_f, R_s)$	1

At the beginning of the dialogue, the distance is maximal between the two rankings – three swaps are needed – whereas, at the end of the dialogue, only one swap is required. The persuasiveness metric is then: $P_{ersuasiveness} = 3 - 1 = 2$.

For an n items ranking, the range of the Kendall τ permutation metric is

$$[-\frac{n \times (n-1)}{2}, \frac{n \times (n-1)}{2}]$$

.

4.2 The Desert Survival Scenario

The EDEN dialogue framework is designed to improve persuasion in human-computer dialogue, both by easing the authoring of argumentation and by providing a novel dialogue management technique that is more responsive to the user than state-of-the-art planned approaches.

The first experiment is designed to validate the EDEN Framework design and to evaluate the persuasiveness achieved by the EDEN Framework and the one achieved by a purely planned approach.

Moon (1998) uses the Desert Survival Scenario to analyse the effect of distance on the computer's persuasiveness. This scenario is also used in the present experiment to evaluate the persuasion of the dialogue system; however a different metric is defined in section 4.1.2 for clarity of interpretation.

The Desert Survival Scenario is an interpersonal communication task often used to evaluate or train team communication and negotiation abilities. In the scenario, the participants are stranded in the desert after a plane crash and have to organise their survival. A set of twelve items can be salvaged from the plane-wreck, the participants must agree on a ranking of items from the most useful to the less useful. The original Desert Survival Scenario uses the following twelve items:

- one litre of alcohol
- a flashlight
- a knife
- a sectional airmap
- a raincoat
- a compass
- a compress kit

- salt tablets
- one litre of water
- a book
- a topcoat
- a mirror

Two issues arise when trying to use this list for the dialogue experiment in a preliminary trial study. First, the participants thought that it was too long to rank all the twelve items. The second issue can be seen as a problem of ethics, to collect relevant persuasiveness values, the system needs to have a goal ranking distant enough from the user ranking; however, this means that it might, for example, have to persuade the user that a *litre of alcohol* is more useful than a *litre of water*. This is never true in a desert survival scenario and it would not be acceptable to persuade the users of such belief. In addition, trying to persuade the user that water is not useful in the desert is already a lost cause in most cases and would bias the persuasive metric.

To simplify the ranking task and avoid an impossible persuasive task, the list is limited to five items:

- a flashlight
- a knife
- a sectional airmap
- a raincoat
- a compass

The participants are also told that they have already salvaged a litre of water. As the system uses five items to evaluate the persuasiveness of the system, the range of the persuasion metric is $[-10, 10]$ which will sometimes be reported normalised between $[-1, 1]$ in the following results analysis.

This experiment is designed to show that the EDEN Framework, by adding a reactive layer in the dialogue model, helps at developing more persuasive dialogues by easing the inclusion of reactive strategies as well as social cues without developping a complex planning model. To test this hypothesis, the participants face two consecutive dialogue sessions, following the same procedure but where the conversation is managed by different versions of the dialogue system:

- One dialogue session is managed by a *limited* version of the dialogue system, with no reactive component. This version is similar to a purely task-oriented system, relying solely on revisions to the plan to react to dialogue failures.

- The second version is the *full* EDEN dialogue management system as described in the previous chapters.

To ensure that the order in which the systems appear to the users does not bias their judgement, the order in which they are shown to each participant is randomly chosen.

4.2.1 Procedure

A within-subject experiment is conduced where each participant faces both types of systems consecutively. The order in which the *full* and *limited* systems are presented is randomly chosen. The procedure for both system is exactly the same for the user, they were not told that two different systems are used and do not know in which order the systems are presented. An informal questioning of the participants showed that they did not realise they were facing different types of systems.

Each participant is presented with the scenario through a web interface:

Step 1

You took a flight with friends to a party in Las Vegas. However, the airplane crashed in the desert. You and another person survived.

In this experiment, you will have a chat with another user taking part in the same experiment. After the chat session, you will fill in a small questionnaire about your experience.

Please be sure to complete each step in order without going back to a previous step.

---

the participant is then explained the task:

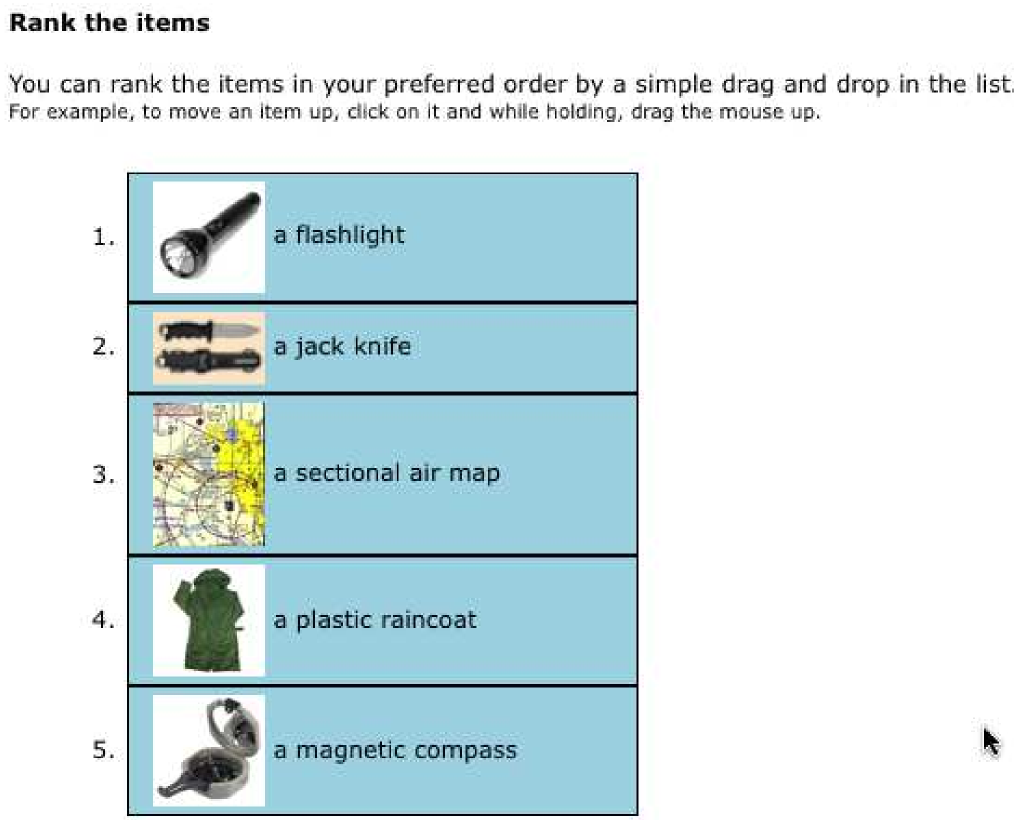
Step 2

Both you and the other person have been told that the purpose of the chat is to agree on what are the most important things to keep.

You are disorientated and you do not know what to do next. However, in the current scenario, you have the opportunity to salvage items from the airplane wreckage. You already saved a litre of water.

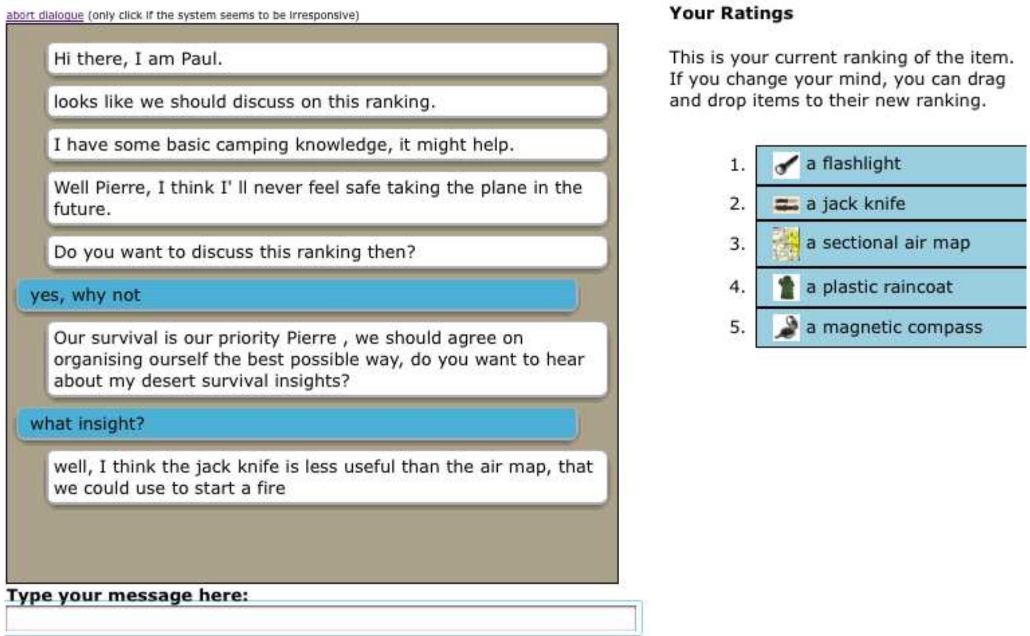
In the second step, the participant is asked for a name and an initial ranking. To enter the ranking, the user can drag and drop items in an ordered list. Each item is presented with its name and a photo as illustrated by Figure 4.1.

When the user is done setting an initial ranking, the system initiates a dialogue session and starts the argumentation to persuade the participant to change the ranking. At any time in the dialogue, the user is able to keep track of the ranking with the items list on the right of the dialogue session (see Figure 4.2).



(Original in colour)

Figure 4.1: Screenshot: Ranking Desert Items

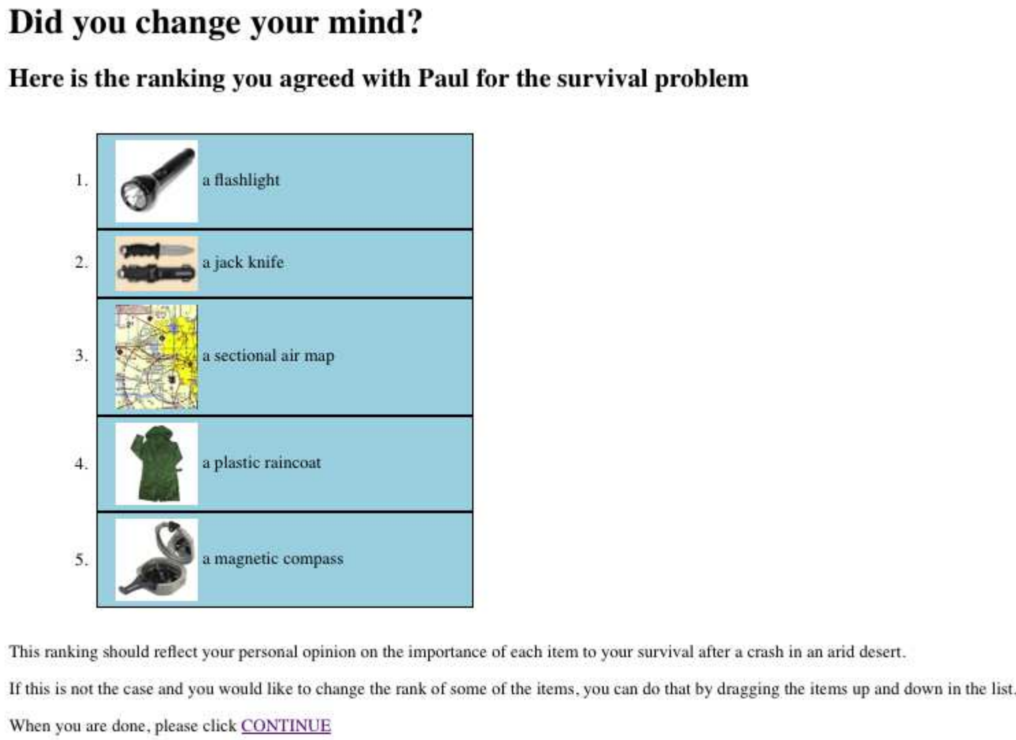


(Original in colour)

Figure 4.2: Screenshot: Start of the Dialogue Session

At the end of the dialogue, the participant is explicitly asked to rerank the items as in preliminary trial runs of the desert scenario design, participants reported not understanding that they could change the ranking during the dialogue, even if they had actually changed their mind.

Participants are then asked to fill in a questionnaire about their interaction with the system. The full questionnaire is reported in Appendix C. The general questions on the quality of the interaction were inspired by the PARADISE (Walker, Litman, Kamm, & Abella 1997) framework. Note that even if the same questions were used, the parameter optimisation described by the PARADISE framework is not used here. Each question (except for the details about the participants and a comment field) is formulated as a statement that the participant can evaluate on a five-points Likert scale of agreement/disagreement: *Strongly Disagree* (0), *Disagree* (1), *Neither agree nor disagree* (2), *Agree* (3),



(Original in colour)

Figure 4.3: Screenshot: Opportunity to Rerank

Strongly Agree (4).

4.2.2 Selection of System Goals

To avoid convincing the user of implausible rankings and giving false information, two fixed system goals are used and are given in Table 4.1.

The goals ranking is randomly selected when the dialogue starts and the belief goals for the planner are produced by finding the minimum number of swaps needed to get from the user's ranking to the system's ranking.

A fixed goal has the disadvantage that not all dialogues have the same persuasive strength. If the user starts with a ranking close – or equal – to the system's ranking, then the dialogue has to present less arguments. This leads to an inter-

Goal 1	**Goal 2**
flashlight	flashlight
raincoat	raincoat
compass	knife
airmap	airmap
knife	compass

Table 4.1: System Goal Rankings

esting behaviour in the users as some of them achieved a number of swaps away of the system's goal ranking when faced with the *limited* system, resulting in a negative persuasiveness of the system as the final ranking is further away from the system goal ranking.

4.3 Results

Sixteen participants took part in the online evaluation and faced both systems. The participants group is of mixed gender and age (from 20 to 59) and has a variety of backgrounds. This is a small group of participants due to the lack of time and resources that were allowed for this experiment, the results should be interpreted in regard to this number of participants. However, the experiment provides a number of significant results that support our initial hypothesis and confirm the interaction guidelines that are laid out in the litterature.

The EDEN Framework is designed to provide more defence to each argument and this led the *full* system dialogues to be significantly longer than the *limited* dialogue ones. The *full* system tries to defend more its arguments and includes more chit-chat, which leads the user to enter more utterances (see Figure 4.4) and the resulting dialogues are thus significantly longer ($\hat{d} = 1.15$) with a mean of

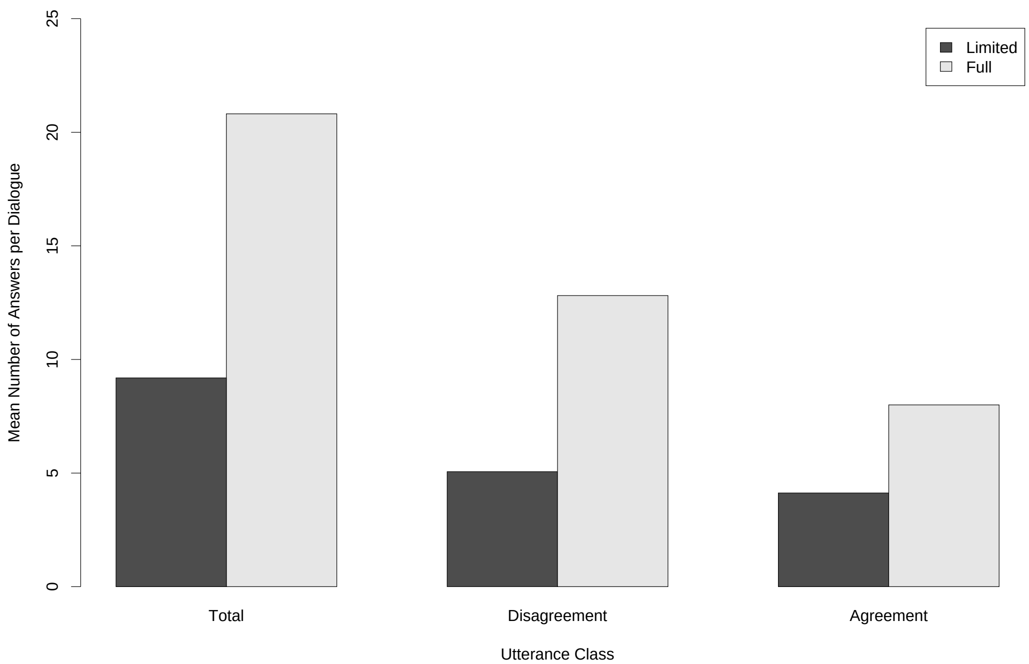


Figure 4.4: Mean Number of User's Utterances per Dialogue, split by agreement classes (see section 3.9).

46.9 utterances[2] ($SD = 24.0$) for the EDEN Framework versus 20.6 utterances[3] ($SD = 11.2$) for the *limited* manager (Wilcoxon $T = 2$, $p < 0.01$).

The repartition of *agreement* and *disagreement* utterances[4] does not change between the *limited* system and the *full* system dialogue. Indeed, there is no reason for this repartition to differ as the arguments defended are the same and only the amount of defence to each differs between systems, thus the number of user's objections to an argument should not differ. No significant differences are measured between the number of *agreement* utterances and *disagreement* utterances from the participants for both systems (Wilcoxon $T = 80$, $p = 0.55$; see Table 4.2).

[2] 43.8% of which are from the user ($SD = 4.5\%$)
[3] 43.2% of which are from the user ($SD = 6.3\%$)
[4] according to the classifier discussed in section 3.9.

	Limited System	Full System
Mean Repartition	-1.4%	-12%
Std. Dev.	62%	36%

Table 4.2: Repartition of *Agreement* vs. *Disagreement* Utterances by Dialogue System. The repartition is computed as the difference between the number of user's *Agreement* and the number of user's *Disagreement* on the total number of user's utterances per dialogue. A negative repartition means that there are more *Disagreements* than *Agreements* per dialogue.

4.3.1 Persuasiveness of the Systems

The goal of the EDEN Framework is to produce more persuasive dialogue and be able to change the user's beliefs and eventually their behaviour. The persuasiveness metric described in section 4.1.2 is used to evaluate the persuasion achieved by each system and is reported in Figure 4.5. The main hypothesis of this research is that the EDEN Framework design eases the persuasion and achieves better results than a standard planning design. The participants in the *limited* system achieve a mean distance of 0.44 swaps ($SD = 2.25$) *away* from the system's goal ranking – with a worst case of 4 swaps *away* – when in the *full* system, 1.44 swaps ($SD = 2.39$) *towards* the system's goal are measured – with a maximum of 8 swaps. The Desert Scenario Experiments allows to measure a significant difference ($t(15) = 2.15$, $p = 0.047$) in the mean persuasiveness distance measured for each system.

The EDEN Framework achieves an 8.25% stronger ($\hat{d} = 0.76$) persuasion than the purely planned system. The reasons that could explain such increase in persuasiveness are detailed in the following sections.

4.3.2 Persuasiveness and Expected Behaviour

When designing the EDEN Framework, the study of research in Persuasive Communication (see section 2.2) was used to set the hypothesis that human-computer

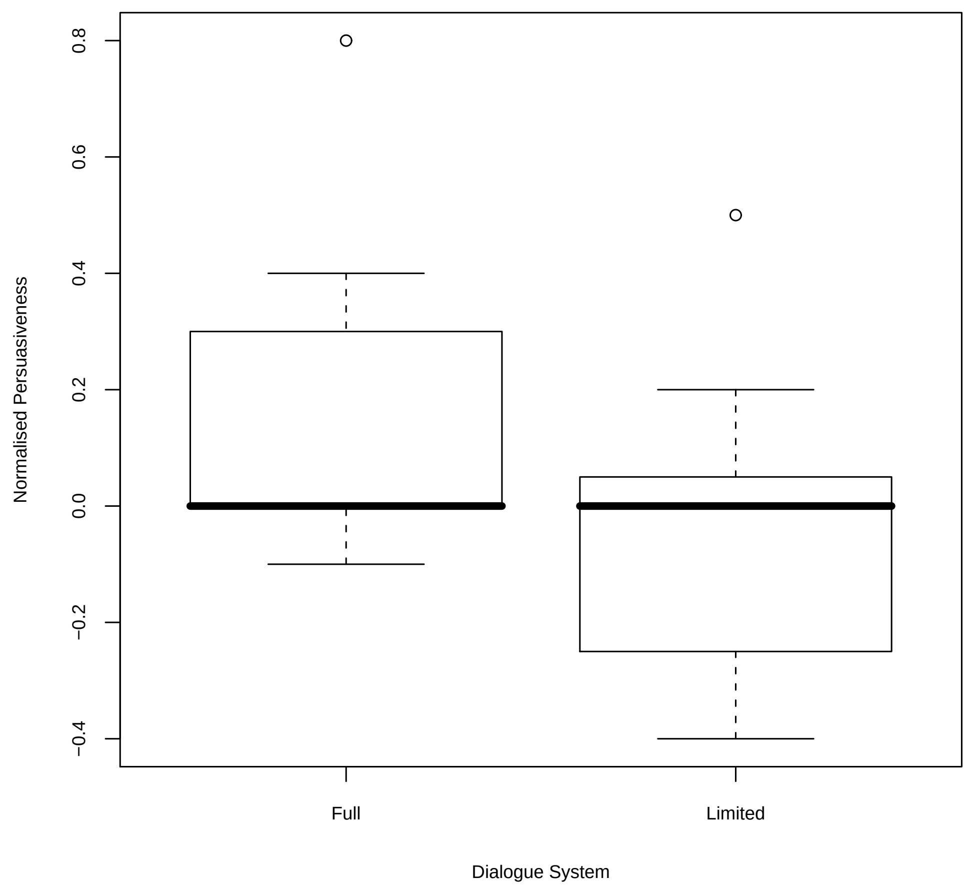


Figure 4.5: Comparison of the Persuasiveness Metric for the *Limited* and *Full* dialogue systems.

dialogue would be more persuasive if the system is able to make the user feel more comfortable by increasing the empathy of its response and using strategies to lower the user's frustration. The EDEN Framework goes towards this goal while focusing on simplifying the implementation of argumentative dialogue.

The participants are asked to give their level of agreement with a list of statements regarding their perception of the dialogue interaction and how easy it is to follow on a Likert scale: *Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree* (see questionnaire in Appendix C).

Some features of the dialogue show little impact on the persuasiveness, in particular the answers to the statements regarding the quality of the dialogue show no correlation with the persuasiveness and no difference of judgement are detected between systems. The participants are told they would interact with another human at the beginning of the dialogue, but none of them was "fooled" and they rapidly detected that they were chatting with a computer system. This might have led the participants to form a low expectancy of the dialogue system and made them give a good evaluation of the dialogue quality. In particular, the *pace of the dialogue* is judged similarly in both systems (Wilcoxon $T = 17$, $p = 0.66$) and has no significant correlation over the persuasion performances (Spearman $\rho = 0.01$, $p = 0.93$). *The ease of understanding of the arguments* show similar evaluation with no correlation (Spearman $\rho = 0.05$, $p = 0.74$) and no significant difference between systems (Wilcoxon $T = 7$, $p = 0.53$).

The *quality of interpretation of the system*, however, is judged differently for the *Full* system, for which the participants disagreed with the statement *"In this conversation, the computer interpreted correctly what you said"* while participants mainly agreed to the statement when using the *Limited* system (see figure 4.6). The repartition of the user's counter-arguments in the two systems is not different, however, the EDEN Framework argues more with the users and some might have understood this as misinterpretation of their utterance; in fact, the users that thought the system had a bad quality of interpretation used *more*

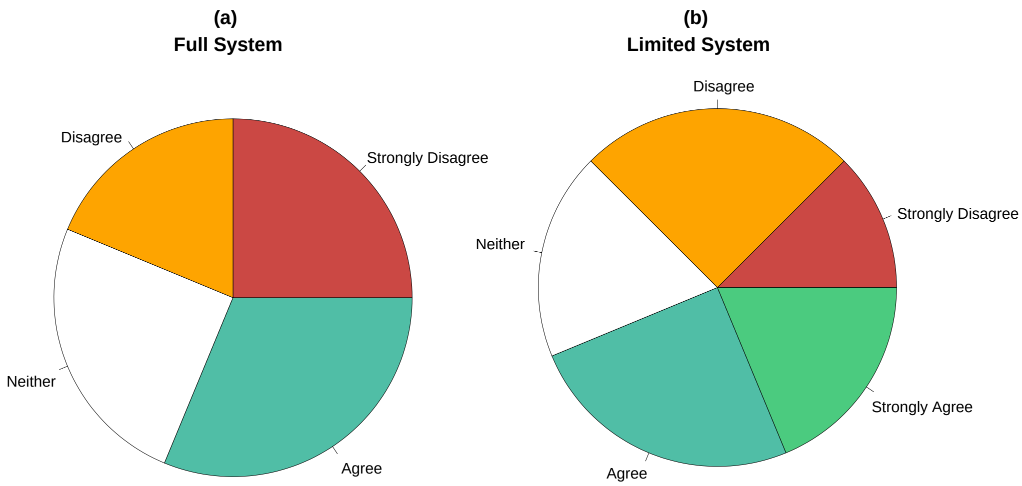


Figure 4.6: Answers to the statement: *"In this conversation, the computer interpreted correctly what you said."* for each system. Users mainly agreed to the statement in the Limited system (b) while they disagreed for the Full system (a).

disagreement[5] per dialogue (Spearman $\rho = -0.57$, $p < 0.01$) and the system had to *drop more persuasive goals* to finish the dialogues (Spearman $\rho = -0.42$, $p = 0.015$). This increased dropping of goals might have created a lack of consistency in the dialogue and generated the difference in evaluation of the systems ($\hat{d} = 0.39$). However, even if the perception of the quality of interpretation is significantly different between systems (Wilcoxon $T = 28$, $p = 0.014$), there is no measurable impact of this judgement over the persuasion performances of the systems (Spearman $\rho = 0.04$, $p = 0.82$).

The answers to the statement *"In this conversation, the other user interacted with you the way you expected it would."* show an interesting evaluation from the participants. However, these results should be interpreted carefully, indeed, the participants were told they were going to face a human interlocutor and might have had higher expectations of the interaction than if they had been told they would interact with a computer. Thus, there might be a bias in their initial expectations and their judgement of the interaction; however, this bias will be the

[5] According to the classifier discussed in section 3.9

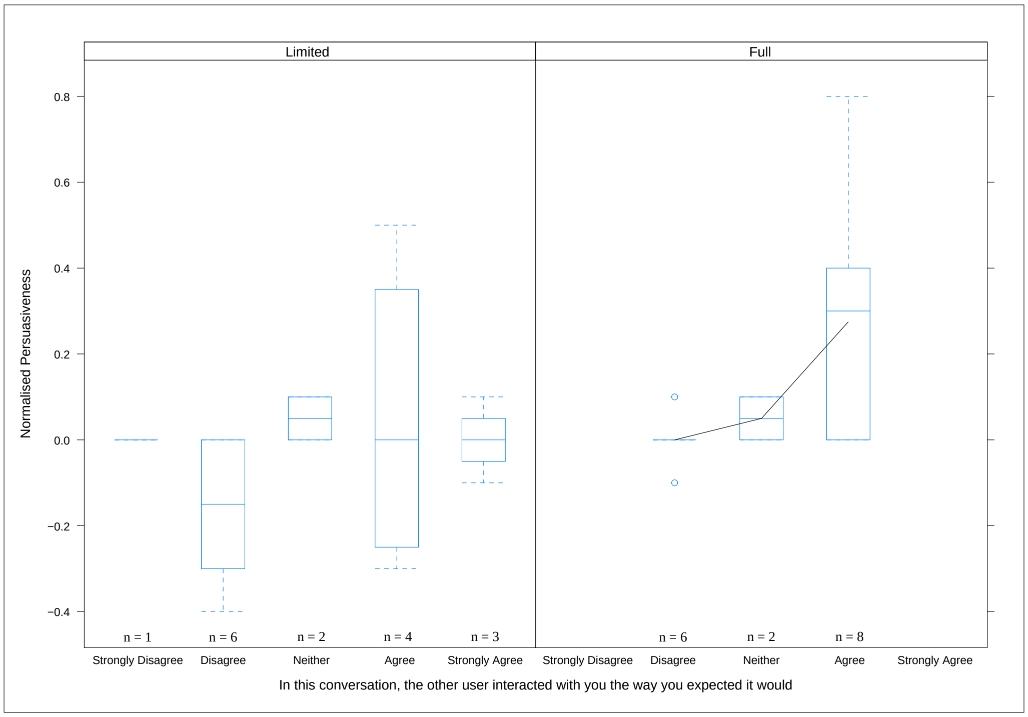


Figure 4.7: Answers to the statement: *"In this conversation, the other user interacted with you the way you expected it would."* and their influence on the persuasiveness distance regardless of the system used. The *Full* system shows a correlation between the answers and the system's persuasiveness (Spearman $\rho = 0.56$, $p = 0.02$).

same for both dialogue systems as they are both initially presented as being human interlocutors.

The participants, having been told that the interlocutor would be human, were unhappy with the computer interaction and this impaired the persuasiveness of the system. The answers to this statement, while showing no significant difference between systems (Wilcoxon $T = 37.5$, $p = 0.94$), have an impact on the dialogue persuasiveness (Spearman $\rho = 0.36$, $p = 0.04$), in particular with the *Full* system (Spearman $\rho = 0.56$, $p = 0.02$) as shown in Figure 4.7. Indeed, the participants that thought the computer system behaved as they had expected (*Agree* and *Strongly Agree*) change more their mind (mean persuasiveness of 0.16, $SD = 0.29$) than the participants that were unhappy with the system interaction (*Strongly Disagree* and *Disagree*; mean persuasiveness -0.07, $SD = 0.15$; $\hat{d} = 0.89$).

The results reported in Figure 4.8 hold a key to the difference in the perception and expectation of the participants. The participant's age seems to have an impact on their perception of the dialogue interaction. Indeed, the younger participants might be more used to interact through computer chats as well as interaction through natural language with computers, in which case they would have known better what to expect from the system; also, they might have formed lower expectations on the quality of the human-human interaction they were going to face as they might be more used to the limitations of communication through chat interfaces. There is a significant correlation (Spearman $\rho = -0.44$, $p = 0.01$) between the age of the participants and their evaluation of the system, in particular with the *Full* system (Spearman $\rho = -0.61$, $p = 0.01$). Indeed, there is an important difference in age ($\hat{d} = 0.97$) between the disagreeing group ($M = 38.8$, $SD = 13.2$) and the group of participants that are happy with the interaction ($M = 27.3$, $SD = 7.55$).

These results confirm part of the initial hypothesis, showing that participants that are less frustrated by the system interaction are easier to persuade.

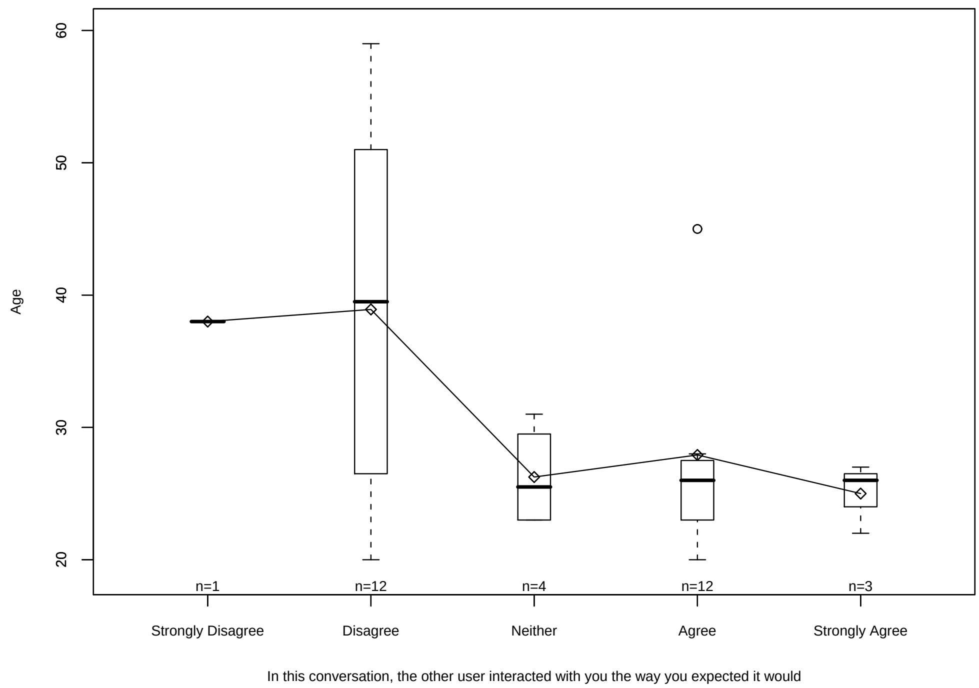


Figure 4.8: Answers to the statement: *"In this conversation, the other user interacted with you the way you expected it would."*. There is a correlation between the participant's evaluation of the system's behaviour and their age (Spearman $\rho = -0.44$, $p = 0.01$). n is the number of participants that used this particular answer.

4.3.3 Trust and Coercion

Another key feature of the persuader that improves persuasiveness is to build the interlocutor's trust (Fogg 2003, Stiff & Mongeau 2002). The trustworthiness of computers is often taken for granted by users (Fogg 2003, van Mulken, André, & Müller 1999) but the system has flaws. Users usually assume that the data presented by a computer are trustworthy even if this is changing with the growth of Internet scams. In this experiment, and in both conditions, the participants were told that they would face a human interlocutor so might have had different expectation on the trustworthiness of the data provided. From the observation of the dialogue session logs and comments from the participants, it appears that some of them were not fooled by the system and detected that they were chatting with a computer. This might have affetected their trust, either negatively as they had been deceived by the instructions, or positively as they would trust more a computer. However, the setup being similar in both conditions (limited system and full system), this factor will have an equal effect and the participants trust answers are thus still interesting to compare.

In the evaluation, there is no significant impact of trust over persuasion (Spearman $\rho = 0.02$, $p = 0.90$). In particular, the participants do not show a difference of judgement between systems (Wilcoxon $T = 24.5$, $p = 0.07$). The *full* system – that is more persuasive and uses longer argumentation with the user – is not judged less trustworthy than the *limited* system and overall, the participants answered to the statement "*The other user was trustworthy.*" between the neutral *Neither Agree nor Disagree* (28.1%) and the agreement (56.3%; see Figure 4.9).

A study of the log of the dialogues using the EDEN Framework shows a mean of 6.1 ($SD = 4.4$) reactive utterances per dialogue – i.e. defences to an argument – versus zero for the *limited* system which does not, by design, insist on an argument if the user disagrees. The previous results show that it does not impair the trust given to the system, however it could appear *coercive* and forceful to the user.

The answers to the statement "*The other user was* ***not*** *forceful in changing*

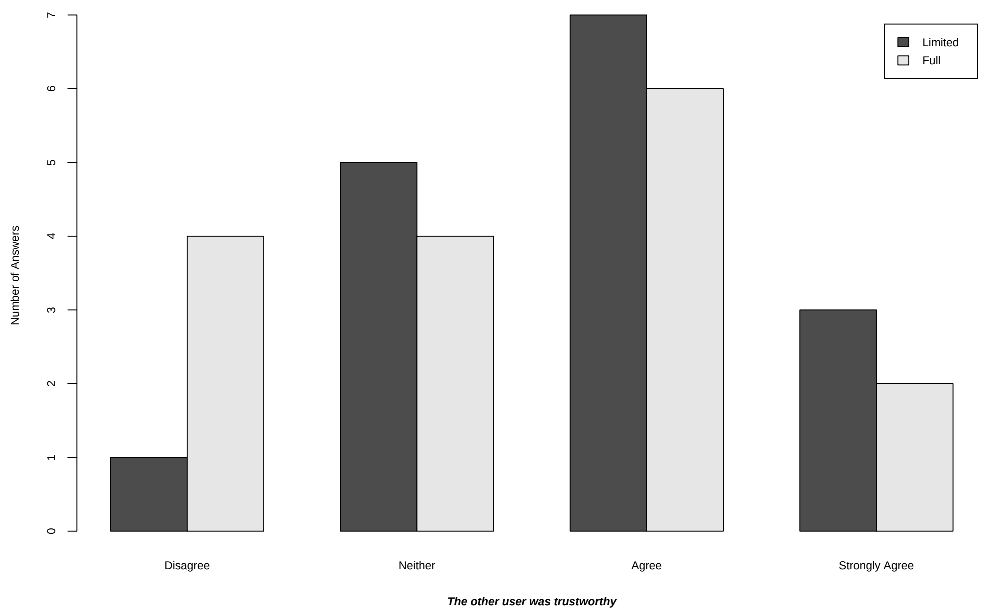


Figure 4.9: Answers to the statement: "*The other user was trustworthy.*" by system.

your opinion." (reported in Figure 4.10) show that participants do not report significantly different feelings about the coercion of each system (Wilcoxon $T = 23$, $p = 0.52$). In average, participants agreed (59.4% *Agree* and 9.4% *Strongly Agree*) that the system does not try to force them to change their ranking and there is no impact of their judgement of coercion on the system's persuasiveness (Spearman $\rho = -0.17$, $p = 0.35$).

4.3.4 Perceived Persuasion

In the questionnaire, the participants were directly asked, for each system, if they thought that "*The other user was persuasive.*". The answers to this statement provide an idea of the *perceived persuasiveness* of the system to compare with the *measured persuasiveness*. When the *full* system achieves a measured persuasive-

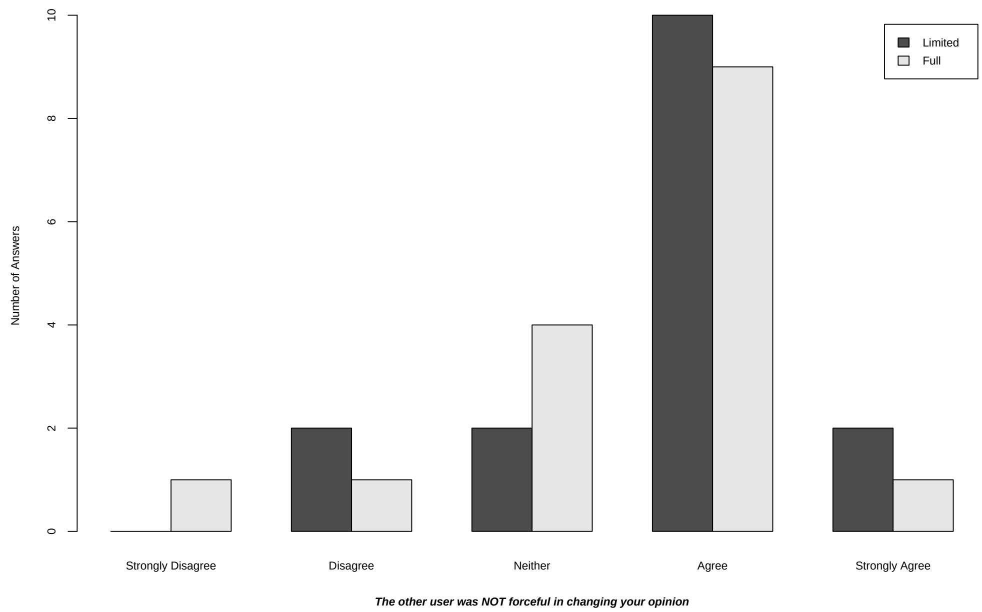


Figure 4.10: Answers to the statement: "*The other user was* **not** *forceful in changing your opinion.*" by system.

ness over seven times higher than the *limited* system, the report of the *perceived persuasiveness* (see Figure 4.11) shows no significant difference between the two systems (Wilcoxon $T = 29$, $p = 0.75$). Participants are mainly split between no specific perception of persuasion (28.1% answered *Neither Agree nor Disagree*) and a little perception of persuasion (46.9% answered *Agree*) and their judgement is not correlated with the measured persuasion (Spearman $\rho = 0.09$, $p = 0.63$).

4.3.5 Conclusion

Persuasiveness requires that the system does not alienate the users by forcing their choices. In particular, if such a system is to be used on a long-term basis for applications requiring multiple dialogues, the dialogue will need to keep the user involved in the interaction by keeping trust and by avoiding coercion. This exper-

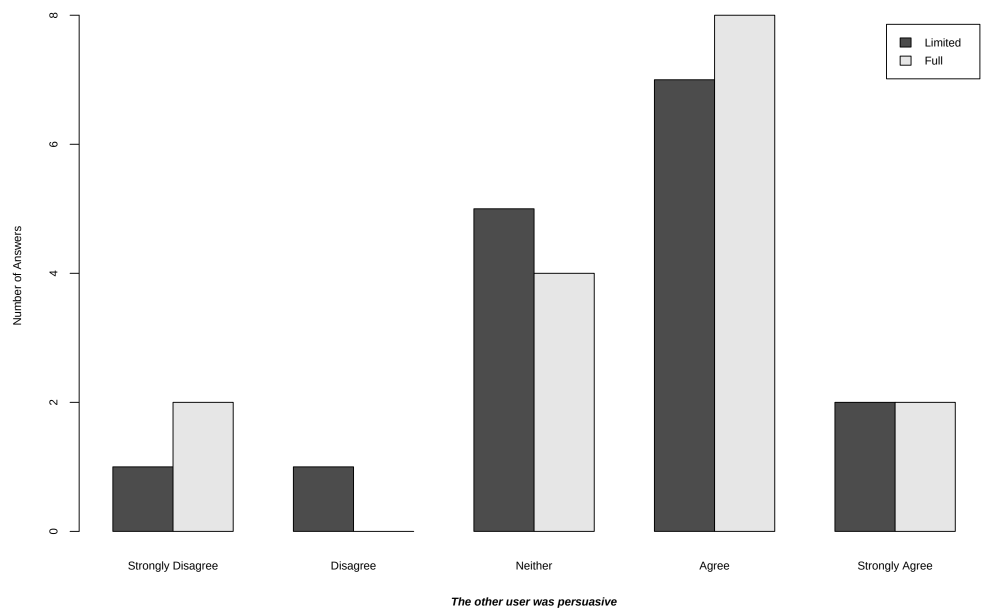


Figure 4.11: Answers to the statement: "*The other user was persuasive.*" by system.

iment shows that the EDEN Framework can be more persuasive than a standard planning approach; this is linked to the new layered approach that allows a better management of discourse level reactions as well as simpler integration of reactive argumentation strategies. The results of the Desert Survival Scenario experiment show that the dialogue model and framework design proposed in this thesis achieve better persuasiveness without influencing the perception of the users, in particular, the user's perception of persuasion or coercion is not influenced by the increase in persuasiveness.

Chapter 5

Personality and Persuasion

The desert scenario is well suited to evaluate the ability of the EDEN Framework to explain complex arguments supported by a detailed *descriptive* belief hierarchy. However, none of the participants had a prior expertise of the domain nor did they form personal preferences about the items. Two experiments were designed, using a restaurant recommendation domain, to explore different parameters of the persuasion, to evaluate the influence of *prescriptive* beliefs, preferences matching and personality on the overall interaction.

5.1 Restaurant Recommendation Scenario

To evaluate the impact of the system's personality, the Personage generator (Mairesse & Walker 2007) is used to replace the canned text generation engine used in the previous experiment (see section 3.7.3 for details). The Personage generator has a set of generation knowledge available for a restaurant recommendation domain and thus the second experiment is designed around this domain.

In this experiment, the Personage generator (Mairesse & Walker 2007) is used to simulate different personalities for the same content. The Personage generator uses trainable Natural Language Generation to map the Big Five personality traits (Norman 1963) to generation parameters. This training is performed in three steps:

1. Over-generate random surface forms for the same semantic content with random generation parameters.

2. Present these generations to expert annotators that rate the perceived personality of the texts with a Big Five traits inventory (Gosling, Rentfrow, & Swann 2003).

3. Use this annotated corpus to train a regression model to map between personality traits and generator parameter.

When evaluated, the Personage generator is able to generate recognisable personalities. In particular, Mairesse & Walker (2007) reports that, for the introvert/extrovert trait taken as a binary value[1] the generation parameter predicts the human annotation with a 10.8% error. The experiment reported also evaluated the *naturalness* of the generation and showed that annotators found the generation natural. However, this generator has been tailored for speech generation and it is used here for text only interaction, so the naturalness of the utterances might be lower in this experiment.

The task of the user is designed to allow the use of the persuasiveness metric (see section 4.1.2) and the dialogue system is not implemented as for a typical recommendation task. Instead, the user's task is to pick an initial ranking of three preferred restaurants out of a detailed list of restaurants. The restaurants database available to Personage contains a large number of restaurants, but to avoid information overload during the choice, only ten randomly picked restaurants are used.

Once the user has entered a preferred ranking R_i, the task of the system is, as for the desert scenario, to change this ranking to its own goal ranking R_s. In

[1]The Big Five extroversion scale ranges from 1 – introvert – to 7 – extrovert –. Mairesse & Walker, in the evaluation, splits the scale in two binary personality traits around the neutral extroversion rating 4.

this task to ease the understanding of the persuasiveness metric, R_s is always the reverse of R_i, so $K_\tau(R_i, R_s)$ is equal to the maximum distance possible:

$$K_\tau(R_i, R_s) = \frac{n \times (n-1)}{2}$$

where n is the number of restaurants ranked: 3. If the system is persuasive enough to make the user invert the initial ranking, the persuasiveness distance of the system is maximum and equal to: $\frac{n \times (n-1)}{2} = 3$. If the system does not succeed in changing the user ranking, then the persuasiveness distance is null. Persuasiveness is reported normalised between $[0, 1]$ for comparison with the previous evaluation results.

The Personage generation framework generates distinct surface forms based on five personality parameters (traits): Openness to experience, Conscientiousness, Extroversion, Agreeableness and Emotional stability. The current experiment is designed to study the influence of the *extroversion* parameter on the dialogue output and the user's perception of the interaction.

The participants were randomly split in two groups facing the same EDEN Framework dialogue system with two different instances of the Personage generator, but with the same knowledge base and available reactions. One group is faced with an *extrovert* generation while the other group is faced with an *introvert* system. All the other parameters of personality are identically set to their positive end.

5.2 Assessing User Preferences

Each restaurant is defined by five attributes: food quality, service, food type, decor and type. To plan the persuasion, the system needs to know about the user's preferences for these attributes (see section 3.5).

At the beginning of the dialogue, the user provides a ranking of its preferred restaurants. Indeed, it would ease the task if the system could infer automatically the user's attribute preferences from this ranking. In the field of marketing and economic psychology, some techniques have been introduced to elicit partworth preferences but they all require additional inputs from the user (Breivik & Supphellen 2003). In the field of Artificial Intelligence, partworth preferences are usually seen as a utility model; Schmitt, Dengler, & Bauer (2003), for example, uses the weight given to individual attributes to rank products with a linear utility model. Eliciting the weight of each attribute from a known ranking can thus be seen as the reverse problem; given the utility model that links attributes weight to a ranking, to elicit attributes preferences, the system needs to find the weight of each attributes leading to the observed ranking.

$$r_i = \sum_{j \in A_i} \beta_j \times w_{ij} \tag{5.1}$$

$$\vec{\beta} = (WW^T)^{-1} W^T \vec{r} \tag{5.2}$$

The problem of estimating the importance of each attribute from the restaurant ranking is defined as a linear regression problem. The user provides a set of observations for each attribute conditioning the rank of each restaurant. If the rank is assumed to be linearly correlated to the restaurant's attributes, Equation (5.1) can be defined for each restaurant where r_i is the rank of the restaurant, w_{ij} the value of the attribute j for the restaurant i, A_i the set of possible attributes and β_j the importance of the attribute j for the user (which is supposed to be independent of the restaurant i). Such equation can be written for each restaurant included in the user's ranking. The system is interested by the user's preference given by the unknown $\vec{\beta}$ vector of weights given to each attribute. Applying linear regression, this vector is resolved with the equation (5.2), where W is the matrix of attribute values, W^T its transpose and $\vec{r}$ the vector of restaurant's ranks.

However, Equation (5.2) supposes that the matrix WW^T is non singular, but as the number of restaurant rankings is smaller than the number of attributes, this matrix is always singular and the linear regression cannot be used.

In this experiment, the user task is to rank three restaurants out of ten. This has been decided to ease the task and make it more realistic – when friends go out to a restaurant, they do not start by ranking twenty different restaurants, but they start with a shortlist and then negotiate on the best restaurant. A simpler approach is thus chosen for the current scenario. The preferences are elicited directly from the user through a questionnaire at the beginning of the dialogue following the method described by Walker et al. (2004) (developed with the SMARTER procedure Edwards 1994) as illustrated in section 5.3.1.

5.3 Initial Personality Experiment

An initial experiment to explore the effect of personality style generation on the system persuasiveness was designed. This experiment was thought for a very similar to the desert survival scenario experiment but with different evaluation question and on a different domain. The following sections discuss the procedure and results of this exploratory experiment.

5.3.1 Procedure

The procedure is similar to the Desert Survival Scenario (see chapter 4), however, in this experiment, each participant takes part in only one dialogue session to set up a *between-subject* experiment. Each participant faces the full EDEN dialogue system using the Personage generator for utterances realisation, the between-subject variable is the personality of the system as participants are randomly assigned to either an introvert system or an extrovert system.

The personality parameter is treated in a binary fashion with two conditions: introvert or extrovert. The Personage generation parameter for this personality trait are set to either of the extroversion scale extremes. The rest of the personality traits are always set to the positive end of the scale.

The participants are first presented with a welcome screen to the scenario:

Restaurant night out Scenario

Thank you for participating in this experiment. You will have to get into a scenario and have a chat with another user through an instant messaging system. Please be sure to read all the instructions and reply to the little questionnaires given to you along the way.

Please, try avoiding going back in the steps or reloading the pages.

Both you and the other person have been told that the purpose of the chat is to agree on a short list of restaurants where you could go out with your friends tonight.

The participant then goes through two short steps assessing its preferences of the restaurant attributes. The user is asked:

Knowing more about your preferences

Imagine that for whatever reason you have had the horrible luck to have to eat at the worst possible restaurant in the city.

The food itself is terrible, it has terrible service, the decor is ghastly, the price is 100£ per head, and you do not like the type of food they have.

Now imagine that a good fairy comes along who will grant you one wish, and you can use that wish to improve this restaurant to the best there is, but along only one of the following dimensions. What dimension would you choose?

- food quality
- service

- decor
- cost
- food type

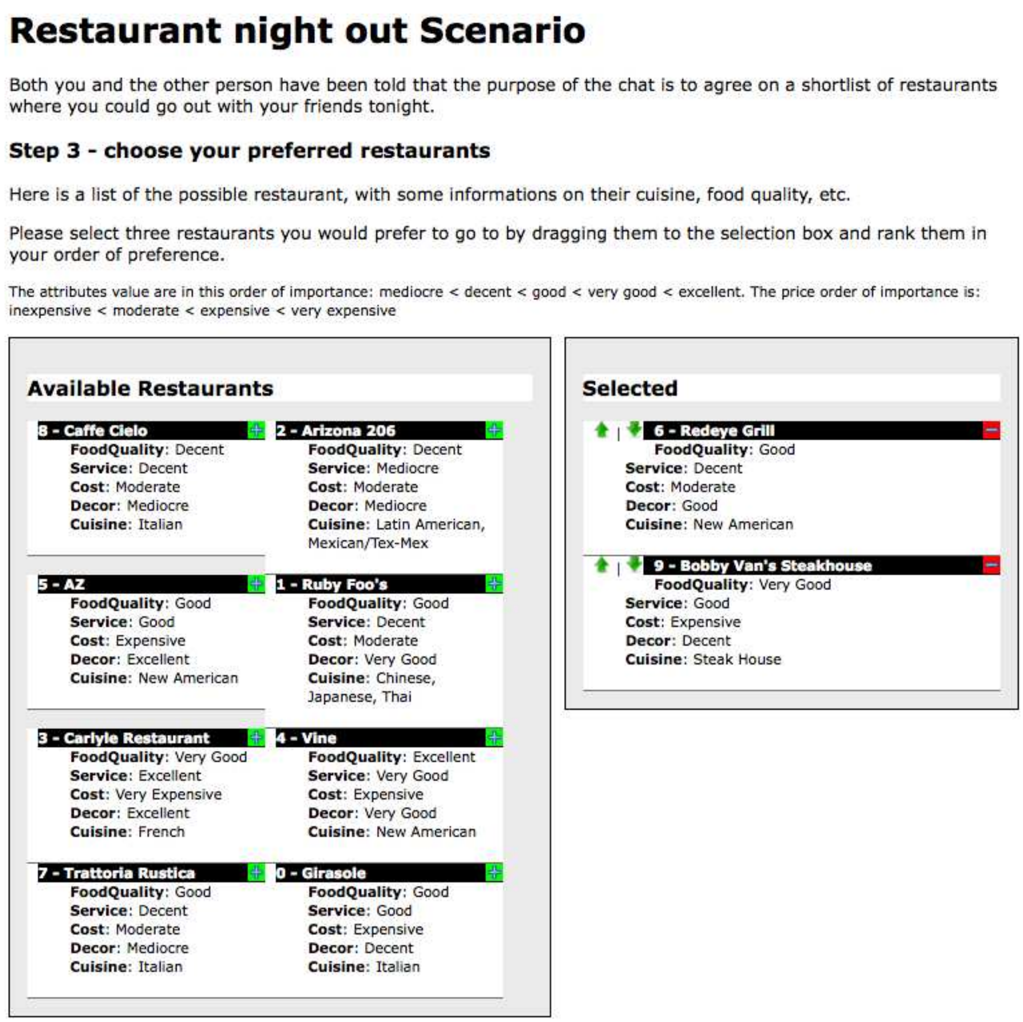
Once the user chooses one attribute, the same question is asked with this attribute removed, until only two attributes are left. This step provides an implicit ranking of attribute preferences that is then used in the dialogue planning.

The system also needs to know about the type of cuisine the user prefers. A list of all possible cuisine types found in the ten randomly selected restaurants is generated and presented to the user who can provide a preference:

Step 2 - Knowing more about your preferences

Could you please tell us your likes and dislikes for cuisine type.

- Chinese, Japanese, Thai
  ☐ Like ☐ Dislike ☐ I don't care
- Steak House
  ☐ Like ☐ Dislike ☐ I don't care
- French
  ☐ Like ☐ Dislike ☐ I don't care
- Italian
  ☐ Like ☐ Dislike ☐ I don't care
- Latin American, Mexican/Tex-Mex
  ☐ Like ☐ Dislike ☐ I don't care
- New American
  ☐ Like ☐ Dislike ☐ I don't care



(Original in colour)

Figure 5.1: Screenshot: Ranking Restaurants

The participant is then presented with the choice of restaurant and asked to choose three of them (Figure 5.1). The system then creates goals – in a manner similar to the desert survival scenario – and plans a dialogue session with the participant.

During the dialogue (see figure 5.2), participants have access to a reranking module to keep track of their current ranking. This is used as a *beliefs monitor* by the system (see section 3.10.1) to keep track of the real beliefs change in the user's mind.

After the dialogue ends, the user is proposed to give a final ranking of the restaurants and finally fills in a questionnaire about the dialogue (the full questionnaire is given in Appendix D).

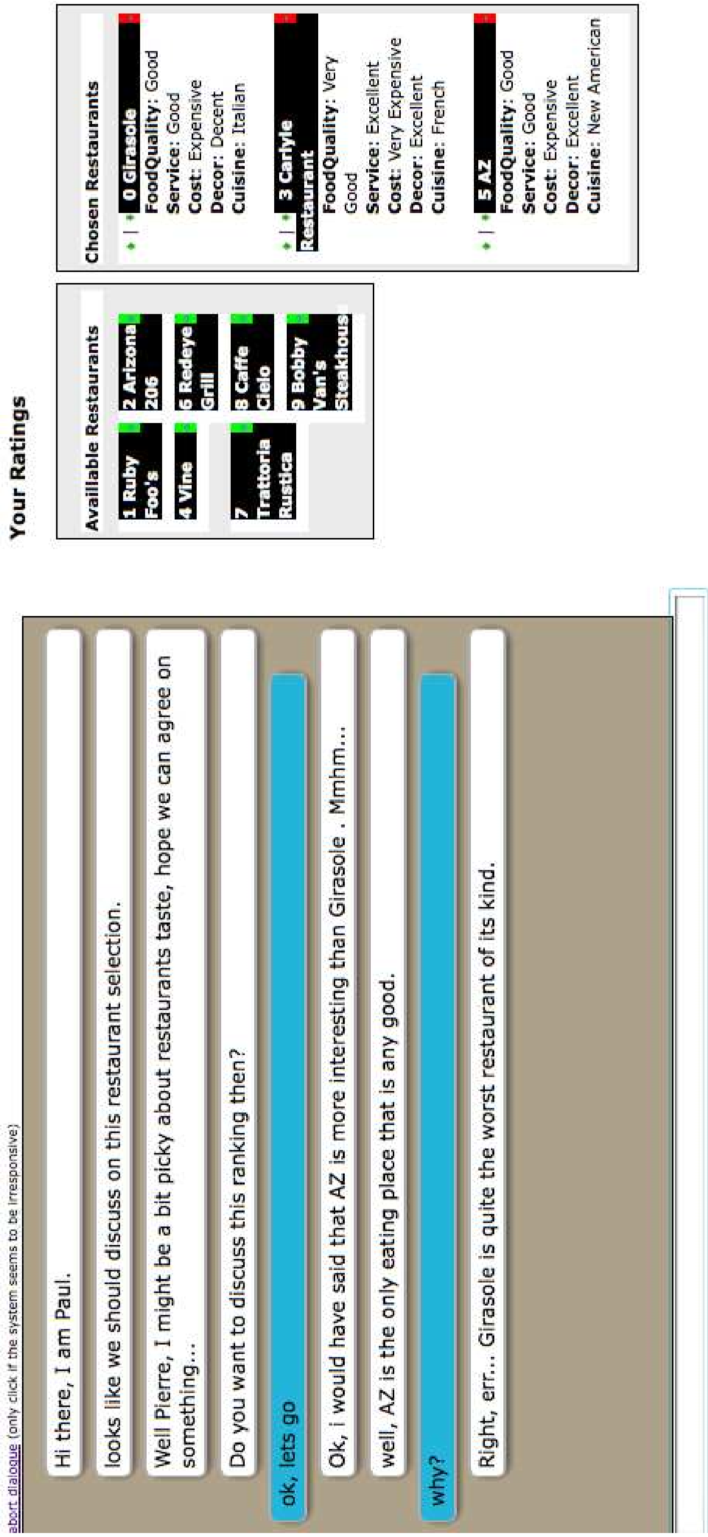
5.3.2 Results

The restaurant study is designed to evaluate the influence of the system's displayed personality, in particular its extroversion, on its persuasiveness. The participants were faced with either an *extrovert* or an *introvert* system; with the difference of personality being managed at the generation level by the Personage framework (Mairesse & Walker 2007).

Twenty-eight participants took part in the restaurant task. They were randomly split in a group of eighteen participants facing the *extrovert* system and a group of ten facing the *introvert*. For each dialogue, the persuasiveness is measured by using the metric described in section 4.1.2.

Figure 5.3 reports the difference in persuasiveness for the two personalities. The system is trying to persuade each participant to perform two moves to achieve a distance of three adjacent swaps but the participants performed an average of 0.68 ($SD = 0.94$) swaps of restaurants towards the system goals. From the collected results, no difference between the *extrovert* and *introvert* systems' persuasiveness can be measured ($t(26) = 0.19$, $p = 0.85$).

While the Desert Survival Scenario did not display a link between how the participants evaluated the persuasion and the actual measured persuasiveness (see



(Original in colour)

Figure 5.2: Screenshot: Restaurant Dialogue Session with Introvert Personality

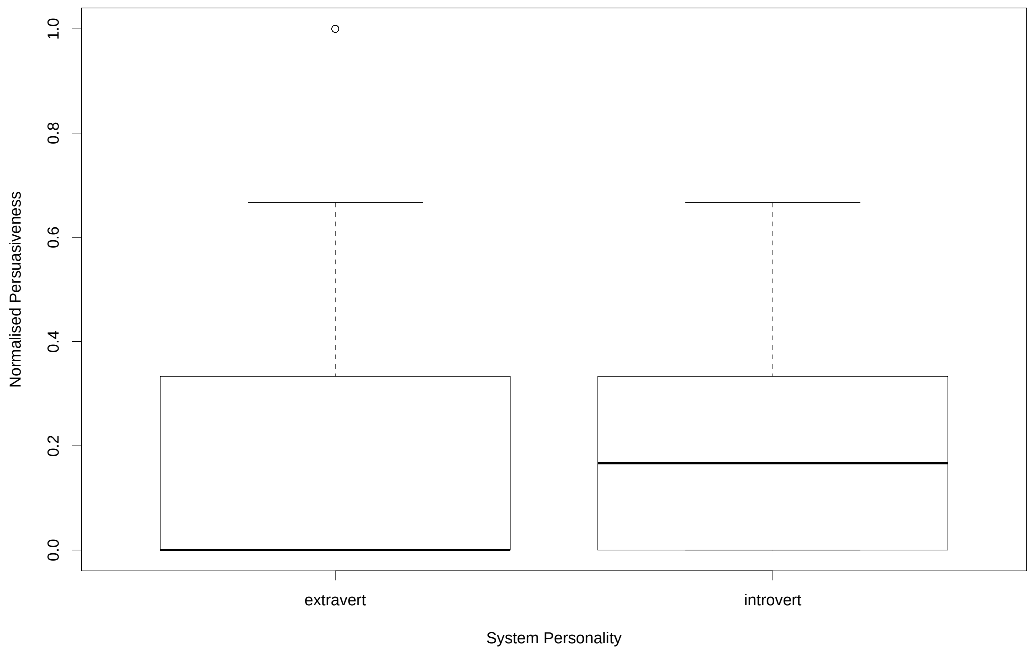


Figure 5.3: Persuasiveness of the system according to its extroversion. There are eighteen observations for the *extrovert* system and ten for the *introvert*.

section 4.3.4), the Restaurant domain has a different effect on the user's perception of persuasion. In the latter domain, users are familiar with the scenario and the items to choose from, in the past, they already had to choose between restaurants and argue about them with friends while they had no particular expertise in the Desert Survival Scenario. In addition, the system used in the Restaurant Scenario argues around users' preferences and this seems to impact on the perception of the system's persuasion. Indeed, the answers to the statement "*The other user was persuasive*" evaluating the perceived persuasiveness show a correlation (Spearman $\rho = 0.70$, $p < 0.01$) with the measured persuasiveness (see Figure 5.4).

In the case of the restaurant domain, the system does not appear as persuasive as during the previous experiment. The following sections try to explain the reasons that might have led participants to be less prone to persuasion.

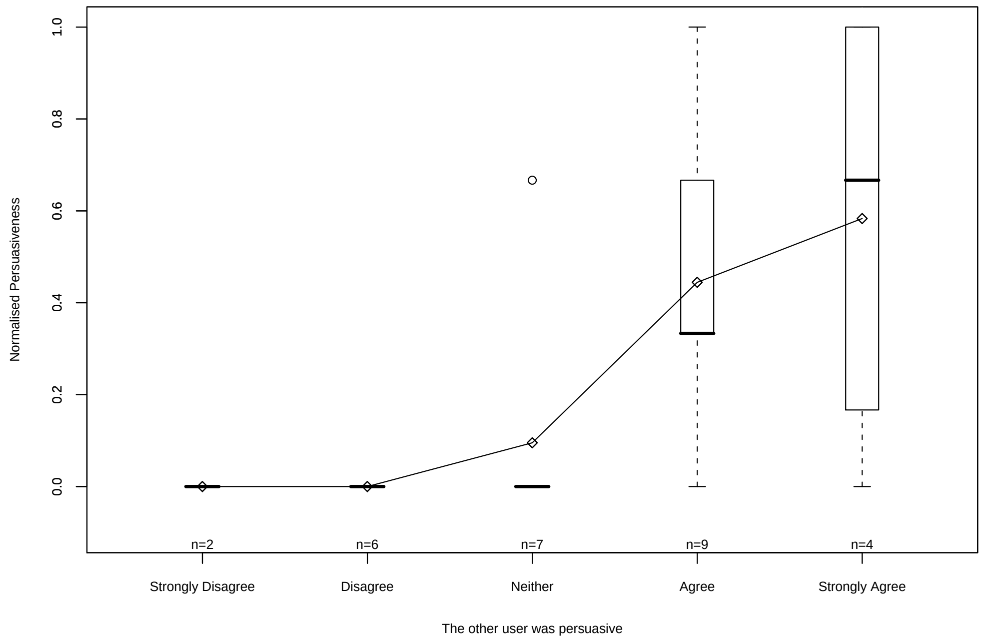


Figure 5.4: Correlation Between the Perceived Persuasion and the Measured Persuasion. n is the number of participants that used this particular answer.

5.3.2.1 Familiarity and Persuasion

Comparing the two controlled conditions showed no apparent influence of the system's extroversion on its persuasiveness. In the questionnaire filled by each participant, the statement *"I liked that the computer was outgoing."* is assessed to evaluate the extroversion perceived by the user. The answers to this statement might help in understanding why the system is not very persuasive and what can actually make a difference in its persuasiveness.

The Personage generator generates different surface forms for each personality, but a main theme is used along the generation. The generated sentences are often outgoing and using many colloquialisms with the user. Figures 5.5 and 5.6 are sample dialogues for – respectively – the *introvert* condition and the *extrovert* condition.

SYSTEM: Ok, i would have said that Redeye Grill is more interesting than AZ. Well, right, Redeye Grill is the only eating place that is any good.
USER: Why is that?
SYSTEM: Right, ok, AZ is quite the worst restaurant of its kind.
USER: okay
SYSTEM: Ok, i would have said that Redeye Grill is more interesting than Arizona 206. Mmhm. . . well, Redeye Grill is the only eating place that is any good.
USER: Can you explain that?
SYSTEM: Right, err. . . Arizona 206 is quite the worst eating place of its kind.
. . .

Figure 5.5: Sample Introvert Dialogue in the Restaurant Domain

The system, in both cases, is perceived to use an outgoing and friendly text that is not appreciated by all the participants (see Figure 5.7). some of the participants like the outgoing characteristic of the system (32.1% *Agree* and 3.6% *Strongly Agree*), 57.1% do not express any preference (*Neither Agree nor Disagree*) and 7.14% clearly state they do not like the system being outgoing (*Disagree*). No significant difference in repartition between the two personality conditions is measured (Mann-Withney $U = 101.5$, $p = 0.55$).

The participants that were happy with the system being *outgoing* accept the persuasion more easily (see Figure 5.8). In the reported comments from the users, two groups are detected. The first group of participants likes that the system is more familiar and outgoing, the other group is disturbed by such behaviour coming from a computer system and the participants state that they do not like a computer “talking” to them this way. When considering the persuasiveness of the system in the light of these stated preferences (see Figure 5.8), a significant

SYSTEM: I thought you preferred food quality over the decor. Caffe Cielo seems to be better in that sense than Redeye Grill. Caffe Cielo seems more like something for you compared to Redeye Grill.

USER: ok, I guess you are right

SYSTEM: Ok, Caffe Cielo seems a better choice than Bobby Van' s Steakhouse. Actually, I know Caffe Cielo is darn alright, you know buddy!

USER: I think I prefer the steakhouse though

SYSTEM: Yeah, actually, I would recommend Caffe Cielo, it' s one of my favourite places, you know mate!

USER: really?

SYSTEM: For real.

USER: any particular reasons?

SYSTEM: Ok, hum, why choose Bobby Van' s Steakhouse over Caffe Cielo? Actually, I know you would adore Caffe Cielo mate, you know!

USER: right

SYSTEM: I am glad we agree.

SYSTEM: Bobby Van' s Steakhouse seems a better choice than Redeye Grill. I believe Redeye Grill isn' t great mate. Actually, the service is just darn friendly and it provides like, nice atmosphere, and adequate food, you see?

USER: can you explain that then?

SYSTEM: well, I would say that Bobby Van' s Steakhouse is better than Redeye Grill. I know you wouldn' t adore Redeye Grill! Actually, the service is just really friendly and this restaurant features like, nice atmosphere, and nice food, you see?

...

Figure 5.6: Sample Extrovert Dialogue in the Restaurant Domain

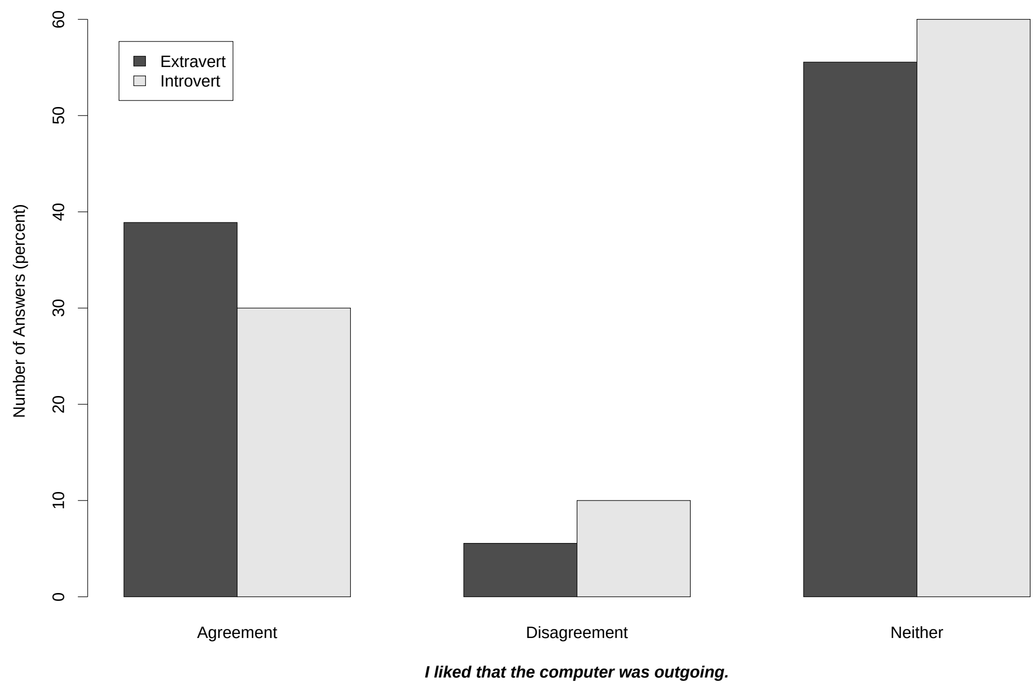


Figure 5.7: Answers to the statement: *"I liked that the computer was outgoing."* by personality

difference in the persuasiveness of the system is measured when the participants express a positive preference and when they do not express any preference or state they do not like the outgoing system ($\hat{d} = 0.66$; Mann-Whithey $U = 30.5$, $p < 0.01$).

Using this observation to parameterise the style of generation used by the system could improve the overall persuasiveness of the dialogue. In the current study, the participants are asked to fill in a short questionnaire to measure their personality traits according to the Big Five inventory (Gosling et al. 2003). In particular, this questionnaire provides a measure of the participants' *extroversion* that ranges between zero (hightly introvert) and seven (hightly extrovert)[2]. The results from the evaluation of the user's personality show that extrovert partici-

[2]Note that the controlled personality condition of the system's is binary: introvert or extrovert whereas the evaluation of the user's personality is on a continuous scale.

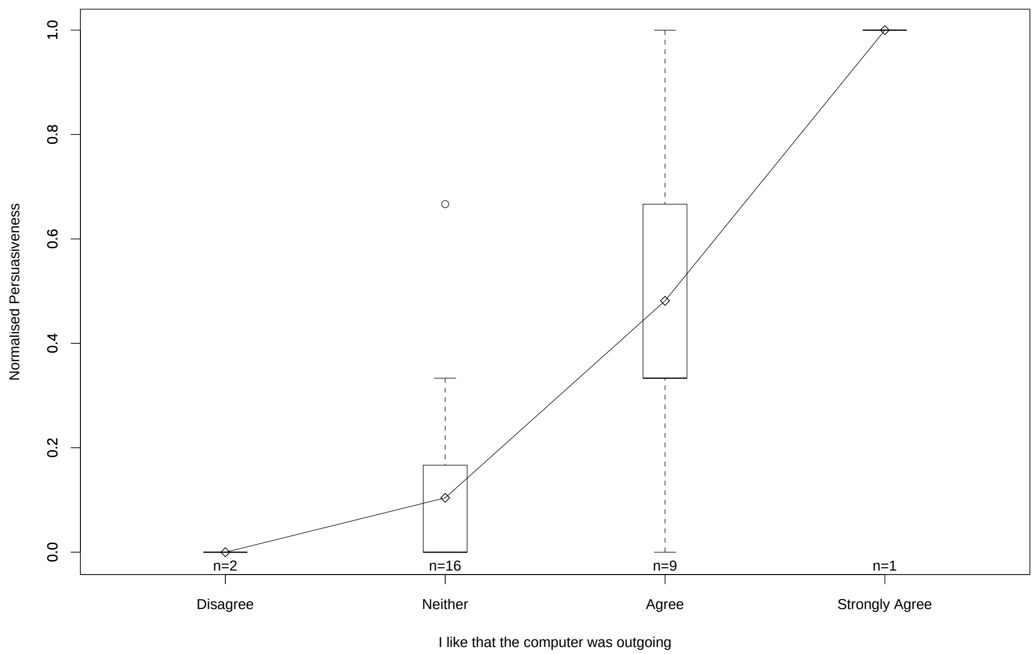


Figure 5.8: Answers to the statement: *"I liked that the computer was outgoing."* and their persuasiveness by personality. The participants that liked the system being outgoing are more easily persuaded (Spearman $\rho = 0.63$, $p < 0.01$).

pants – participants ranking high on the extroversion scale – prefer the "outgoing" type of system (see Figure 5.9). Indeed, there is a significant correlation between the extroversion scale and the answers to the statement "I liked that the computer was outgoing." (Spearman $\rho = 0.39$, $p = 0.04$). In particular, all the participants that disagree with the statement "I liked that the computer was outgoing." rank on the introvert side of the extroversion scale ($M = 3$, $SD = 1.4$).

The initial aim of the "I liked that the computer was outgoing." statement was to evaluate the link between the perception of the system's personality by the user and its persuasiveness. The assumption was made that "outgoing" was a colloquial synonym for "extrovert" that would be easier to understand by the users; however, the answers to this statement show that it might not map directly to the extraversion parameter of the generator as the answers to this statement are similar for both systems. The answers to this statement are still very interesting

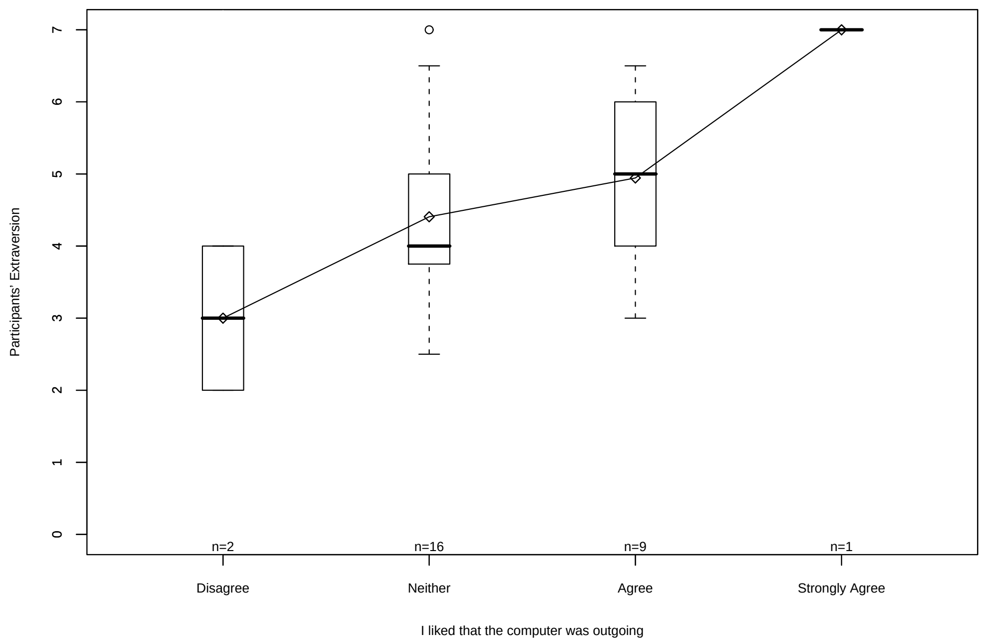


Figure 5.9: Correlation between the participants' extroversion and their answers to the statement: "*I liked that the computer was outgoing.*". The extroversion scale is a measure of the participant's extrovert/introvert personality trait within the Big Five inventory Gosling et al. (2003); seven is the maximum on the scale and corresponds to highly extrovert participants; zero corresponds to introvert participants. The line shows the correlation between the mean measures of extroversion and the participants' answers (Spearman $\rho = 0.39$, $p = 0.04$). n is the number of participants that used this particular answer.

as they show a significant correlation between the users' personality, how they like the system's displayed personality and its persuasiveness. This shows that personality tailoring will have an impact on the system's performances.

While there are not enough observations available to discover all the features of the users that would predict which type of generation to use, it seems that the participant's own personality might have an importance in their preference. This confirms the observations reported by Cassell & Bickmore (2002) that found that there was an interaction between the user's extroversion and the system style – in the case of Cassell & Bickmore, being either passive or initiating. These results also strengthen the observations of De Boni, Richardson, & Hurling (2008), which sets out to evaluate the effect of personality matching between the system and the user but couldn't find a significant influence in all cases.

In addition, this experiment does not provide data to show if an "outgoing" system would be more persuasive than a "normal" system towards the participants that like this feature. A new experiment design is needed to validate the real benefits of an outgoing system. In the meantime, the observations from this experiment support that a non-outgoing system might be preferable as 100% of the participants that stated not liking an outgoing system were not persuaded at all and in the case of the *extrovert* system, 75% of the participants that did not express any preference were not persuaded at all.

5.3.2.2 Influence of Personality on the Perception of the Dialogue

The controlled condition used for the generator did not allow finding a correlation between the extroversion parameter and its persuasiveness even if other features of the system personality make a difference in the actual persuasion. However, the extroversion parameter illustrates the difference of perception from the user.

The questionnaire designed for the Restaurant experiment contains a second statement relating to the system relationship with the user. The participants were asked to rate the statement: "*The computer acting like he knew you was irritating.*" to asses their perception of the system's familiarity.

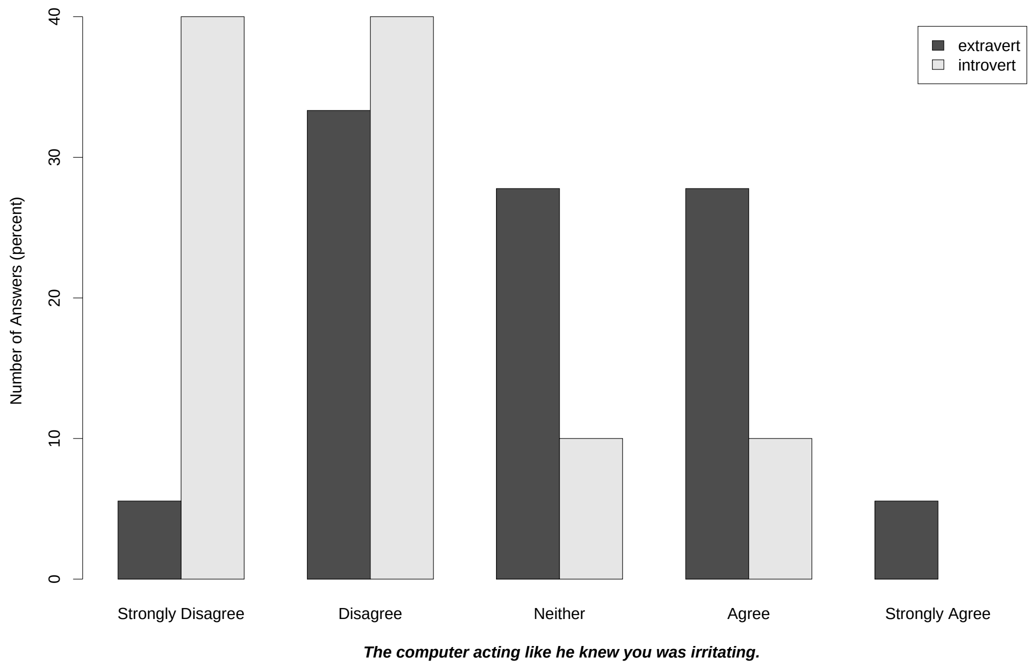


Figure 5.10: Answers to the statement: "*The computer acting like he knew you was irritating.*" by system's personality

The participants facing the *extrovert* system are more irritated by the system familiarity than the users using the *introvert* system. In fact, the answers to this statement (see Figure 5.10) show a significant difference ($\hat{d} = 0.92$) between the two generation conditions (Mann-Withney $U = 138$, $p = 0.02$). This is due to a similar effect as the *outgoing* difference showed before, as the *extrovert* system tends to call the user "mate" or "buddy" and to use gregarious invectives like "…you know!". Participants do not like a computer to use this level of familiarity with them and prefer a more reserved system. The *introvert* system, by design, generates shorter sentences using less bounding social cues and thus might seem less familiar to the participants. The familiarity of the system, even if it irritates the participants, does not show any correlation with the measured persuasiveness of the systems (Spearman $\rho = 0.02$, $p = 0.92$).

The Desert Survival Scenario showed that the age of the participants has an

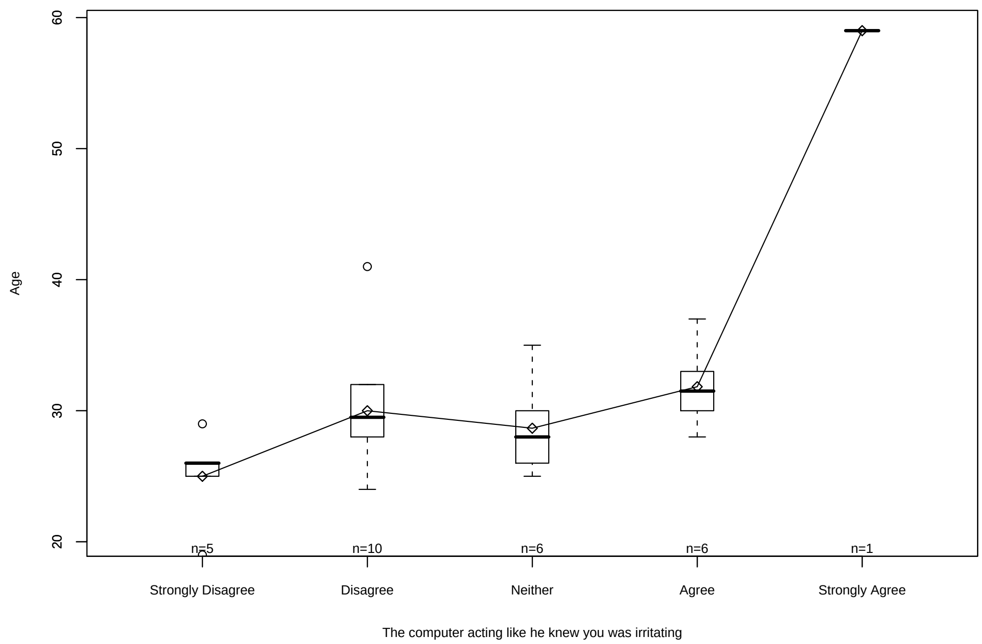


Figure 5.11: Correlation of the answers to the statement: *"The computer acting like he knew you was irritating."* by Age. n is the number of participants that used this particular answer.

impact on their expectation of the style of interaction in the dialogue (see section 4.3.2. A similar effect is found in the Restaurant experiment where older participants judged the system's familiarity more irritating than younger participants (see Figure 5.11). There is a weak but significant correlation between the judgement of the participants and their age (Spearman $\rho = 0.51$, $p < 0.01$) that could be explained, as before, by the fact that young adults might be more used to interacting through chat interfaces and other alternative human-computer interfaces.

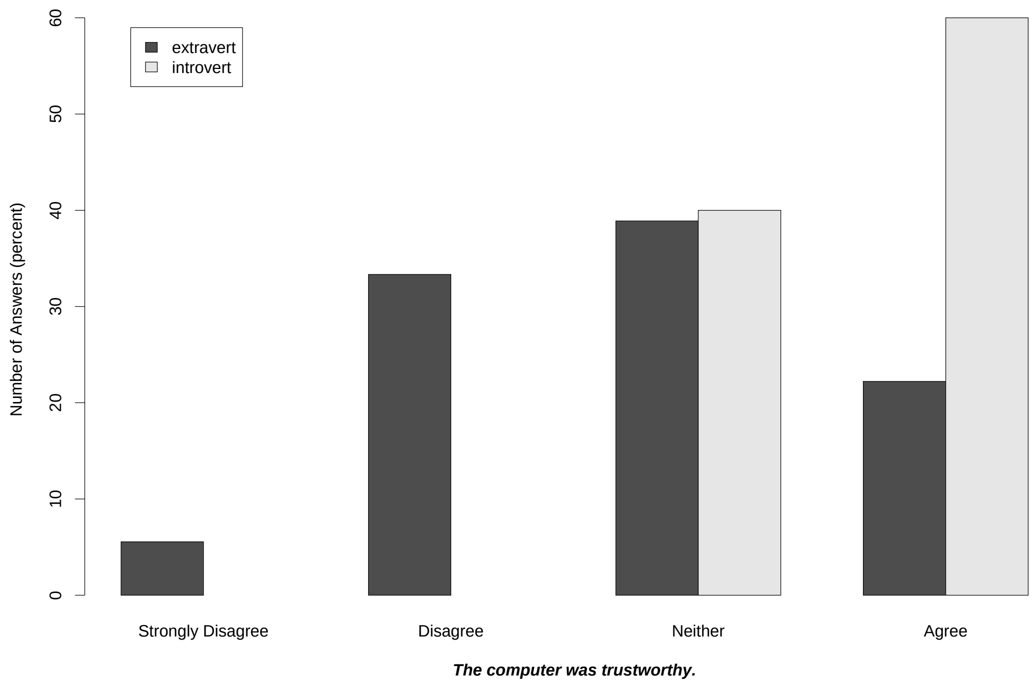


Figure 5.12: Answers to the statement: *"The other user was trustworthy."* by system's personality

The generation personality parameter also influences other features of the interaction. The participants found the system less trustworthy if they were facing the *extrovert* personality ($\hat{d} = -0.96$). In fact, a minority of the participants facing this personality answered *"Strongly Disagree"* (5.6%) or *"Disagree"* (33%) while 60% of the ones facing the *introvert* personality answered *"Agree"* and the rest were neutral when answering to the statement *"The other user was trustworthy."* (see Figure 5.12). This significant difference (Mann-Withney $U = 42$, $p = 0.01$) in the trustworthiness of the personalities might be explained by the results observed before; if the users are not happy with the way the system interacts, they do not see the system as trustworthy and the system is less persuasive. Indeed, the answers to the latter statement show a weak, but significant, correlation with the persuasiveness (Spearman $\rho = 0.38$, $p = 0.05$).

The behaviour of the *extrovert* system is also criticised by the participants. The plan and the reactive arguments available to both systems are exactly the same, each dialogue has to perform a similar persuasion, trying to achieve the same number of swaps; there is thus no difference in the dialogue management, the sole difference being created at the generation stage by the Personage generator. However, the judgement of the participants is different depending on which system they used; indeed to the statement *"In this conversation, the computer interacted with you the way you expected it would."*, 22.2% of the participants using the *extrovert* system disagreed in some way while 80% of the ones using the introvert system agreed ($\hat{d} = -0.82$; Mann-Whitney $U = 45$, $p = 0.02$). However, this difference of evaluation of the dialogue interaction is not reflected in its persuasion and no correlation is measured between the users' answers and the measured persuasiveness (Spearman $\rho = -0.06$, $p = 0.76$). As for the desert scenario experiment (see section 4.3.2), answers to this statement might be biased by the fact that the system was initially presented as being a human. The participants will have formed higher expectations than if they had been told it was a computer initially, however, the results reported here show a significant difference between the two generation personalities which were both presented identically to the participants, thus the bias in the expectation should have no effect on the results.

The Persuasive Communication textbook by Stiff & Mongeau (2002) as well as results in the computer science field (Fogg 2003) demonstrated that these features of interaction are important to achieve persuasion and keep the user in-

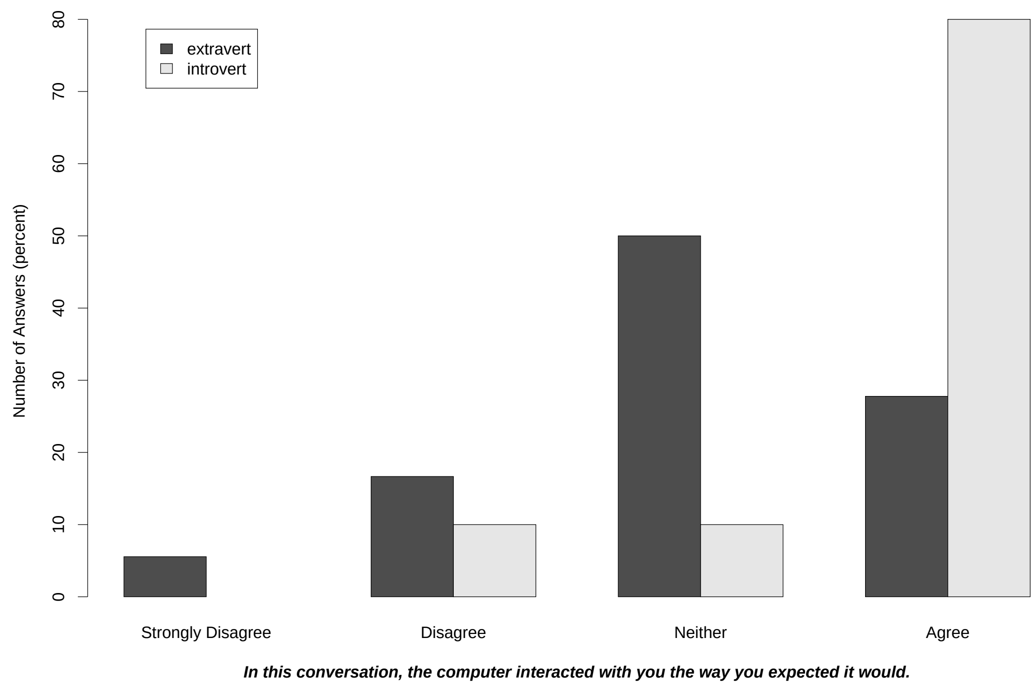


Figure 5.13: Answers to the statement: *"In this conversation, the computer interacted with you the way you expected it would."* by system's personality

volved in the interaction. In the group that is studied, the perception of the quality – as evaluated by the participants – of the dialogue is linked to the personality parameter of the generator but does not display a large observable impact on the measured persuasiveness.

5.3.3 Limits of the Restaurant Domain

The restaurant experiment is designed to use the Personage generator (Mairesse & Walker 2007) and study the impact of the system generated extroversion on the persuasiveness of the dialogue. The results that were collected during the experiment suggest a correlation between the extroversion parameter and the perception of the dialogue that does not translate in an observable impact on the persuasiveness of the EDEN Framework. The persuasiveness is showed to be influenced by another feature of the interaction and of the personality of the user. If the user likes the computer to be *outgoing*, then there is a direct impact on the dialogue persuasiveness.

The design of the restaurant experiment limits the conclusions that can be drawn from the results. In particular, the *extroversion* parameter does not appear to be the best control variable to observe the impact of the system's personality on the dialogue persuasiveness. To confirm the impact of the *outgoing* preference a different design is needed to control the right parameters. According to the prior studies in persuasive communication and human-computer interaction, the influence of the *extroversion* on the system *trust* and the user's perception of the interaction should have an impact on the persuasiveness of the dialogue. To confirm this hypothesis, a wider experiment has to be run involving a larger number of participants to be able to control their personality parameters and find a correlation with their perception of the interaction.

The restaurant domain has two drawbacks in its implementation that lead to an overall poor persuasiveness of the dialogue. The participant having chosen a ranking of three restaurants – for example $R_1 > R_2 > R_3$ –, the task of the dialogue system is to convince the user to achieve two moves: `reorder`$(R_3 > R_1)$ and `reorder`$(R_2 > R_1)$.

The first problem with the restaurant domain regarding these goal swaps is the number of available argumentation strategies. The Personage generator provides three operators for generating statements about the restaurant: a *recommendation*, a *comparison* and a *negative advice*. The system has two main arguments to make and has the choice of five generation operations for the whole dialogue:

- `recommend`(R_3)
- `recommend`(R_2)
- `negative_advice`(R_1)
- `compare`$(R_3 > R_1)$

- `compare`$(R_2 > R_1)$

If the participant does not argue with the system, there is no repetition, however, if the system has to generate counter-arguments for each goal argument, at least one of the generation operations has to be repeated.

In the 28 dialogues used to evaluate this experiment, a mean of 2.3 ($SD = 1.9$) defences per dialogue is found which requires at least one repetition of an operator per dialogue. The domain implemented in the Personage generator provides lexical variations for each operator as well as a variation on the restaurant attributes used in each generated statement. However, the structure of the statement remains very similar along each generation and the participants commented that there was "too much repetition".

The repetition in the dialogue might be the reason for the low persuasiveness of this experiment. To improve the quality of the dialogue, the variation in the generations needs to be increased as well as the possible argumentation strategies for this domain.

The available implementation for this domain also raised a difficulty in arguing about the user's preference. Each restaurant in the randomly selected list presented to the user has a different *cuisine type*. In the study of the dialogue sessions for this dialogue, participants rank as first restaurant one that matches their preferred cuisine. The system is then unable to use this preference to hint at a reordering of two restaurants, it is limited to use either a logical (descriptive) argument that an attribute is better than another one or a preferential (prescriptive) argument on the attributes. To these arguments, participants answered utterances similar to "but I feel like a French restaurant tonight" or "well, I prefer Italian over a steakhouse", the cuisine type seems to be a more valued preference than the other attributes, rendering the argumentation of the system difficult.

The experiment design should be changed to have control of this attribute and evaluate the persuasion in a smaller domain, rendering the evaluation too distant from a natural dialogue.

5.4 Extended Personality Experiment

The initial experiment showed an effect of the system personality style on the persuasiveness of the whole dialogue and found a significant correlation between the participants preferences for this personality style and their extroversion.

However, the phrasing of the questionnaire and of the initial scenario briefing have given results that might be unclear. In particular, the use of the “outgoing” term as a colloquial synonym for “extrovert” did not seem to correlate with the user’s understanding of this word. In fact, the results show that the preference for an outgoing system does not significantly correlate with the controlled extroversion generation parameter of the system.

In this second exploratory experiment, a similar procedure is used for the dialogue sessions as for the previous personality experiment (see section 5.3):

- The users are discussing with the dialogue system over a choice of three restaurants.
- Each participants faces a random system personality, that is either set to *introvert* or *extrovert* in the Personage generator; all other personality parameters are set to be neutral.
- Before the dialogue session, the restaurant’s attributes preferences are assessed through the same ranking of attributes as in the previous experiment.

However, the participants’ briefing as well as the questionnaire are slightly modified. In particular:

- The users are now told that they will interact with a computer dialogue system and not with a human:

> Thank you for participating in this experiment.
>
> You will have to participate in a scenario and have a chat with a simulated interlocutor through an instant messaging system.

> The interlocutor with whom you will be chatting is controlled by an automated computer dialogue system.
>
> Please be sure to read all the instructions and reply to the little questionnaires given to you along the way.
>
> At the end of the chat session, you will be asked a few questions about your impressions of the interaction with the automated system. The questionnaire also contains questions to evaluate your personality along the big five traits. This questionnaire will be treated anonymously: no personal data will be used when reporting results or shared with a third-party.
>
> Please, try to avoid going back in the steps or reloading the pages.
>
> The purpose of the chat is to agree, with the automated system, on a short list of restaurants where you could go out with your friends tonight. For the research purposes, the objective of the chat is to get to an agreement with the automated system.

- The restaurants' cuisine type is not displayed anymore and the users' cuisine preferences are not assessed before the dialogue (see Section 5.3.3).
- The questionnaire is changed to include a Big Five personality traits inventory for the users to report their perception of the system's personality (see Appendix E).
- The questionnaire now uses a phrasing compatible with this personality traits inventory to evaluate the participants' preference for the extroversion of the system's generation style:

> I found the computer extrovert, enthusiastic tone irritating.

5.4.1 Results

This second experiment was designed to confirm the initial results that there is a significant influence of the user's personality on the style of generation that should be used to achieve better persuasion. The experiment is also designed to clarify the "outgoing" style factor that was observed in the initial experiment and discover if it is directly linked to the system's generation parameter.

115 participants logged in the web based interface for this experiment, however, they did not all complete the dialogue session or fill in the questionnaire after the session. Thus, only 53 dialogue sessions can be analysed for reporting the results of this experiment.

Each participants filled the questionnaire described in the Appendix E and the evaluation of the dialogue persuasiveness is measured through a ranking task as described in section 4.1.2. The participants are of different background but through the setup of the web based experiment are limited to users that can use a computer and that have an Internet access. The participants age ranges from 17 to 60 years old ($M = 33.8$, $SD = 11.4$), distributed between 25 women and 28 men.

5.4.2 Perception of the System's Personality

The system's personality generation parameter is randomly chosen between either an *extrovert* generation personality or *introvert* generation personality. 24 participants faced the *extrovert* generations while 29 faced the *introvert* system. At the end of the dialogue, to test if this generation parameter has an effect on the perception of the system's personality, the users are asked to fill in a Big Five personality traits inventory evaluating their perception of the system. For consistency with the evaluation of the Personage generator discussed by Mairesse & Walker (2007), the same inventory is used here to evaluate the perceived system personality (Gosling et al. 2003).

For this experiment, the extroversion trait of the Big Five inventory should ideally correlate with the generation parameter. The generation parameter is a binary choice between extrovert and introvert corresponding to the extremes of each personality whereas the Big Five inventory provides a numerical evaluation of the perception of extroversion ranging from one – for very introvert – to seven – for very extrovert – with a neutral position at four.

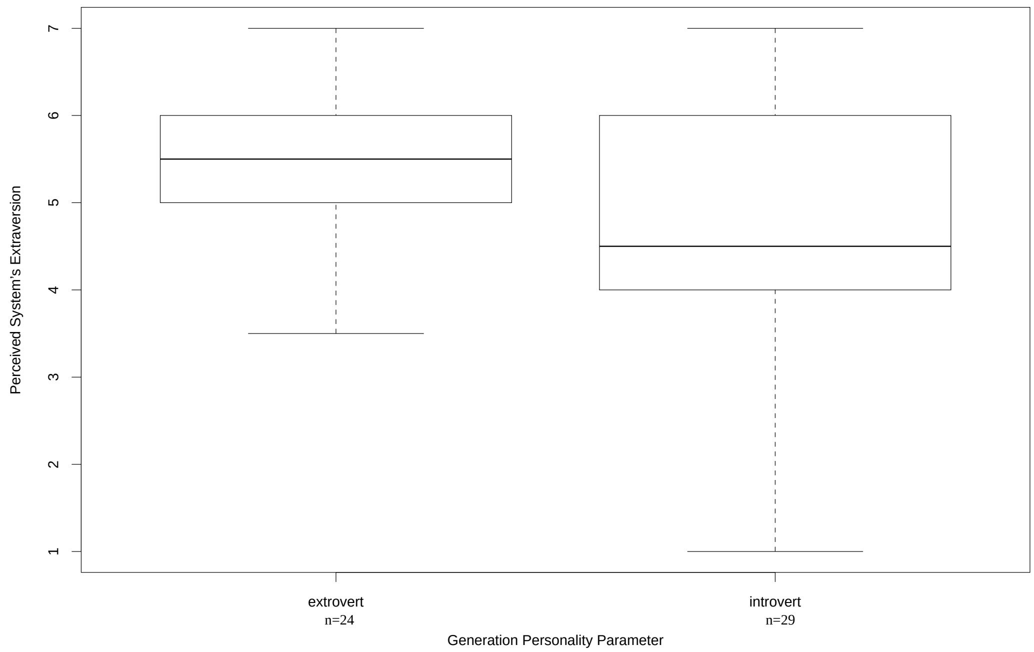


Figure 5.14: The difference in the participants' perception of the system's extraversion does not seem to correlate with the personality trait selected for the generation.

Figure 5.14 shows the perception of the participants of the extroversion of the system. No significant difference can be observed between the generation parameters (Mann-Withney $U = 438.5$, $p = 0.10$). In both cases, the participants rated the system in the *extrovert* side of the scale: $M = 4.77$, $SD = 1.39$ for the *introvert* generations and $M = 5.33$, $SD = 0.90$ for the extrovert generation, only 10.3% of the participants who faced the *introvert* generation rated the system's extroversion under the neutral point, with $M = 2.33$, $SD = 1.26$.

Even if both conditions produce different utterances' styles (see examples in Figures 5.5 and 5.6) that are individually perceived as introvert or extrovert (Mairesse & Walker 2007), the participants perceive the whole dialogue sessions as extrovert. The following sections try to discover the factors that have influenced this perception of the system's personality.

5.4.3 Dialogue Behaviour and Personality Perception

The Personage generator (Mairesse & Walker 2007) was evaluated on individual utterances and showed that users were able to perceive the difference in the system's extroversion personality trait. However, this is not what is observed in our study and there might be multiple factors explaining this discrepancy. In particular, the experiment discussed in this section evaluates the perception of the system's personality *over a whole dialogue* while Mairesse & Walker are testing the generator on *single utterances* outside of a dialogue context.

Thus, the dialogue interaction might affect the perception of the system's personality even if each utterance uses a distinguishable personality style. An adhoc hypothesis might be that a system trying to do persuasion is intrinsically perceived as extrovert, however, there is no data in the current experiment to support this hypothesis.

However, the data collected during the dialogue sessions supports that the behaviour of the dialogue system has an influence on the participants perception of its personality. In particular, the perception of the system's extroversion is influenced by the number of supports used by the system for each arguments (Spearman $\rho = -0.35$, $p = 0.01$). When the system uses more supports as reactions to the user's disagreement, the participant perceive the system as less extrovert (see figure 5.15).

The number of supports and the difference of personality perception cannot be shown to correlate with the number of goal drops in the dialogue. This means that the user accepted the arguments whether there was more support or not and the planner did not have to drop goals to finish the persuasive dialogue[3]. Thus,

[3] See section 3.8 for details on the dialogue management behaviour.

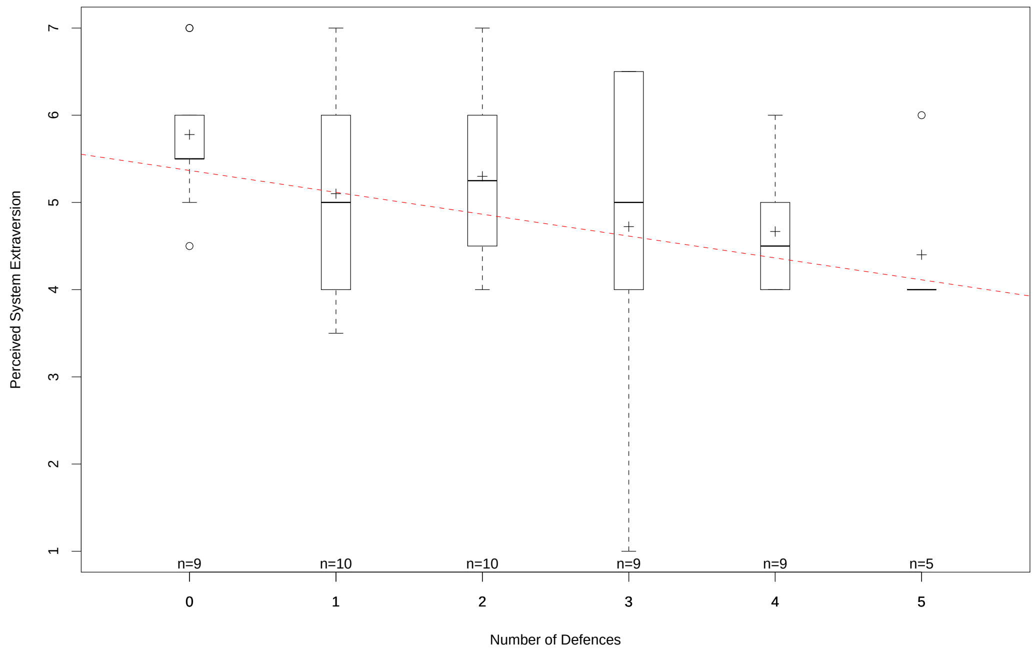


Figure 5.15: The number of supports used by the dialogue manager influences the participant's perception of extraversion (Spearman $\rho = -0.35$, $p = 0.01$).

the participants that are persuaded by the first utterance of the argument and do not require more support to be convinced found the system more extravert, while the participants that needed more support for each argument[4] seem to find the system less extrovert. This difference in perception is identical for both generation parameters, thus the content and presentation of the arguments do not seem to influence this perception and it might come from the personality of the participants.

If we look at the participants that faced the *extrovert* generations, an interesting effect is observed that could explain the fact that some participants rated the system as less extrovert when it uses more supports. With the *extrovert* generation condition, participants that rated themselves low on the *conscientiousness* personality trait needed significantly more supports to be persuaded (Spearman

[4]on average.

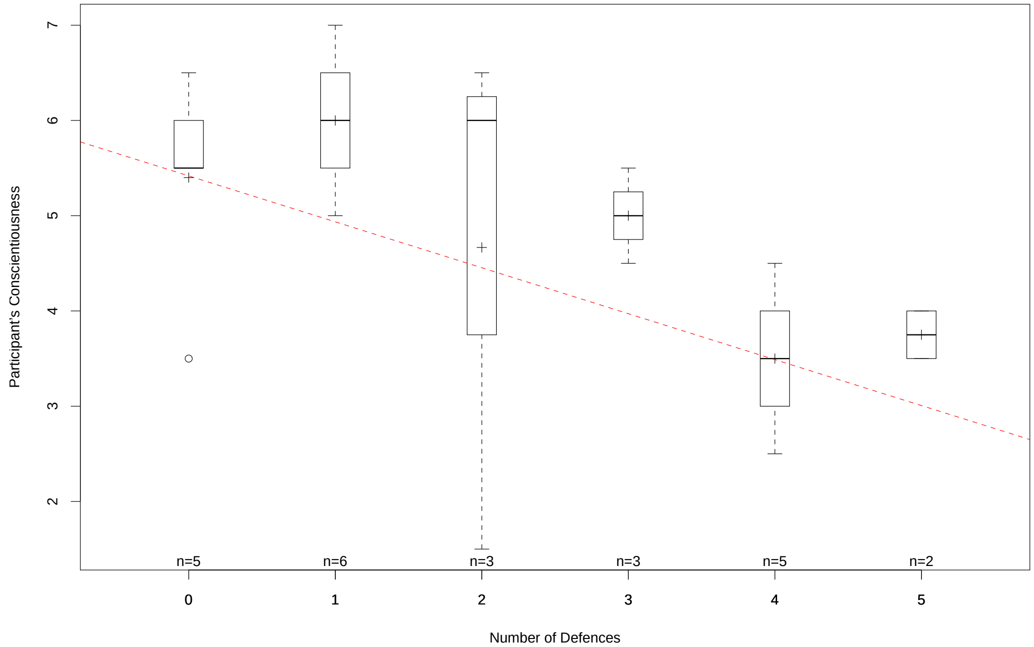


Figure 5.16: Participant's that rated themselves as less conscientious need more support from the system to be persuaded when they face the *extrovert* generation (Spearman $\rho = -0.59$, $p < 0.01$).

$\rho = -0.59$, $p < 0.01$; see figure 5.16).

In the Big Five personality inventory, the conscientiousness trait corresponds to the participant's tendency to show self-discipline and organisation. Participants rating high on the conscientiousness scale are labeled as *dependable* and *self-disciplined* while those that rate low are labeled as *disorganised* and *careless* (Gosling et al. 2003; see Appendix E).

The link between the need for defences and the user's conscientiousness might be explained by the fact that participants with low self-discipline might need more reassurance. In addition, the fact that the system uses more supports makes it appear generally less extrovert as it might seem that it needs to assert its position more.

Other traits of the system's perceived personality are also influenced by what happens during the dialogue. In particular, while the generator's parameter for

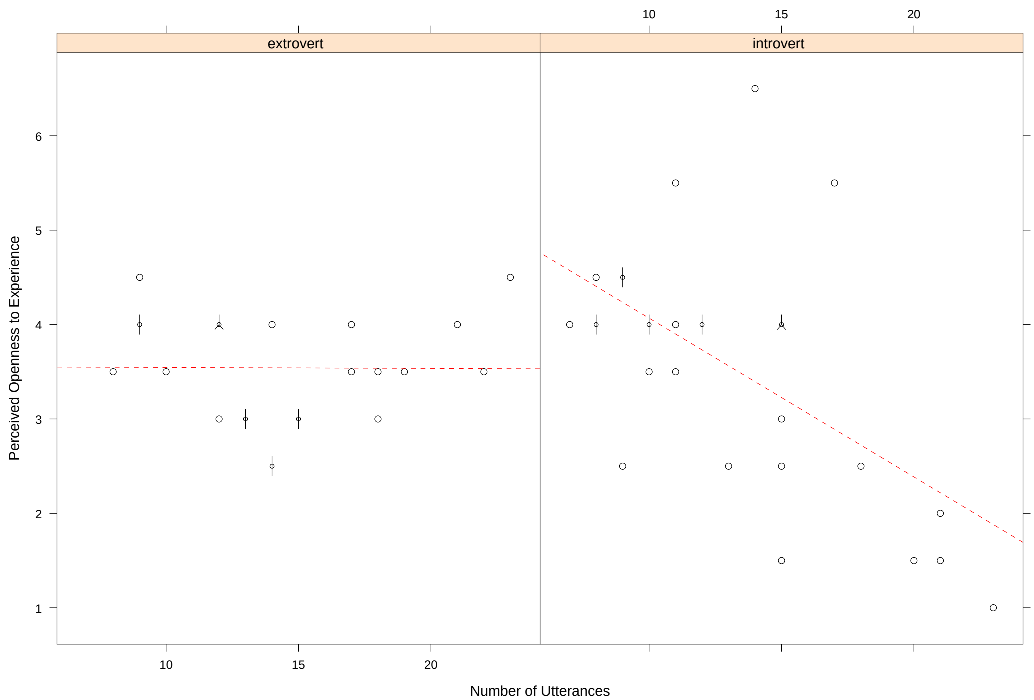


Figure 5.17: The length of the dialogues influences the participants' perception of its openness to experience (Spearman $\rho = -0.42$, $p < 0.01$). Looking at the underlying groups, we can see that this correlation is only due to the *introvert* condition (Spearman $\rho = 0.52$, $p < 0.01$).

the system's *openness to experience* was always the same for all dialogue sessions, this personality trait seems to be perceived differently by the participants. If the participants face the *extrovert* system, there is no significant difference in how they perceive this personality trait and it is close to the neutral point that was set as parameter ($M = 3.53$, $SD = 1.03$). However, for the participants that face the *introvert* system, the *openness to experience* personality trait is significantly influenced by the length of the dialogue. If the dialogue session has a lot of utterances, the system's *openness to experience* trait is perceived significantly lower than for shorter dialogues (Spearman $\rho = -0.42$, $p < 0.01$; see figure 5.17).

This difference in perception of the *openness to experience* between the two generation conditions might be explained by the perception of repetition. Partic-

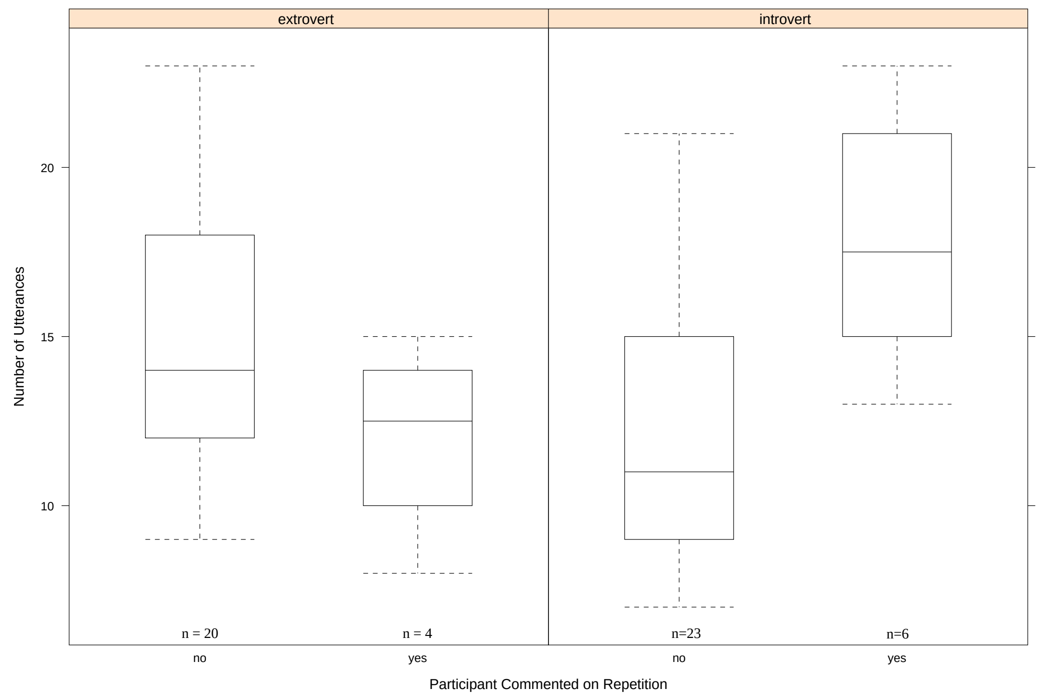


Figure 5.18: When faced with the introvert generations, the participants perceived more repetition in the dialogues if they were longer (Mann-Withney $U = 19.5$, $p < 0.01$).

ipants facing the *introvert* system commented significantly more on the repetitiveness of the dialogue when they were facing longer dialogues (Mann-Withney $U = 19.5$, $p < 0.01$) while this effect cannot be observed with the *extrovert* system (see figure 5.18). This increased impression of repetitiveness might be due to the amount of persuasive moves available to the dialogue manager (see section 5.3.3) and the variability in their generation. The *extrovert* generations tend to generate more verbose sentences (see Mairesse & Walker 2007) which allows more variability in the lexical and syntactic realisation for the same content, however, *introvert* generations are shorter and thus have less possibility for variability in their realisation, leading to more repetition if the same persuasive move has to be used more than once (see section 5.4.4 and figure 5.19 in particular).

When the dialogue becomes longer, there is more chance that the system will have to reuse a similar persuasive move and that, for *extrovert* generations, the generator will create the same realisations. The user then feels like the system is less creative and when filling the personality inventory (see Appendix E) will rate it as *conventional* and *uncreative*, thus affecting the measured *openness to experience*.

5.4.4 Repetitions

In this experiment, no question was formulated in the questionnaire (see Appendix E) to evaluate the user's perception of repetition, however, ten participants used the free comment box to express their frustration over the repetitiveness of the system during the dialogue. This observation has to be considered with caution as it does not mean that these participants are the only one that perceived repetition, and other participants might have perceived it but not wanted to comment. However, it does provide some interesting insight on the effect of repetition on the perception of the system's personality and general behaviour.

No significant difference in the participants' personality can be observed between the ones that commented on the repetitiveness of the dialogue and the ones that did not. Their might be another trait of these participants that cannot be observed in this experiment that influences the fact that they commented or not. However, a measure of the amount of repetition in the dialogue shows that the participants that commented actually faced significantly more repetitive dialogues.

To evaluate the amount of repetitions in the dialogues, two measures have been used. The first measure uses the word based Levenshtein *edit distance* (Levenshtein 1966) to measure the average distance between system utterances in each dialogue session. The interest of this distance is that it evaluates the similarity between utterances' structure. If the utterances have similar content and use a similar syntactic structure, the distance is small. However, this measure is dependent on the length of the utterances. From this measure, a significant difference can already be observed between the participants that commented and

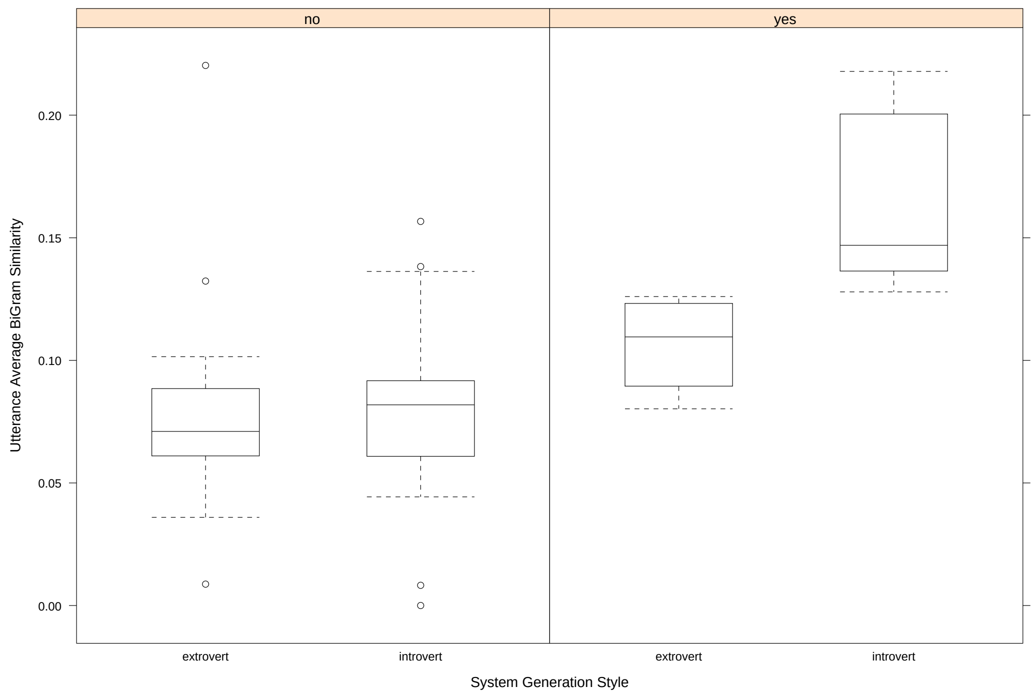


Figure 5.19: Participants that commented on the repetition faced significantly more repetitive dialogue (Mann-Withney $U = 53$, $p < 0.01$). In addition, the utterances generated by the *introvert* generation parameter, were significantly more similar than the one generated by the *extrovert* system in the dialogues where participants commented on the repetitiveness (Mann-Withney $U = 0$, $p < 0.01$).

those that did not (Mann-Withney $U = 103.5$, $p = 0.01$).

The second measure evaluates the average in similarity of the utterances' content for each dialogue sessions. This measure uses bi-grams to measure similarity between utterances: if two utterances have a similar content, their bi-gram signatures will be close and the similarity will be high. A stronger effect ($\hat{d} = 1.33$) can be observed with this measure and a significant difference is measured between the users that commented and those that did not (Mann-Withney $U = 53$, $p < 0.01$; see figure 5.19). As discussed earlier, this effect is stronger for the *introvert* generation where the similarity between utterances is significantly higher than the *extrovert* generation when the participants commented on the repetitiveness (Mann-Withney $U = 0$, $p < 0.01$).

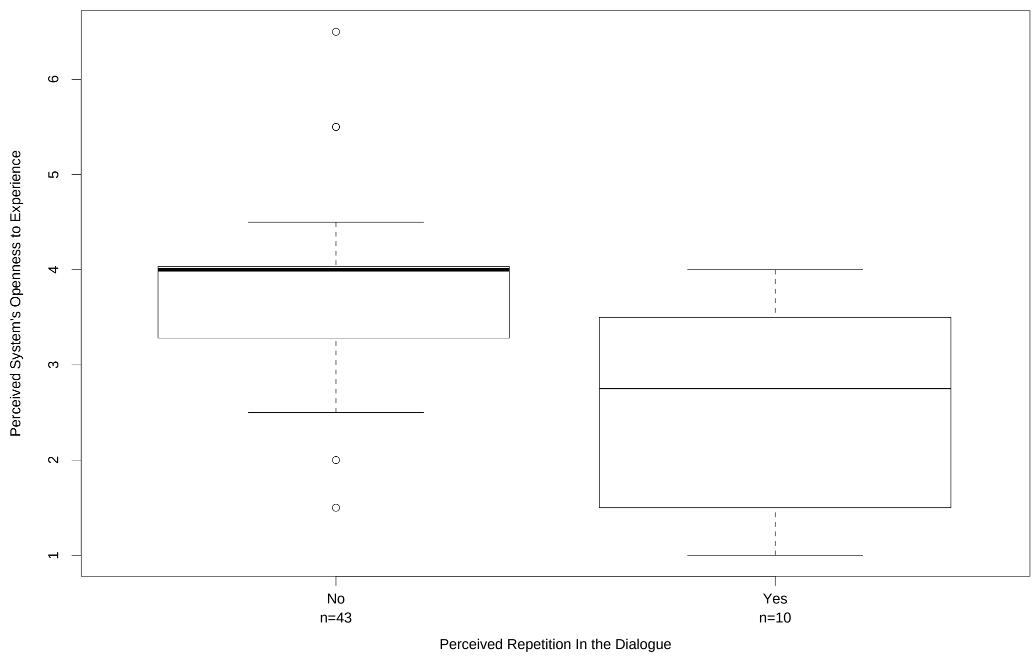


Figure 5.20: The impression of repetition changes the partitipants perception of the system's openness to experience (Mann-Withney $U = 336.5$, $p < 0.01$).

The participants that perceived and commented on the repetitiveness of the dialogue also reported significantly different perceived personality traits for the system. In a similar way to the number of defences or the length of the session,

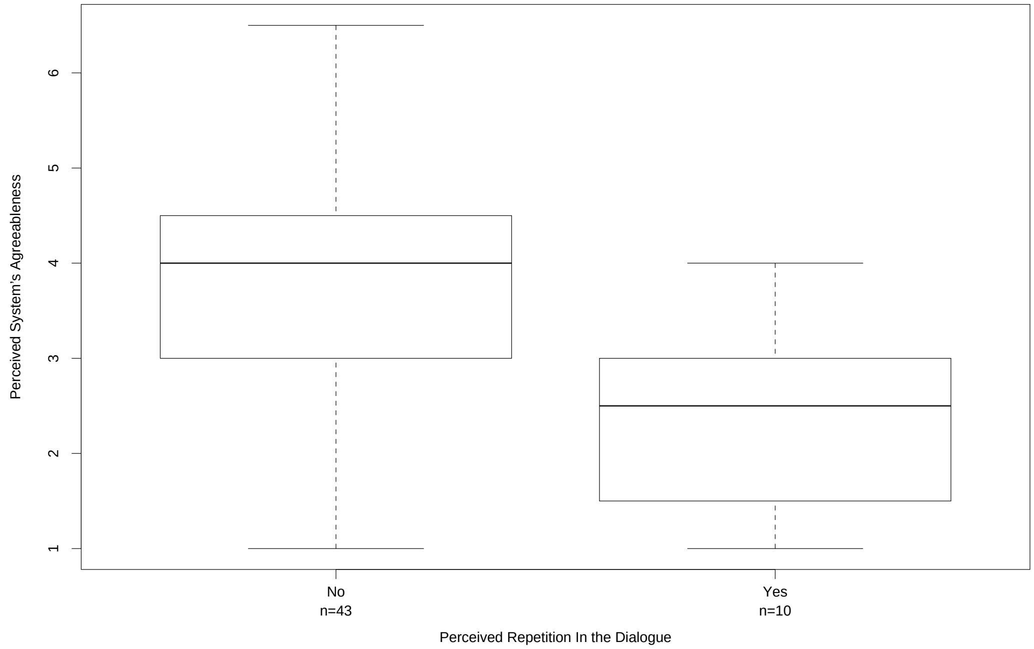


Figure 5.21: The participants perception of repetition also influences their perception of the system's agreeableness (Mann-Withney $U = 337.5$, $p < 0.01$).

the perception of repetition in the dialogue is a trait of the system's behaviour that influences the perception of its personality. In particular, even if all dialogues have the same generation parameters for these traits, the participants that commented on the system's repetitiveness rated its *openness to experience* and *agreeableness* significantly lower than the users that did not comment – Mann-Withney $U = 336.5$, $p < 0.01$, $\hat{d} = 0.31$ and Mann-Withney $U = 337.5$, $p < 0.01$, $\hat{d} = 0.36$ respectively (see figures 5.20 and 5.21).

5.4.5 Perceived Personality and Dialogue Perception

Effect of Repetition

The impression of repetition also influences the final perception of the dialogue's output. In the previous experiment, the impact of different aspects of the sys-

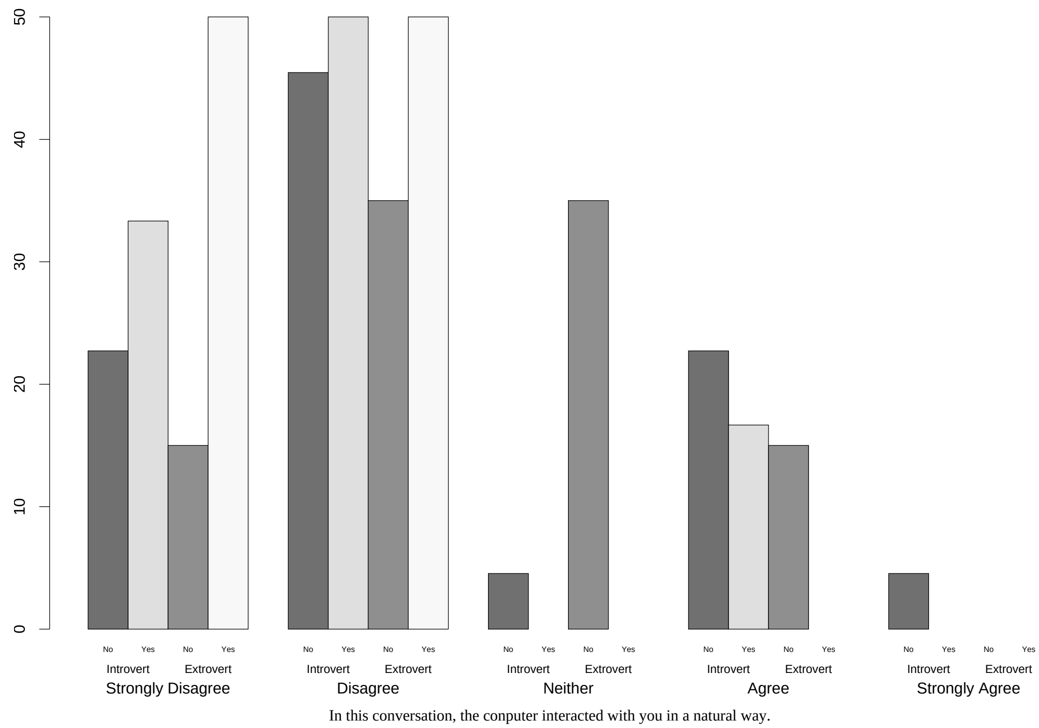


Figure 5.22: Repetition and Perception of Naturalness

tem's behaviour on the participants' perception of its trust, naturalness and persuasiveness has already been observed. In this experiment, similar results can be measured, however, different factors can be distinguished as affecting the user's perception of the dialogue system effectiveness to create natural dialogues.

Indeed, The participants that commented on the repetitiveness of the dialogue answered the statement "*In this conversation, the computer interacted with you in a natural way*" in a significantly more negative way (see figure 5.22), in particular, when they were facing the *extrovert* system, 50% disagreed with the system and 50% strongly disagreed (Mann-Withney $U = 64$, $p = 0.05$). A similar influence of the repetitiveness can be observed on the trust attributed to the system. When the participants commented on the repetitiveness of the system, they also answered significantly more negatively to the statement "*The computer was trustworthy*" (see figure 5.23). In particular, when faced with the *extrovert* system,

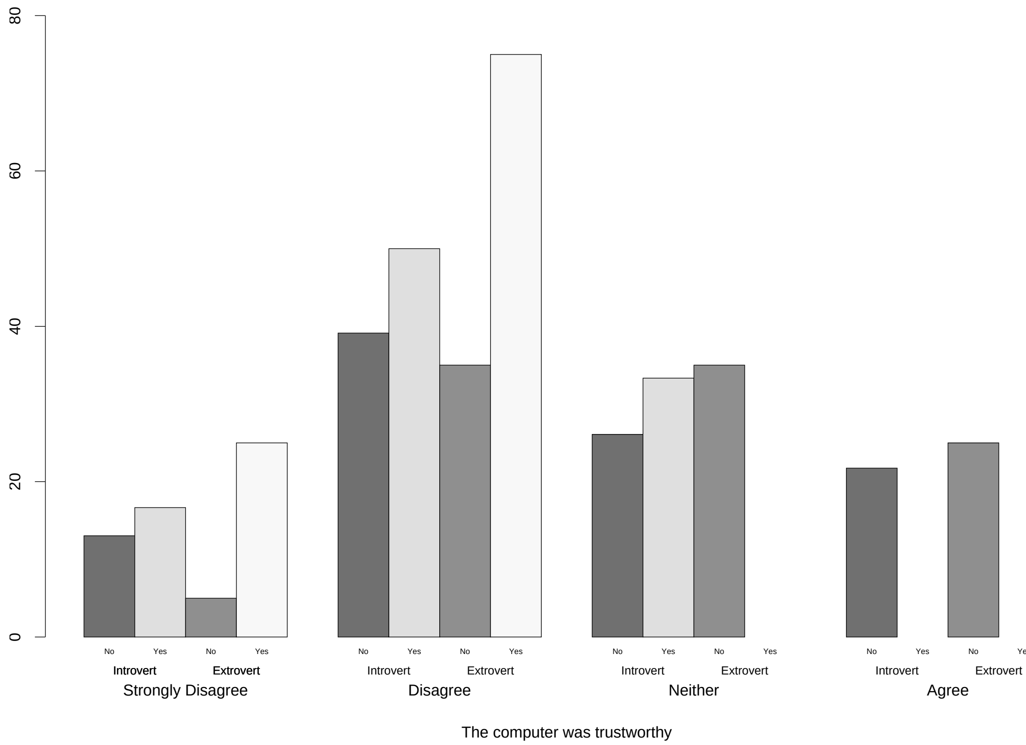


Figure 5.23: Repetition and Perception of Trust

25% answered strongly disagree and 75% disagreed (Mann-Withney $U = 66$, $p < 0.05$).

These effects of repetition on the perception of trust and naturalness seem to be stronger with the *extrovert* generation parameter. There is not enough observations to find why this is; the number of participants in each generation group that commented on the repetition is similar and there is no observable difference in their personality or any other measured aspect of the dialogue.

Effect of the Perceived Personality

The trust attributed to the speaker has been identified as an important factor of the interaction in persuasive communication (see section 2.4.1), Fogg (2003) showed that in human-computer interaction, this trust was linked to aspects of the computer system such as the quality of the graphical interface.

With an interactive persuasive dialogue, the graphical interface is minimum as the interaction with the computer goes through dialogue. The perception of the system through this dialogue is thus important for its trustworthiness. In this experiment, the perceived system's personality seems to have an influence on the trust given by the participants to the system.

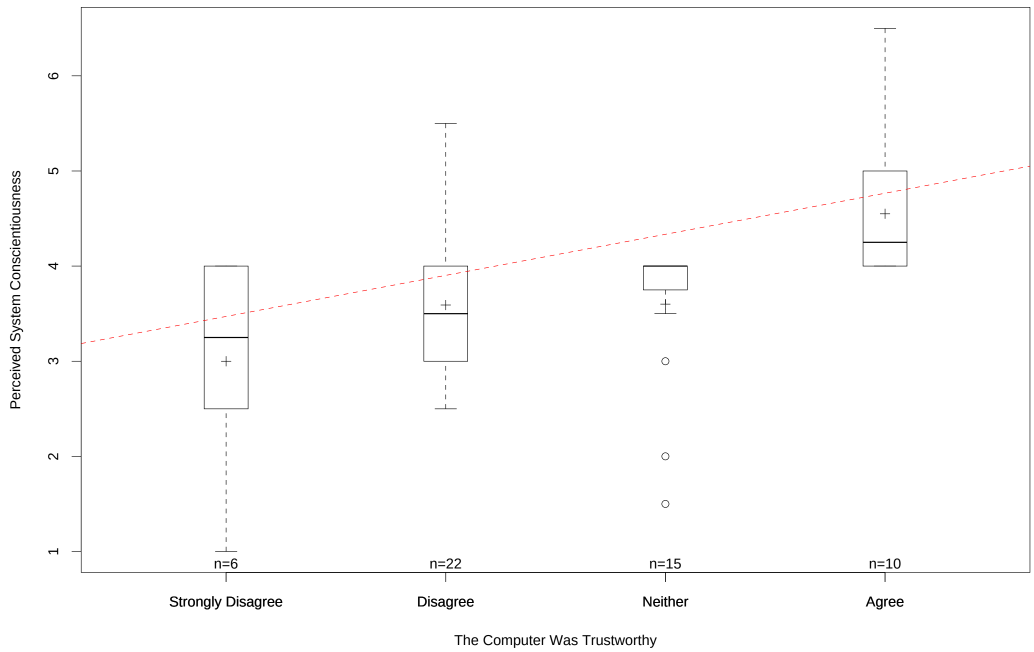


Figure 5.24: Participants that perceived the system as having a higher than neutral conscientiousness found it significantly more trustworthy (Spearman $\rho = 0.46$, $p < 0.01$).

The perceived system *conscientiousness* has a significant influence on the trust given to the system by the participants (Spearman $\rho = 0.46$, $p < 0.01$; see figure 5.24). According to the Big Five personality traits theory (Pervin & John 2001), people rating high on the conscientiousness trait can be defined as *self-discipline* and *competent* while people with low conscientiousness are considered *disorganised* and *careless*.

When the participants perceive the system as having a low conscientious-

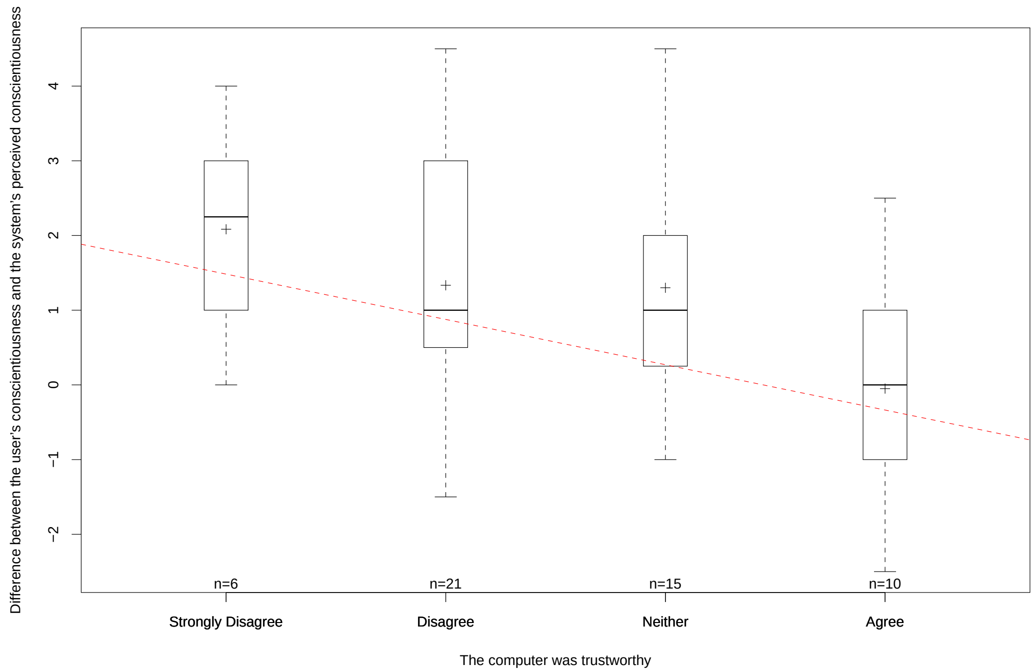


Figure 5.25: The participants that rated their own conscientiousness as lower than the system's one appear to trust more the computer.

ness, they find it disorganised and careless, and thus find it less trustworthy. This observation is supported by a second effect observed during this experiment: Participants that find the system more conscientious than they think they are significantly give it more trust (Spearman $\rho = -0.312$, $p = 0.02$). Thus, participants that have a higher esteem of the system's competence than of their own competence find the system more trustworthy (see figure 5.25).

In a similar manner, the perceived *agreeableness* trait has an influence on the system's trust. The participants that found the system more agreeable significantly found it more trustworthy (Spearman $\rho = 0.44$, $p < 0.01$). The agreeableness trait corresponds to the sympathy of the system, in the inventory used in this experiment, high agreeableness corresponds to *sympathetic* and *warm* persons while the low end of the scale corresponds to *critical* and *quarrelsome* persons. A majority of the participants that rated the system's agreeableness under the neu-

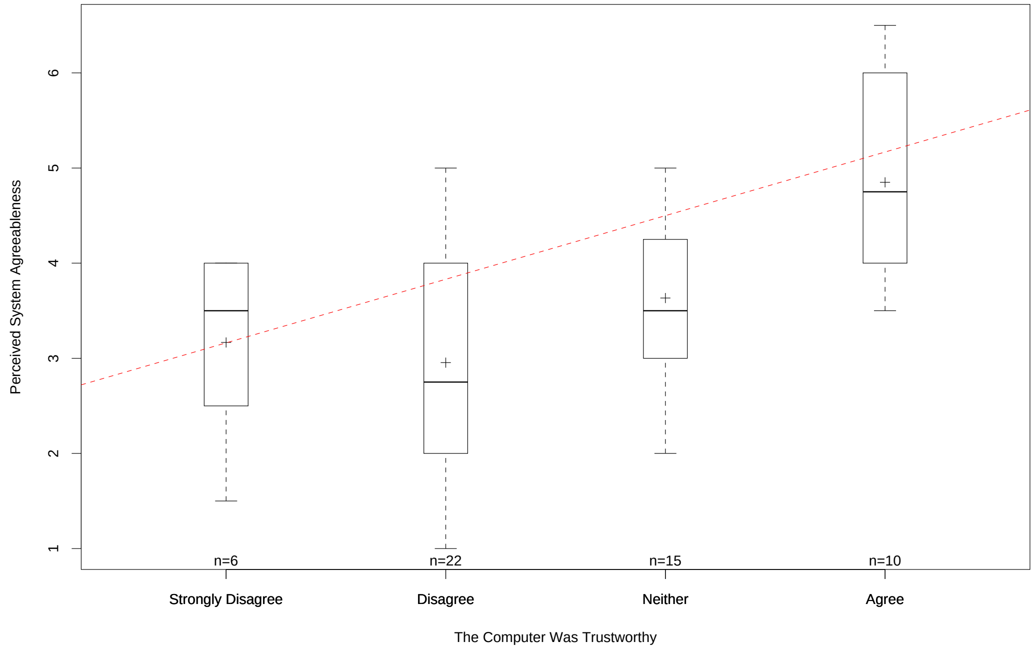


Figure 5.26: The perceived agreeableness of the system influences the trustworthiness given to the system by the participants.

tral point of the scale answered to the statement "The computer was trustworthy" with either *Disagree* (51.2%) or *Strongly Disagree* (11.1%) (see Figure 5.26). In this experiment, there are no observable factors that can explain the difference in perception of these traits, however, it is important to have a control over this perception to achieve better persuasiveness and additional experimentation might be needed to find the related factors.

The perceived extroversion also seems to have an influence on the judgement of the dialogue naturalness. For the participants that faced the *extrovert* generation, the dialogue sounded significantly less natural when the system was perceived as very extrovert (Spearman $\rho = -0.47$, $p = 0.03$; see figure 5.26). Participants commented that a computer that seemed too extrovert did not seem natural to them.

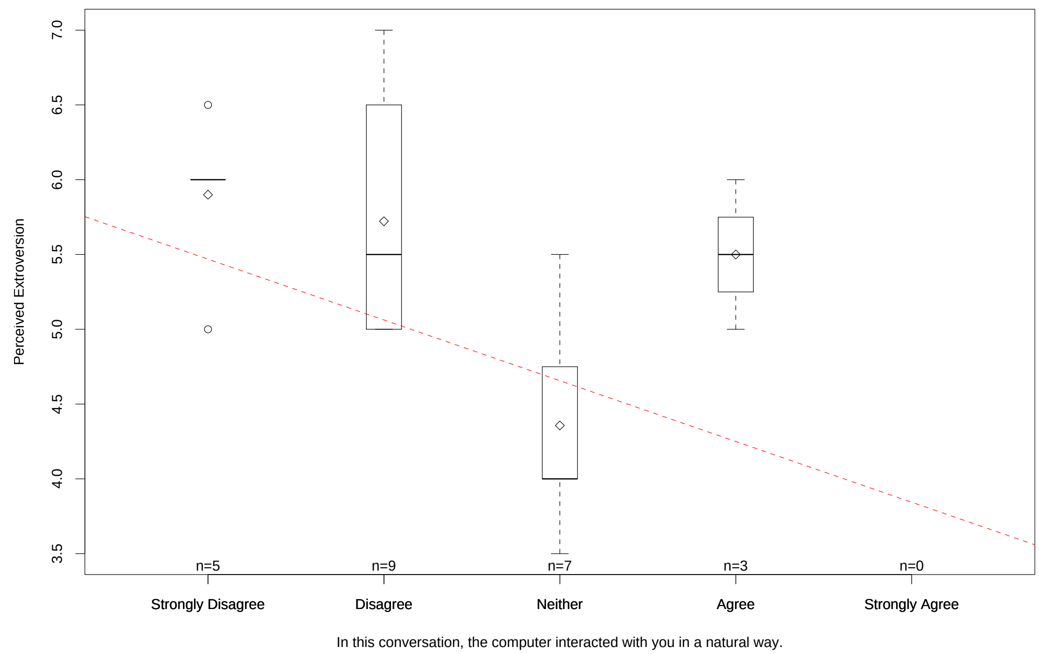


Figure 5.27: Participants that faced the *extrovert* system found the system less natural if they perceived it as more extrovert.

5.4.6 Dialogue Perception and Persuasiveness

In this experiment, the participants are achieving the ranking task defined in section 5.1 and the persuasiveness metric is used to measure the power of persuasion of the system. As with the previous experiment, the answers to the statement "The computer was persuasive" significantly correlate with the persuasiveness measure (Spearman $\rho = 0.38$, $p < 0.01$). However, in this experiment, the persuasiveness measure does not correlate significantly with other factors observed in the dialogue sessions. In particular, there is no correlation with the answers to the statement "I found the computer's extrovert, enthusiastic tone irritating" in this experiment and the *outgoing* factor observed in the previous experiment does not seem to correspond to the perception of extroversion of the system. In addition, no significant correlation can be observed between the perceived system's personality traits and its persuasiveness.

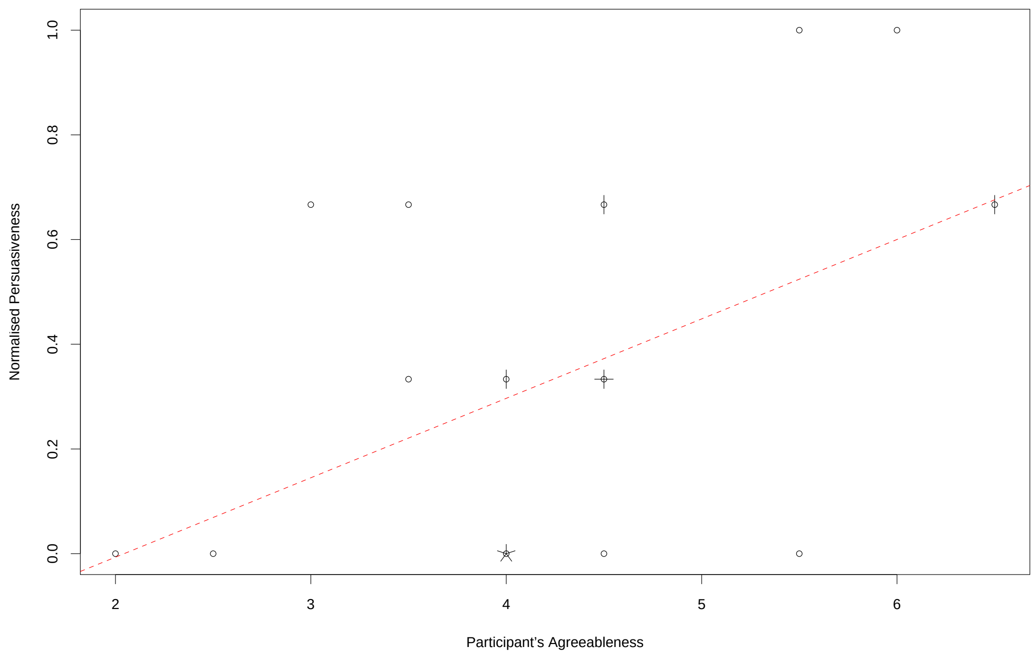


Figure 5.28: Participants that faced the *extrovert* system were more responsive to the persuasion if they rate high on the agreeableness personality trait. (stars represent overlapping observations)

The only factor that can be observed to significantly influence the system's persuasiveness is the participants' *agreeableness* personality trait. Participants that faced the *extrovert* generation and that rate high on the agreeableness scale are significantly more persuaded than if they rank low on the scale (Spearman $\rho = -0.45$, $p = 0.02$). The Big Five personality inventory gives high agreeableness rating to the participants that are less *critical* and *quarrelsome*, in addition, Pervin & John (2001) defines the agreeable personality with the adjectives *trust* and *compliance (not stubborn)* that might explain that more agreeable users are more ready to be persuaded by the computer as they are intrinsically more trustworthy and compliant (see figure 5.28).

While no correlation can be found between the impression of repetitiveness and the measured persuasiveness of the system, the participants that actively commented on the system's repetitiveness also answered significantly more neg-

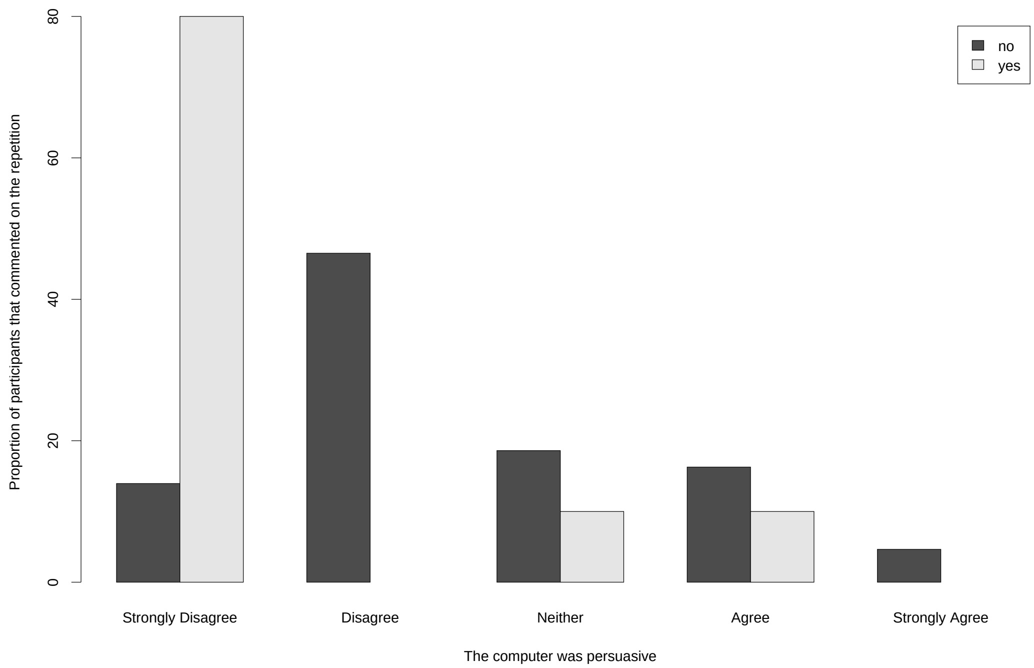


Figure 5.29: Repetition and Perception of Persuasion

atively to the statement "The computer was persuasive" and 80% of them answered "Strongly Disagree" (Mann-Withney $U = 338.5$, $p < 0.01$; see figure 5.29). This might be directly due to the actual repetitiveness of the system that, as shown before, influences the perception of the system and thus makes it seem less persuasive. However, there is no observable correlations between the repetitiveness measures and the answers to the question to this statement, and as we cannot be sure of what made these participants comment on the repetitiveness, there might be a hidden reason why they rated the system as less persuasive.

In the questionnaire, the participants were also asked to answer to the statements "I found the fact that the system acted as if it knew me was irritating" and "I found the computer's extrovert, enthusiastic tone irritating". While the answers to these statement do not seem to correlate with the persuasiveness of the system, they do seem to strongly correlate together (Spearman $\rho = -0.45$, $p = 0.02$). The participants that were irritated by the extroversion of the system were

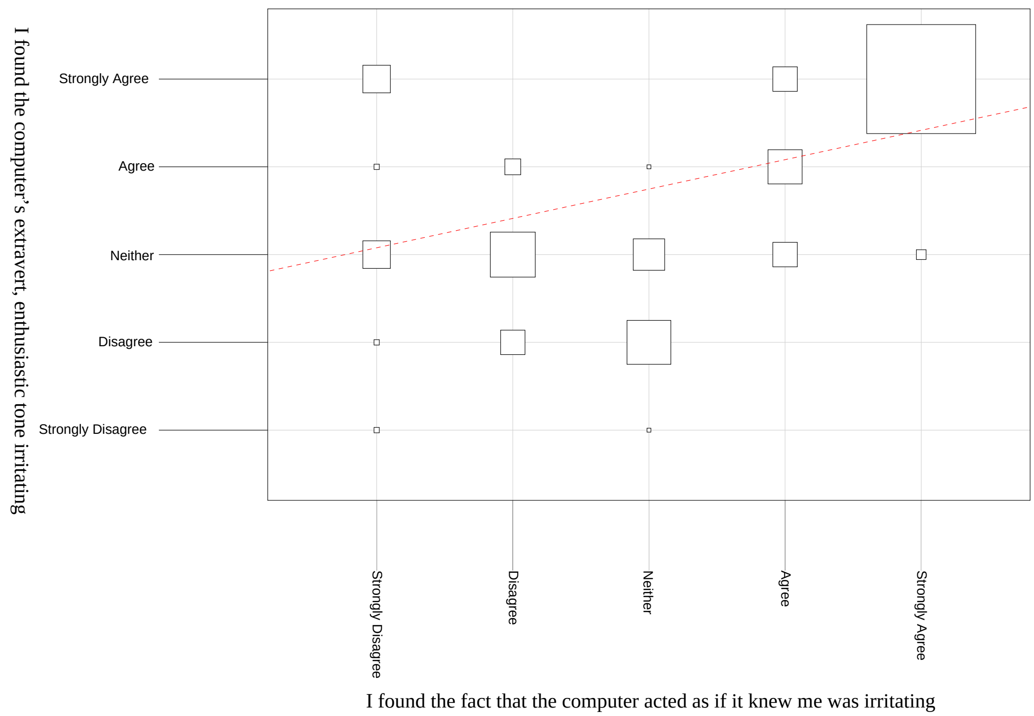


Figure 5.30: Participants consistently found the system to be irritating when acting in an extrovert manner and being too familiar.

also irritated by its familiarity and 80% of the participants that strongly agreed with the second statement also strongly agreed to the first (see figure 5.30).

Chapter 6

Conclusions

6.1 Contribution of the Thesis

Persuasive Communication is a wide field of research in sociology and philosophy and many theories have been formulated to improve the persuasion. Sociologists concentrate on the identification of features of communication that can influence persuasiveness, studying the receiver's motivation, intentions and values and the correlation with the output of persuasion. Philosophers have, long before, studied rhetoric to formalise the techniques of selecting and presenting data in a persuasive way.

After having studied these researches and their current applications in the field of computer science, the aim of this research is to contribute novel techniques to improve human-computer interaction and improve its persuasiveness. In particular the research presented concentrates on the development of persuasion through human-computer dialogue.

The current axis of research on persuasion in human-computer interaction is to improve the user's perception of the computer to build-up trust and relationship. The exploration of the available dialogue management techniques shows inherent drawbacks in their design for persuasiveness, motivating the development of a novel dialogue management framework.

The EDEN Framework was developed around the hypothesis that the management of persuasive goals and of social cues have to be separated to ease the domain authoring and to offer more empathy to the user. Planning an argumentation dialogue a priori can only lead to plans containing support argumentation, without taking into account unpredictable counter-arguments from the user. In addition, including social cues in this planning framework would render its authoring complex and the reasoning computationally heavy.

The first contribution of this thesis is to provide a layered framework, able to plan and guarantee persuasive goals achievement while staying reactive to the user counter-argumentation and being able to include social cues and dialogue backchannel without prior planning.

The second contribution of the thesis is an improved persuasiveness metric that is used to evaluate the dialogue performance in controlled experiments where external influences on behaviour are limited. In these experiments, the EDEN Framework is shown to achieve a better persuasiveness output than standard planned approach to dialogue management. The research reported also shows that such increase in persuasiveness is not done at the expense of the user's perception of the dialogue quality. The EDEN Framework achieves a stronger persuasion while staying as trustworthy as the planned approach and no significant difference is measured in the user's rating of the interaction.

The reactive approach to argumentation and social cues management is also used to explore the impact of the system's displayed personality on the persuasiveness output as well as on the users' perception of the interaction. The experiment design of this thesis shows that the system's extroversion has an impact on the trust given to the dialogue system and on the general perception of the interaction by the user. The experimentation also shows that the adoption of familiatity in the generation by the system has an impact on its persuasiveness. The users are split between a group appreciating this familiarity and a group that is alienated by such form of discourse from a computer.

The research performed in this thesis shows that there is a need for a novel approach to dialogue management to develop dialogues whose performance is not only tied to the achievement of a particular task but also to the dialogue's reactiveness to the user.

6.2 Evaluation Results

Two experiments are presented to evaluate the impact of different features of the interaction on the persuasiveness of a computer dialogue system. These features are used as the drawing line of the new EDEN Framework and appear to lead to an efficient design for the dialogue framework.

The Desert Survival Scenario study shows that the EDEN Framework is able to be more persuasive than a purely planned dialogue management. This improved persuasiveness can be explained by two features of the new dialogue management:

- The EDEN Framework is designed to create smoother dialogues including more argumentation without impairing the trust or the perception of the iteration. The design of the framework also avoids an increase of the perception of coercion.

- The EDEN Framework is more suited to author persuasive domains and can generate persuasive dialogues without adding complexity to the planning component or losing consistency in the dialogue.

The Restaurant Domain experiment is designed to study the impact on the persuasion of a system displaying personality. Even if the experiment did not show a direct link between the system's extroversion and its persuasiveness, the experiment is able to confirm that the personalisation of the system has a direct impact on its perception by the user. De Boni et al. (2008) shows similar results, confirming that features of the dialogue such as humour and relationship maintenance have an important impact on the perception of the dialogue system.

De Boni et al. conclude by asking whether this perception will have an impact on the system's performance. The Restaurant Domain study gives a hint as to what the answer can be for persuasiveness; the observations collected during this experiment illustrate the difference of system's performance when the user is happy to face a familiar dialogue management system and when this irritates the participants. The experiment also shows a possible correlation between this preference and the user's personality and age.

The extended personality experiment with the Restaurant Domain allowed to discover more factors influencing the perception of the dialogue. In particular, it appears that the utterance level personality style selection provided by the Personage generator (Mairesse & Walker 2007) is not enough to simulate an extrovert or introvert dialogue. In fact, factors coming from the interaction behaviour of the dialogue, such as the number of defences or its length influence the users perception of the system's personality. These results correspond to human-human interaction rules and goes towards the *Media Equation* theory (Reeves & Nass 1996) that states that humans interact with computers following rules close to the social rules of interaction between humans.

The second personality experiment also showed the importance of the repetition on the perception of the dialogue and ultimately its persuasiveness. It is thus important to develop dialogue systems that have large enough knowledge base for the interaction management and the generation of diverse utterances.

6.3 Future Research

The EDEN Framework shows a novel approach to dialogue management that can be extended and improved by the addition of more complex planning and generation techniques. During the development of the framework, the research concentrated on the design of a novel approach and its evaluation by using techniques of the state-of-the-art dialogue management field and no resources were available to evaluate novel and specialised techniques for dialogue planning or content generation.

In the field of persuasion planning, research is now focusing on planning the ethics of persuasion (Guerini & Stock 2005) and complex argumentation strategies (Mazzotta et al. 2007). The EDEN Framework modular design could experiment on novel domains and include new strategies through more complex planning techniques.

The framework developed for this thesis uses a canned text approach to content generation that is extended by the Personage (Mairesse & Walker 2007) framework to generate personalities. As discussed in section 5.3.3, this extended generation has limits and the use of a specialised content planner for persuasion (for instance Reed 1998) might improve the variation in the structure of the generated arguments and avoid the repetition that lowers the persuasion of the overall dialogue.

The reactive component of the Eden Framework relies on the detection of the type of utterances entered by the user. The system needs to know if the user is agreeing or disagreeing and currently relies on a discreet classifier. The extension of this classifier to provide a more detailed evaluation of the user's agreement on a continuous scale will help the system to select more appropriate reaction to the user's counter-arguments, adapting the length and type of the argumentation to the level of disagreement of the user. As discussed by Gilbert et al. (2003), perfect argumentation scheme detection would allow tailoring even more the system's reaction. However, neither of these extended approaches appears to be currently possible due to the lack of annotated corpus available for training such complex argumentation system. The work of Grasso (2003) might lead to such corpus in the domain of health communication.

6.4 Conclusion

The research performed for this thesis proved that a more flexible approach to dialogue management is needed to create more natural dialogues able to discuss complex issues that cannot be tackled easily with the current dialogue management systems.

The EDEN Framework does not use complex algorithms or techniques to model a more reactive dialogue, but it can still achieve a better persuasiveness without impairing the comfort of the user in the dialogue.

Improving the dialogue reactiveness and the user's perception of the interaction is not only a requirement of argumentative dialogue as other applications such as tutoring require more involvement from the user and could benefit from the layered approach proposed by this thesis. The EDEN Framework could thus be extended and applied to different application domains that require a more natural dialogue to improve the dialogue outcome.

Appendix A

Finding Goals in chatbots

A.0.1 Persuasion and chatbots

In trying to change the users' behaviour and attitude, techniques of persuasion – including argumentation and rhetoric – have to be used, but there is no existing framework integrating them in computer based dialogue.

Indeed, persuasive communication textbooks (Stiff & Mongeau 2002) suggest that it is often necessary to create social bonds – trust and credibility for example –, by using social cues, in order to be effective in the persuasion. In some ways, chatbots are designed to display empathy and entertain the user. These management systems are not tailored to any particular task and can chat freely with the user. This *freedom of speech* makes them appear to be using some social cues.

From this perspective, developing natural argumentation on the basis of a chatbot system seems feasible. Argumentative dialogue needs to be more open than traditional task-oriented dialogue. It needs to leave some freedom to the user and needs to use social cues to effectively deliver the argument (see chapter 2).

Freedom of speech is also the weakness of chatbots that do not focus on any particular subject and change topic each time they cannot understand the user because of a lack of dialogue history and context. Therefore, the users cannot achieve a well specified task and can be frustrated or disappointed by this lack of *discourse consistency*. This is also a barrier to achieving continuity in the discourse, as chatbot systems have short memories[1] and cannot adapt the discourse in regard to the topics already discussed.

In fact, argumentative dialogue is not just *chitchat* with the user. Argumentation always tries to eventually achieve a goal. In a persuasive argumentation, the goal is to persuade the users of the speaker's point of view, to change their beliefs – and eventually their behaviours.

Indeed, the desired system needs to foresee the dialogue moves to accomplish in order to reach a point in the dialogue where the user is convinced. In the following section, results that show how chatbot systems are not suited for building a natural argumentation system with such characteristics are discussed.

As a first step to the thesis, the study of a state-of-the-art chatbot system[2] based on pattern matching has been performed. The aim of this study was to find underlying structures to the pattern matching system to be used to add goal management for the *apparently unrestricted* pattern matching tree.

The AliceBot chatbot (Wallace 2004) is the winner of the Loebner prize[3] for the years 2000, 2001 and 2004. This prize evaluates dialogue systems with the Turing test by comparing them to open domain human-human interaction. This success is rendered possible by the simple design of the system and the large community extending the knowledge base.

This community knowledge of "good" dialogue interaction is encoded in a sparse structure of reaction rules with no direct information for task achievement or goal planning. This renders the use of the AliceBot system for a goal oriented dialogue – such as persuasive dialogue – difficult as it is not possible to plan the achievement of goals and guarantee consistency. However, the knowledge encoded in the reaction database of the chatbot is very interesting as it contains

[1] usually, one or two utterances.
[2] AliceBot: `http://www.alicebot.org`
[3] `http://www.loebner.net/Prizef/loebner-prize.html`

information about social cue use and chit chat that both make the dialogue appear more natural.

A structure analysis is discussed in this chapter to try to find implicit finite-state machines within an existing chatbot engine. This one contains a large amount of knowledge about "proper" interaction and it would be interesting to discover if the community that developed the sparse knowledge structure implicitly integrated reusable task oriented knowledge in addition to the discourse level reactions.

A.0.2 Pattern Matching System

The AliceBot system is not based on any state-transition machine or planning system, but on a pattern matching system similar to the technique introduced in Eliza (Weizenbaum 1966). The chatbot answers to the user by searching in its database for a pattern that matches the user's utterance, this pattern is linked to a template that the chatbot fills in from the user's input to produce an answer.

The database is composed of *categories* representing possible reactions to the user, which are defined by:

a *pattern* that matches user utterances. A pattern can use two types of wildcards: "*" matches any words if there does not exist another category with a more specific pattern – i.e. with a specific word in the same place as this wildcard – and "_" matches any words even if a more specific category exists.

a *template* that is used to generate an answer. The template can reference parts of the user input matched by wildcards and apply standard transformations – e.g. replacing pronouns to transform "you" in "me" (see Figure A.3).

a *topic* limits the use of a category to a particular context. The current topic of the dialogue can be changed by a category.

a *"that"* is also a pattern that maintains short-term memory by matching the last utterance from the chatbot to select a new category.

```
<category>
    <pattern>A *</pattern>
    <that>IT REFERS TO SOME BAD *</that>
    <template>
       It refers to
    </template>
</category>

<category>
    <pattern>IT IS MY FAVORITE *</pattern>
    <template>
       <srai>
            MY FAVORITE <star/> IS <get name="it"/>
       </srai>
    </template>
</category>


<category>
    <pattern>MY FAVORITE CAR IS *</pattern>
    <template>
       Oh yeah, I like <star/> too.
    </template>
</category>
```

Figure A.1: Example of AIML Categories.

For example, Figure A.1 shows two different categories in the special markup language (AIML) developed for AliceBot. The first category matches any sentence starting by “a . . . ” given in answer to a chatbot utterance starting by “it refers to some bad . . . ”.

The second category shows some of the possibilities offered by the AIML transformations. In the answer template:

- <*star*> is replaced by the content matched by the pattern's wildcard.
- <*get name="it"*> is replaced by the variable "it" that was set by a previous category earlier in the dialogue.
- The <*srai*> tag is a special command that tells the category searching mechanism to look for another category where the pattern matches the newly constructed sentence. It can recursively construct an answer. For example if the user is talking about cars, then the chatbot finds the answer given in the third category.

The categories are then parsed to build a "*matching tree*" where top-down searches are performed for matching the user's input and find the correct template to use. The tree is divided in three layers (Figure A.2):

The topic layer where the current *topic* is matched.

The *that* layer for specialising categories to the context.

the *pattern* layer that matches the user's input and provides a generation template.

The Figure A.3 is an example of a short conversation with the Alicebot. Utterances 5 and 6 show the pattern matching and the construction of an answer. The chatbot uses a category that matches "do you know about . . . " and construct an answer "I haven't heard of . . . ". From Utterance 10, the dialogue becomes senseless. The chatbot has no understanding of the sentence it is constructing, it is just applying the pattern matching and replacements. Utterance 14 shows how the chatbot mindlessly replaces a "you" in the user input by a "me" in the answer, generating a non grammatical sentence.

A.0.3 Analysis

Adding planning and goal seeking in this dialogue framework is not obvious at the beginning, as the different categories are not explicitly linked together –

except by topic. Wallace (2004) made an analysis of the amount of the knowledge contained in the database by visualizing the search trees resulting from the pattern matching system[4].

The aim of the current study is to detect an implicit dialogue planning. This planning structure is searched in the transitions structure between each category. In fact, the database, even if not constructed with a state-oriented perspective should contain some implicit knowledge – coming from the authors – about the right dialogue moves to have a entertaining chitchat with the user. Even if the authors did not build the categories as following each other in an obvious state-transition machine structure, the use of *topic* and *that* might create isolated sequences of utterances.

These sequences of categories would contain implicit state-machines describing strategies of dialogue that make the user comfortable. The goal of the analysis described in this section is to find such implicit structures in the publicly available collection of AIML patterns.

Patterns to States

The search tree can be seen as an unrestricted notation to encode large finite-state machines. The wildcards system allows to encode more than one state transition at once and can help develop large dialogue state-machines with generic reaction patterns.

The analysis presented here consists of linking the categories together in a graph; transforming the search tree in a state-transition machine. The assumption is that:

1. each "template" – representing one possible answer of the chatbot – can be considered as a node in the graph,

2. the "that" defines a link with a previous node,

[4]see `http://www.alicebot.org/documentation/gallery/`

3. the “pattern” is the trigger for a transition from the node defined by the “that” and the one defined by the “template”.

Following these rules, a state-machine is constructed to represent the intrinsic reasoning achieved by the chatbot. The database that is analysed in this study contains around 25,000 categories and has been manually constructed by the AliceBot community by studying the logs of chatbot conversations.

Results

The extracted graph has been studied with standard graph visualisation and analysis tools. Figure A.4 shows an example of a connected graph extracted from AliceBot’s knowledge base. “*U*” represents the user’s input – the “pattern” – and “B” the bot’s answer template.

A first observation is that not all the categories are interconnected, there are a few small isolated transition graphs (see Figure A.4). This means that the chatbot does not continue on the same topic as it has to move to another graph at the next user utterance. Hence, the chatbot loses the “memory” of the last dialogue moves and can fall into *loops*. For example, in the little state-machine described by Figure A.4, the user can close the dialogue with salutations, the chatbot replies “see you later”; However, the user can then follow with another random sentence, the chatbot then jumps to another part of the graph without remembering that it already closed the dialogue.

A few *small*, isolated, graphs of similar size and content as the one given in the example of Figure A.4 were found and contain utterances that do not give useful insight on complex dialogue move strategies.

With the exception of these *small* graphs, the whole “state-machine” is interconnected in one large graph containing thousands of nodes and is impossible to print-out in this thesis. A graph analysis tool was used to extract the strongly connected components and a limited number of strongly interconnected sub-graphs were found. These graphs are still too large – containing many states and transitions – to be of interest.

The strong interconnection between nodes and sub-graphs is produced by the overmatching pattern system, where patterns are composed of one single wildcard "*" or very under-defined patterns such as "A *". When referring to the dialogue context (*that*), these patterns produce connections between many states by matching almost any prior category.

A.0.4 Conclusion

Studying the state-transitions between all the possible context/answer pairs in the system's knowledge base produced results that did not fit the hypothesised implicit planning structure of chatbot dialogue systems. The pattern matching system shows no internal structured state graph. Extracted state-transitions are either:

- in small (2 or 3 states) clusters (see Figure A.4 for example) that give no information on any global dialogue strategy,
- in large, strongly interconnected clusters that displayed no logic in the dialogue "path" the chatbot could follow.

It seems therefore impossible to extract any pertinent knowledge of how the system is "planning" its moves and using long-term context. Categories are not developed in a "state-machine" perspective and do not seem to implicitly create one. It is therefore impossible to see a simple reuse of this knowledge-base to construct a planned dialogue system that can guarantee the achievement of persuasive goals.

These observations confirm the assumption that such chatbot frameworks have really short context memory and more advanced techniques need to be created to enable a chatbot system to achieve explicit goals.

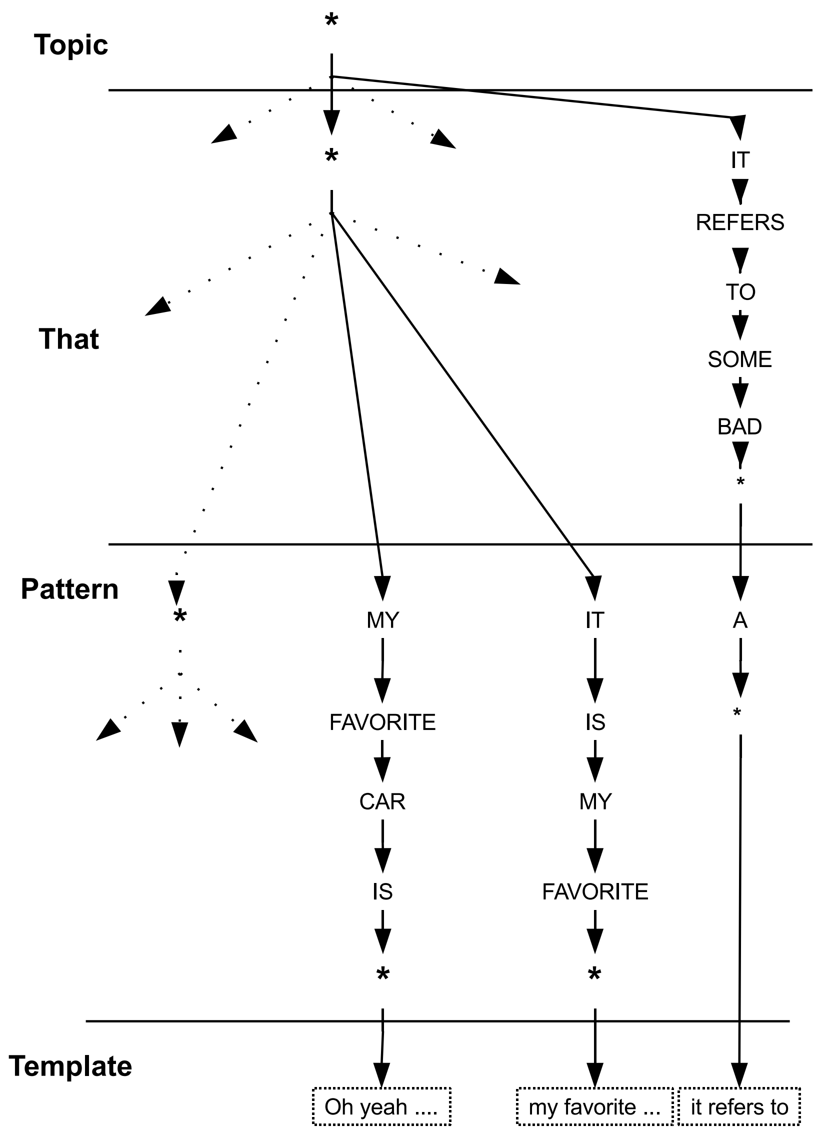


Figure A.2: Sample AIML Search Tree.

1 HUMAN: Hello
2 ALICE: Hi there!
3 HUMAN: How are you?
4 ALICE: Everything is going extremely well.
5 HUMAN: do you know about train times?
6 ALICE: I haven't heard of train times.
7 HUMAN: ok, so who are you?
8 ALICE: Hey. I am ALICE. Who are you? Tell me more.
9 HUMAN: I am Pierre, what do you wanna know?
10 ALICE: Do you mind if I tell other people you are Pierre what do me want to know. [*the chatbot pattern matching misunderstands the input.*]
11 HUMAN: not at all, but it's not completely correct you know?
12 ALICE: Would you ever do it?
13 HUMAN: do what?
14 ALICE: at all but it is not completely correct me know.

Figure A.3: Sample AliceBot Dialogue

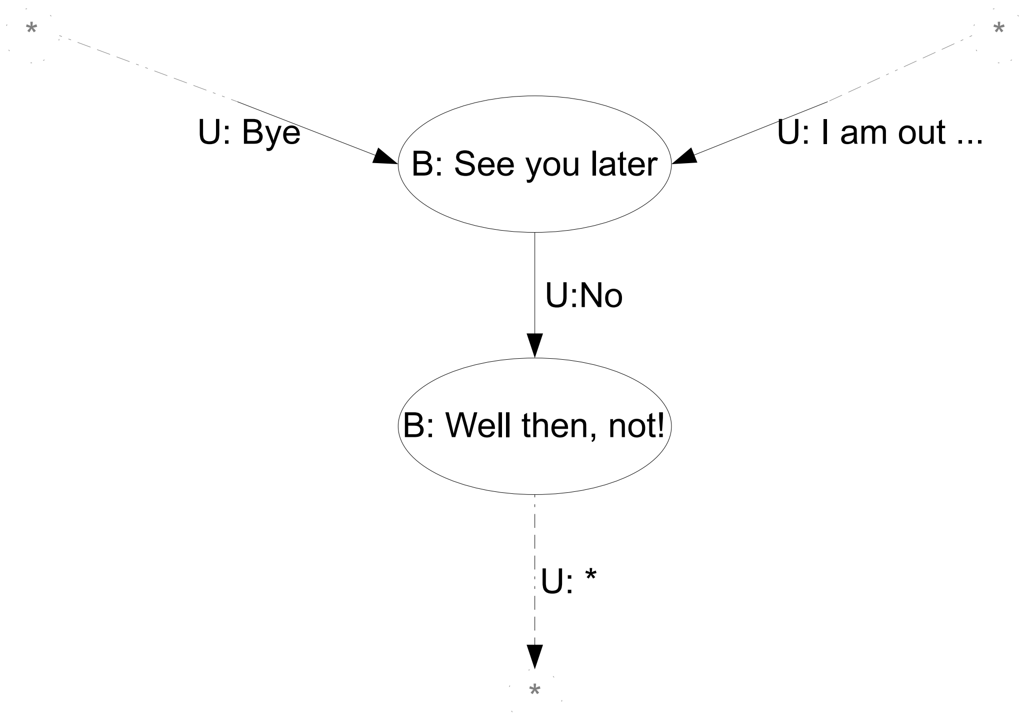


Figure A.4: Sample of the Chatbot Internal State Transitions. The “Bye” and “I am out . . . ” transitions are not connected to any previous state by a context specification (*that*). They can thus be triggered at any time in the dialogue.

Appendix B

Argumentation Scheme Classification

In the theoretical "Persuasion Machine" proposed in Gilbert et al. (2003), a major step in persuasive discourse is to identify which scheme is used by the user. According to a rhetorical framework, the dialogue manager chooses a counter-argument scheme and identifies the weak points of the opponent's argument. Different taxonomies have been proposed for the classification of argument in schemes
(Hastings 1963, Katzav & Reed 2004, Perelman & Olbrechts-Tyteca 1958; see section 2.3.1).

If the dialogue system is able to classify the scheme used by the user's argument in one of these taxonomies, it can decide which argument strategy to use to achieve stronger persuasion in its answer. In the search for a design of a persuasive dialogue framework, the feasibility of such a classifier was studied by looking at shallow features of utterances that would define their argumentation scheme.

To train and evaluate the classifier, an existing classification of argumentative texts is needed. Only one framework (Katzav & Reed 2004) was identified – at the time of the study – as being developed for Artificial Intelligence treatment of the arguments (see section 2.1.2). Katzav et al. (2004) have developed a corpus around the Araucaria software[1] to promote the research in natural argumentation. This database provides a collection of 679 small texts that contain an argument and its scheme(s).

For example:

> Vice Chancellor Brown has claimed that semesterisation would lead to a reduced workload for staff, more flexibility for students, and simpler administration for the university. It seems to me, however, that semesterisation is going to involve an enormous amount of work and should be avoided at all costs.

is tagged with the scheme "Argument from Position to Know".

These texts are used to train and evaluate the feature extraction and the classification algorithm of an argument classifier.

B.0.5 Feature Extraction

Feature extraction (FE) is performed with the FEX[2] scripting language that can easily be plugged in the SNOW[3] multi-class classifier. The feature extraction is designed to extract simple n-gram features for each example argument provided by the database; n-grams are consecutive set of *n* words from the text example that provide feature entailling the structure and content of the text. These features are easy to extract and with large corpora can provide enough information to learn the differences between classes; however, the extraction of n-grams in this application gave a classification accuracy of 11%.

A limitation of n-grams is that they limit the features to the specific lexicon provided by the training examples, which is not often complete enough to span the lexicon of all the testing examples – in particular with small training corpora.

[1] `http://araucaria.computing.dundee.ac.uk/`
[2] `http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=FEX`
[3] `http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=SNOW`

By extracting n-grams composed of the Part of Speech (POS) tags of a text, this lexicon issue can be abstracted and the learning is based on the gramatical structure of the text that is less specific than the lexicon. However, the classifier trained with this feature only achieves an accuracy of 7% – i.e. 93% of the classified utterances were assigned the wrong scheme.

These simple and shallow features may not support characteristics that define an argument and the classification results show that such simple features are probably not powerful enough to describe differences between argumentation schemes. In the scheme definition given by Perelman & Olbrechts-Tyteca (1958), scheme identification needs a deep understanding of the content of the argument, when other taxonomies are created from pragmatic features of the argument (Hastings 1963, Katzav & Reed 2004). In fact, Katzav & Reed conclude that:

> "Two arguments are of the same type if and only if they represent the same relation of conveyance and, further, represent it as ordering the arguments' conveying and conveyed facts in the same way."

This consideration can be interpreted in multiple ways, but by taking into account theories of argument structure analysis – the Rhetoric Structure Theory (RST) for example (Mann & Thompson 1988; see section 2.3.2) –, this *"relation of conveyance"* can be considered as supported by the underlying structure of the argument.

Some approaches have been proposed to automatically extract the structure of an argument (Corston-Oliver 1998b, Marcu 2000, Soricut & Marcu 2003). The first step of the *"parsing"* proposed in Marcu (2000) is the segmentation of the text into the different components of the RST structure. To perform this segmentation, Marcu divides the text around *discourse markers*, which are non-content words found in sentences that carry pragmatic meaning by structuring the sentence. These markers are words like "since", "therefore", or more complex structures such as "if ... then ..." (see Schiffrin 1988).

In this study, the initial research hypothesis that, if *discourse markers* can be used to parse RST trees and if argument schemes are determined by their structure, then such markers can be used to identify the scheme of an argument. *Discourse markers* can therefore be used as features to classify texts in different argument schemes.

Hypothetically, a supervised learning classifier will be able to learn, from lexical features, what is a discourse marker and which are the relationships between them that create the argument structure if the learning algorithm is provided with an appropriate learning corpus. Thus, there should be no need to produce a complete RST parsing of the argument or even annotate the discourse markers during the preprocessing.

For the developed classifier, simple features that contain *candidate discourse markers* tokens and details on their *ordering* are extracted from the arguments. The FEX extraction framework is not powerful enough to perform the feature extraction needed and the Minorthird (Cohen 2004) framework is used to implement a feature extractor and classifier based on the Araucaria corpus.

The candidate tokens are selected according to their presence in the WordNet[4] ontology. This ontology contains only words with meaning and therefore does not contain the functional tokens that are used as discourse markers. Hence, if a token is *not present* in the ontology, then it is considered a candidate discourse marker. A second filtering, based on part-of-speech tagging is applied to exclude the proper noun that could also be absent from the WordNet database but are not discourse markers.

The ordering of the markers component words is kept by creating features containing *sparse* n-gram formed with candidate discourse markers separated by any block of other tokens. For example, the sentence: “If God does not exist, then objective moral values do not exist.” will produce the feature: “(if).(then).”, where the “.” indicates the presence of some text block that contains no markers.

[4]`http://wordnet.princeton.edu/`

B.0.6 Results

The classifier using these features achieves an overall maximum accuracy of 15%, depending on the algorithm used for the classification – the best results being obtained with a nearest-neighbour classifier. This classifier, however, achieved an accuracy superior to 50% for specific classes – i.e. argumentation schemes.

These results can be explained – at the time of the study – by an underlying problem with the Araucaria database when used as a learning corpus. In Araucaria, more than one argumentation framework is used and independent sets of schemes – called schemesets in Araucaria – are used to annotate the arguments. The issue is that the schemes in the separate schemesets cannot be directly mapped on a common schemeset.

In fact, the database contains 192 different schemes in three identified schemesets for 650 examples. This results in some classes – or schemes – that have only one instance in the learning corpus. Therefore, it is impossible to efficiently train and evaluate the classifier on these classes and the overall accuracy is affected.

To improve the accuracy and resolve this problem, an automatic clustering method, using the same features, could be used on the database to find the relationships between the schemes in the different schemesets. Combining the results of this with the knowledge of the scheme typology, one might be able to find a mapping between the different schemeset.

In addition, the selected features are not distinctive enough to train a classifier on such a small corpus. The classification could be extended to use the parsed RST trees as features, using Support Vector Machines and Tree Kernel techniques to include complex tree structures in feature-based learning algorithms (see Moschitti 2006; for an example of parse tree classification).

Appendix C

Questionnaire for the Desert Survival Scenario

Dialogue Evaluation

Thank you for your patience and for helping with this experiment.

You are almost done, but we would like to ask you a few questions about the interaction you just had. This is the last step, just fill in the form on this page and press finish and you'll be done.

The following questions are given in the form of statements regarding the other user and your interaction. We would like you to give your level of agreement with each statement on a scale going from "Strongly Disagree" to "Strongly Agree".

please answer all questions.

About you

- **Your Email** (The email field is required to keep track of the answers. All results will be anonymised when used.)

- **Your Age**
- **Your Gender** ☐ F ☐ M

About the Conversation

- **The other user was easy to understand in the conversation.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree
- **In the conversation, the other user interpreted correctly what you said.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree
- **The pace of interaction with the other user was appropriate in this conversation.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree
- **You knew what you could say at each point of the dialogue.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree
- **The other user was sluggish and slow to reply to you in this conversation.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree
- **In this conversation, the other user interacted with you the way you expected it would.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

About the other user

- **The other user was trustworthy.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The other user was friendly.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The other user was not forceful in changing your opinion.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The other user was persuasive.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

Other Comments

C.1 Participants Answers to the Questionnaire

Table C.1 contains the answers for the questionnaire in the Desert Survival Scenario experiment (see appendix C). The *method* corresponds to the type of system they were facing (1 for *full* system, 0 for *limited* system). Each participant faced both systems, the paired results are reported consecutively. *state* is the order in which users faced each condition.

The results for *asr*, *behaviour*, *coercive*, *expertise*, *friendly*, *pace*, *persuasion*, *response*, *trust* and *tts* are on the Likert scale as described in the appendix C where 0 is “Strongly Disagree” and 5 is “Strongly Agree”.

See chapter 4 for interpretation of these results.

age	asr	behaviour	coercive	expertise	friendly	itemsType	method	pace	persuasion	response	gender	state	tau_after	tau_before	trust	tts
27	2	1	3	3	3	0	0	4	2	0	f	1	4	4	2	3
27	2	3	2	3	3	1	1	4	3	0	f	0	5	5	2	3
38	0	0	3	0	2	0	0	3	0	0	f	0	6	6	2	3
38	0	1	2	0	2	1	1	1	2	0	f	1	7	7	2	3
59	1	1	3	1	3	0	0	3	3	3	m	1	4	1	3	4
59	0	1	2	1	3	1	1	2	3	3	m	0	4	5	3	4
31	1	1	3	1	1	1	1	1	2	3	f	1	3	3	1	1
31	1	2	3	2	3	0	0	3	3	0	f	0	2	3	2	3
27	3	3	3	3	3	1	1	3	3	1	m	0	2	6	2	4
27	4	4	3	4	4	0	0	4	3	0	m	1	2	3	4	4
23	1	1	3	3	3	0	0	1	3	1	f	0	8	4	3	2
23	0	3	3	3	4	1	1	1	0	1	f	1	3	5	4	3
20	0	1	3	1	2	1	0	1	2	3	f	0	5	5	1	1
20	0	3	0	1	3	0	1	3	3	1	f	1	1	9	1	3
28	3	3	1	3	3	0	0	3	3	2	m	1	3	8	3	3
28	3	2	1	3	3	1	1	2	2	2	m	0	5	5	2	3
51	1	1	1	1	1	0	0	3	3	1	m	0	8	5	2	4
51	1	1	3	3	3	1	1	3	0	1	m	1	7	7	1	3
23	2	2	2	2	2	0	0	2	2	2	m	1	3	3	2	2
23	1	3	3	4	3	3	1	4	3	0	m	0	0	0	1	4
23	2	2	4	3	4	1	1	4	4	0	f	0	4	5	3	4
23	4	3	4	4	4	0	0	4	4	0	f	1	5	2	4	4
41	2	1	3	2	3	3	0	3	2	1	m	1	0	0	3	2
41	2	1	3	2	3	3	1	3	2	1	m	0	0	0	3	2
26	2	1	3	3	3	0	1	3	3	3	f	0	3	2	3	3
26	3	3	3	3	3	1	0	3	3	1	f	1	3	5	3	3
22	3	3	3	3	3	0	1	4	3	1	m	1	2	2	3	4
22	3	4	3	3	3	1	0	4	4	4	m	0	3	2	3	4
45	3	3	2	3	2	0	0	3	2	1	m	1	3	1	3	3
45	3	3	3	3	3	1	1	3	4	1	m	0	3	7	3	4
26	4	4	4	2	4	1	0	4	1	1	m	0	5	5	4	3
26	3	3	2	3	3	0	1	4	3	1	m	1	4	8	4	3

Table C.1: Answers to the Questionnaire for the Desert Survival Scenario

Appendix D

Questionnaire for the Initial Restaurant Scenario

Dialogue Evaluation

Thank you for your patience and for helping with this experiment.

You are almost done, but we would like to ask you a few questions about the interaction you just had. This is the last step, just fill in the form on this page and press finish and you'll be done.

please answer all questions

About you

- **Your Email** (The email field is required to keep track of the answers. All results will be anonymised when used.)
- **Your Age**
- **Your Gender** ☐ F ☐ M

About your Personality[1]

Here are a number of personality traits that may or may not apply to you. Please select the extent to which you agree or disagree with the statement. You should rate the extend to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

- **Extraverted, enthusiastic.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Critical, quarrelsome.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Dependable, self-disciplined.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Anxious, easily upset.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Open to new experiences, complex.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Reserved, quiet.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

[1] This questionnaire evaluates the big five personality traits; from Gosling et al. (2003)

- **Sympathetic, warm.**
☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Disorganized, careless.**
☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Calm, emotionally stable.**
☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Conventional, uncreative.**
☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

About the Conversation

The following questions are given in the form of statements regarding the computer during its interaction. We would like you to give your level of agreement with each statement on a scale going from ”Strongly Disagree” to ”Strongly Agree”.

- **The computer was easy to understand in the conversation.**
☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **In the conversation, the computer interpreted correctly what you said.**
☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The pace of interaction with the computer was appropriate in this conversation.**

☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **You knew what you could say at each point of the dialogue.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer was sluggish and slow to reply to you in this conversation.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer acting like he knew you was irritating.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **In this conversation, the computer interacted with you the way you expected it would.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

About the computer

- **The computer was trustworthy.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer was friendly.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer was not forceful in changing your opinion.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **I liked that the computer was outgoing.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer was persuasive.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

What did you think of the computer ”personality”

Other Comments

D.1 Participants Answers to the Questionnaire

Table E.1 reports the answers given by the participants in the first experiment on personality and persuasion reported in Chapter 5. *agreeableness*, *conscientiousness*, *emotional_stability*, *extraversion* and *openness* are the Big Five personality traits of the participants from the TIPI inventory (Gosling et al. 2003) (*p2* to *p10* being the answers to the individual statements of the inventory). *sysperso* is one of the two system's personality conditions randomly chosen for the generator.

The results for *asr*, *behaviour*, *coercive*, *expertise*, *friendly*, *pace*, *persuasion*, *response*, *trust*, *outgoing* and *tts* are on the Likert scale as described in the appendix D where 0 is "Strongly Disagree" and 5 is "Strongly Agree".

age	agreeableness	asr	behaviour	coercive	conscientiousness	emotional_stability	expertise	extraversion	familiarity	friendly	openness	outgoing	p10	p2	p3	p4	p5	p6	p7	p8	p9	pace	persuasion	response	gender	sysperso	tau_after	tau_before	trust	tts
26	6	3	2	3	6.5	6	3	4	2	3	6	2	6	6	7	6	6	2	6	6	6	3	3	1	f	extravert	2	3	3	3
35	2.5	0	3	0	6	6.5	3	5	2	0	7	2	7	3	7	7	7	6	2	5	6	3	0		f	extravert	3	3	0	3
30	6	1	3	1	2.5	3	1	7	2	2	6.5	2	6	5	2	3	7	7	7	3	3	2	2	0	m	extravert	3	3	2	3
30	6	1	1	1	2	4	1	7	3	2	6.5	4	6	5	2	5	7	7	7	2	3	3	4	1	m	extravert	0	3	1	3
32	3.5	2	2	2	5	5.5	4	4	1	2	6	2	6	2	6	6	6	4	5	4	5	2	2	0	m	extravert	3	3	2	4
26	3.5	2	2	3	7	4.5	3	6.5	2	3	4.5	2	3	1	7	4	6	6	6	7	5	4	1	1	m	extravert	3	3	1	3
33	4.5	3	3	1	4.5	4.5	3	4.5	3	2	5	2	4	3	6	3	6	3	6	3	6	3	3	1	f	extravert	3	3	2	4
30	4.5	3	3	3	4	6	2	6	2	3	5.5	3	5	5	4	6	6	6	4	4	6	3	3	2	m	extravert	1	3	3	3
41	6	2	1	1	6.5	5	0	4	1	2	6.5	2	7	6	6	4	6	2	6	7	6	2	1	1	m	extravert	3	3	1	3
29	5	1	2	2	5.5	4	2	4.5	1	3	5	3	3	5	6	4	7	3	5	5	4	3	4	1	f	extravert	3	3	1	1
24	5	3	2	1	4	4.5	3	4	1	3	5	3	5	4	4	5	5	3	6	4	4	3	4	1	f	extravert	0	1	1	3
28	4.5	3	2	2	1.5	3	1	4	3	3	3	2	3	3	2	5	3	3	6	1	1	3	2	1	m	extravert	3	3	2	3
31	4	3	2	1	4.5	5	3	3.5	3	3	3	2	4	3	3	5	2	3	5	6	5	2	3	2	m	extravert	2	3	2	3
25	6.5	3	2	3	4.5	2	4	3.5	1	4	6	3	5	6	6	2	7	1	7	3	2	3	1	0	f	extravert	3	3	1	3
28	3.5	3	2	1	6.5	5	2	6	1	3	3	3	5	3	6	6	1	5	4	7	4	1	3	1	f	extravert	0	3	3	3
32	5.5	2	1	2	5.5	3.5	3	6	3	3	5.5	3	5	6	5	3	6	6	5	6	4	3	3	0	m	extravert	2	3	2	3
59	4.5	1	3	3	4	3.5	3	2	4	3	5.5	1	5	3	5	3	6	1	6	3	4	3	1	2	m	extravert	3	3	3	3
29	4	1	0	3	4	6	3	5.5	0	3	6	2	5	6	2	6	7	6	2	6	6	3	2	0	m	extravert	3	3	2	3
32	4	2	3	2	4.5	4	2	5	1	3	5	3	5	3	4	5	5	5	5	5	3	3	3	1	f	introvert	2	3	3	3
25	5	4	3	1	4	4	3	3.5	2	3	5	2	5	4	4	3	5	2	6	4	5	4	3	1	f	introvert	1	3	3	4
30	4.5	3	3	1	5.5	5	3	5	1	3	5.5	2	6	3	4	5	5	6	6	7	5	3	2	1	m	introvert	3	3	2	3
25	4.5	3	3	4	5.5	5	3	3	0	3	5	3	4	3	6	4	6	1	6	5	6	3	2	1	m	introvert	1	3	3	3
26	6	0	1	1	6.5	5	1	4	0	2	6	2	6	6	7	4	6	3	6	6	6	3	1	1	f	introvert	3	3	2	3
26	6	3	2	2	6	5.5	3	4.5	0	2	6	2	6	6	6	5	6	3	6	6	6	3	3	1	f	introvert	2	3	3	3
31	4.5	1	3	2	6.5	4	2	3	1	3	5.5	2	5	6	7	4	6	4	3	6	4	4	0	1	m	introvert	3	3	2	3
37	3.5	3	3	3	4.5	4	2	4	3	3	4.5	1	7	4	3	4	2	2	3	6	4	3	2	1	f	introvert	3	3	2	3
19	4.5	3	3	3	7	3.5	2	2.5	0	2	4	2	3	5	7	3	5	2	4	7	4	3	1	0	f	introvert	3	3	3	4
28	5	3	3	3	6	7	3	6.5	1	3	7	3	7	3	6	7	7	6	7	6	7	4	4	0	m	introvert	2	3	3	3

Table D.1: Answers to the Restaurant Questionnaire

Appendix E

Questionnaire for the Extended Restaurant Scenario

Dialogue Evaluation

Thank you for your patience and for helping with this experiment.

You are almost done, but we would like to ask you a few questions about the interaction you just had. This is the last step, just fill in the form on this page and press finish and you'll be done.

please answer all questions

About you

- **Your Email** (The email field is required to keep track of the answers. All results will be anonymised when used.)
- **Your Age**
- **Your Gender** ☐ F ☐ M

About your Personality

Here are a number of personality traits that may or may not apply to **you**. Please select the extent to which you agree or disagree with the statement. You should rate the extend to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

- **Extraverted, enthusiastic.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Critical, quarrelsome.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Dependable, self-disciplined.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Anxious, easily upset.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Open to new experiences, complex.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Reserved, quiet.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree
- **Sympathetic, warm.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Disorganized, careless.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Calm, emotionally stable.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Conventional, uncreative.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

About the Conversation

The following questions are given in the form of statements regarding the computer during its interaction. We would like you to give your level of agreement with each statement on a scale going from ”Strongly Disagree” to ”Strongly Agree”.

- **I found the fact that the computer acted as if it knew me was irritating.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **In this conversation, the computer interacted with you in a natural way.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer was trustworthy.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **I found the computer's extravert, enthusiastic tone irritating.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **I believe that the computer did not strongly influence me to change my opinion.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

- **The computer was persuasive.**
  ☐ Strongly disagree ☐ Disagree ☐ Neither agree nor disagree ☐ Agree ☐ Strongly Agree

About the computer

Here are a number of personality traits that may or may not apply to **the simulated interlocutor**. Please select the extent to which you agree or disagree with the statement. You should rate the extend to which the pair of traits applies to the simulated interlocutor, even if one characteristic applies more strongly than the other.

- **Extraverted, enthusiastic.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Critical, quarrelsome.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Dependable, self-disciplined.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Anxious, easily upset.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Open to new experiences, complex.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Reserved, quiet.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Sympathetic, warm.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Disorganized, careless.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Calm, emotionally stable.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

- **Conventional, uncreative.**
  ☐ Disagree Strongly ☐ Moderately Disagree ☐ Disagree a little ☐ Neither agree nor disagree ☐ Agree a little ☐ Agree moderately ☐ Agree

Other Comments

E.1 Participants Answers to the Questionnaire

Table E.1 reports the answers given by the participants in the first experiment on personality and persuasion reported in Chapter 5. *agreeableness*, *conscientiousness*, *emotional_stability*, *extraversion* and *openness* are the Big Five personality traits of the participants from the TIPI inventory (Gosling et al. 2003) (*p2* to *p10* being the answers to the individual statements of the inventory). *sysperso* is one of the two system's personality conditions randomly chosen for the generator.

The results for *asr*, *behaviour*, *coercive*, *expertise*, *friendly*, *pace*, *persuasion*, *response*, *trust*, *outgoing* and *tts* are on the Likert scale as described in the appendix D where 0 is "Strongly Disagree" and 5 is "Strongly Agree".

age	agreeableness	behaviour	coercive	conscientiousness	defence	drop	emotional_stability	extrairitating	extraversion	familiarity	neg	openness	p10	p2	p3	p4	p5	p6	p7	p8	p9	persuasion	pos	ps10	ps2	ps3	ps4	ps5	ps6	ps7	ps8	ps9	hasrepeat	repetition2	repetition2Cnt	repetitionCnt	sex	sysperso	tau_after	tau_before	trust	usr	woz	repetition	p1	ps1
58	4.0	2	4	4.0	5	0	4.0	1	2.0	1	5	5.0	4	4	4	3	6	3	4	4	5	3	3	4	3	4	6	3	4	2	5	4	no	1.01e-01	2.44e-01	8.89e-02	f	extrovert	3	3	1	8	10	21.511	1	4
60	4.5	2	2	3.5	4	0	3.5	3	1.0	3	6	5.5	5	3	5	2	6	1	6	2	5	1	2	3	3	5	5	3	5	5	5	5	no	5.31e-02	8.89e-02	2.44e-01	m	extrovert	1	3	3	8	10	18.111	1	3
52	5.5	1	3	4.5	3	0	6.0	3	4.0	3	4	3.5	3	6	6	6	4	3	5	3	6	1	2	3	2	4	2	2	6	5	3	2	no	7.29e-02	1.79e-01	1.07e-01	m	extrovert	0	3	2	6	8	23.321	5	7
22	5.0	3	3	5.0	2	0	5.0	4	4.5	0	3	3.0	3	5	5	5	3	5	5	5	5	1	3	5	3	3	5	3	5	3	5	3	no	8.37e-02	1.39e-01	1.39e-01	m	introvert	3	3	1	6	9	24.806	4	5
30	3.5	1	1	6.5	4	0	4.5	1	7.0	1	5	6.5	7	2	6	6	6	7	5	7	3	2	3	3	2	3	3	2	4	6	3	4	no	1.25e-01	2.44e-01	4.44e-02	f	introvert	3	3	1	8	10	23.667	7	5
25	2.0	1	4	2.5	4	2	4.0	3	3.0	0	10	3.0	3	2	2	3	3	4	2	3	5	1	1	5	5	4	5	4	6	4	4	3	no	7.77e-02	1.21e-01	1.82e-01	m	extrovert	3	3	1	11	12	22.288	2	5
28	7.0	4	2	6.5	3	0	6.5	1	6.5	2	5	6.0	6	7	6	7	6	6	7	7	6	3	1	7	6	6	7	6	7	6	7	4	no	9.31e-02	2.50e-01	7.14e-02	m	introvert	2	3	3	6	8	28.429	7	6
39	5.5	0	4	3.0	5	1	6.0	4	3.0	0	9	7.0	7	5	3	6	7	3	6	3	6	0	2	1	3	4	4	1	6	1	4	4	yes	1.28e-01	2.58e-01	1.52e-01	m	introvert	3	3	2	11	12	18.985	3	6
51	2.5	1	1	5.0	0	0	4.0	1	5.0	2	0	4.0	6	3	4	6	2	6	2	6	2	1	3	6	6	2	6	2	6	2	6	2	no	7.15e-02	1.00e-01	1.00e-01	f	introvert	3	3	1	3	5	20.700	4	6
28	4.0	3	4	5.5	1	1	4.5	2	1.0	1	3	5.0	5	3	5	3	5	1	5	6	6	0	1	4	7	4	6	5	6	6	5	5	no	6.06e-02	1.00e-01	1.00e-01	m	extrovert	3	3	3	4	5	27.100	1	6
43	4.5	1	1	6.5	2	0	6.5	1	2.0	1	2	5.0	4	5	7	7	6	2	4	6	6	1	3	4	2	4	4	4	6	5	4	5	no	6.14e-02	9.52e-02	1.43e-01	m	extrovert	2	3	3	5	7	22.762	2	7
29	4.5	3	3	6.0	3	0	4.0	2	5.0	1	3	5.5	5	3	6	4	6	4	6	6	4	2	4	4	5	3	3	2	3	6	6	1	no	1.38e-01	3.57e-01	7.14e-02	f	introvert	3	3	3	7	8	24.643	6	2
43	4.5	2	3	1.5	2	0	6.0	3	3.0	1	3	6.5	7	5	2	6	6	3	4	1	6	1	3	3	3	4	4	3	5	5	4	4	no	6.91e-02	1.43e-01	4.76e-02	f	extrovert	2	3	3	6	7	21.190	3	5
23	1.5	0	4	6.0	0	2	5.0	4	1.5	4	5	7.0	7	1	6	4	7	1	2	6	6	0	1	1	4	4	7	4	7	1	4	2	yes	1.39e-01	1.90e-01	2.86e-01	m	introvert	3	3	2	6	7	22.952	2	7
26	4.0	1	3	5.0	3	0	4.0	2	5.5	0	7	6.0	6	6	2	3	6	6	7	1	3	3	2	2	1	3	7	1	1	1	4	3	yes	1.36e-01	2.73e-01	2.18e-01	m	introvert	0	3	1	9	11	20.709	6	1
28	5.5	1	3	4.0	2	1	6.0	0	3.0	0	4	6.5	6	3	6	2	7	6	5	4	6	1	3	5	2	4	4	6	7	5	4	2	no	8.54e-02	1.33e-01	1.11e-01	m	introvert	3	3	3	7	10	20.356	5	5
35	4.5	2	1	6.0	1	0	6.0	2	3.5	0	5	6.0	6	5	5	6	6	2	6	3	6	1	2	2	2	2	6	3	4	4	2	4	no	6.86e-02	2.38e-01	4.76e-02	f	extrovert	2	3	2	7	7	26.143	4	3
28	3.0	3	3	6.0	0	0	6.5	1	4.5	2	1	6.5	6	3	6	5	7	2	6	6	7	2	3	4	3	4	4	4	5	5	4	4	no	8.72e-03	0	0	m	extrovert	1	3	2	4	5	23.600	5	5
30	4.0	0	4	4.0	1	2	4.0	0	4.0	2	0	5.5	5	3	6	6	6	4	3	6	7	3	1	4	5	4	4	4	4	6	6	5	no	1.03e-01	1.33e-01	6.67e-02	m	introvert	1	3	3	1	6	25.400	5	4

age	agreeableness	behaviour	coercive	conscientiousness	defence	drop	emotional_stability	extrairitating	extraversion	familiarity	neg	openness	p10	p2	p3	p4	p5	p6	p7	p8	p9	persuasion	pos	ps10	ps2	ps3	ps4	ps5	ps6	ps7	ps8	ps9	hasrepeat	repetition2	repetition2Cnt	repetitionCnt	sex	sysperso	tau_after	tau_before	trust	usr	woz	repetition	p1	ps1
25	3.5	0	3	7.0	4	2	3.0	1	6.5	2	5	4.0	4	4	4	4	4	4	4	4	4	0	0	4	4	4	4	4	4	4	4	4	no	0	0	3.00e-01	m	introvert	3	3	0	5	5	23	4	4
52	4.5	3	3	5.5	0	1	6.0	1	4.5	2	9	6.0	7	1	7	2	5	6	6	7	4	0	1	1	4	1	7	3	4	4	7	7	no	6.50e-02	1.09e-01	1.27e-01	f	introvert	1	3	0	10	11	24.345	7	5
26	4.0	0	4	4.0	1	0	2.0	2	4.5	1	3	6.0	6	3	7	6	6	4	6	4	6	3	2	4	4	4	4	4	4	5	4	4	no	9.03e-02	2.00e-01	1.33e-01	m	introvert	1	3	2	5	6	27.067	5	6
23	4.0	2	3	3.5	5	0	3.0	3	5.0	3	1	7.0	7	3	5	2	7	4	5	3	2	0	3	7	4	1	7	1	7	1	7	1	no	6.30e-02	1.00e-01	2.00e-01	f	extrovert	3	3	1	4	5	18.200	5	1
26	5.5	2	1	5.0	3	1	5.0	2	3.0	1	7	7.0	7	2	4	3	7	4	6	3	3	4	3	3	2	5	3	4	2	4	3	3	no	8.83e-02	2.12e-01	7.58e-02	m	extrovert	3	3	2	10	12	26.985	6	7
53	6.5	1	1	4.5	4	0	6.0	4	5.5	3	7	7.0	7	5	5	5	7	3	6	5	5	4	1	4	3	4	6	3	5	4	4	4	no	2.20e-01	2.78e-01	3.06e-01	f	extrovert	1	3	2	8	9	16.361	3	5
25	5.0	1	1	4.0	1	0	5.5	2	5.0	2	4	3.0	3	7	6	6	3	5	6	3	6	1	2	4	2	4	3	4	5	2	3	3	no	8.30e-02	1.67e-01	8.33e-02	m	introvert	3	3	1	6	9	24.028	6	6
35	4.0	1	4	4.0	3	0	2.5	2	6.5	2	2	3.0	2	4	5	5	4	5	6	3	6	1	2	3	3	4	4	4	4	5	4	4	no	5.67e-02	6.67e-02	1.33e-01	m	introvert	3	3	2	4	6	24.467	5	4
33	5.0	1	4	4.5	2	0	5.0	2	4.0	3	4	7.0	7	1	5	1	7	6	7	3	4	0	3	4	4	4	4	4	4	4	4	4	yes	1.54e-01	3.21e-01	1.79e-01	f	introvert	3	3	1	7	8	18.964	7	4
54	4.5	0	4	2.5	4	0	4.0	4	2.5	3	3	6.5	6	5	4	4	7	3	5	5	6	2	1	4	4	4	4	4	4	4	4	4	no	6.60e-02	2.00e-01	1.33e-01	m	introvert	2	3	1	4	6	27.733	5	4
26	4.0	1	1	5.0	4	0	3.0	2	4.5	2	6	6.5	6	5	3	5	7	3	4	2	3	0	1	1	4	4	4	4	4	4	4	4	yes	2.00e-01	4.29e-01	2.14e-01	f	introvert	3	3	0	7	8	20.321	2	4
25	4.5	3	3	4.5	2	0	4.0	2	4.5	3	5	7.0	7	3	7	3	7	3	5	3	3	1	1	2	3	4	4	1	4	3	4	4	no	8.19e-02	2.22e-01	2.78e-02	f	introvert	0	3	2	6	9	27.444	6	4
40	7.0	0	4	4.0	3	0	6.5	4	6.5	0	3	6.0	7	2	4	4	5	4	7	5	4	3	1	4	4	4	6	4	5	4	4	4	no	5.05e-02	1.07e-01	3.57e-02	m	introvert	0	3	2	4	8	22.179	5	5
30	3.5	3	3	6.0	0	0	4.5	2	3.0	3	7	7.0	7	7	6	6	7	6	7	2	7	0	3	2	2	2	3	1	7	1	5	2	yes	2.18e-01	3.82e-01	2.55e-01	m	introvert	3	3	1	10	11	20.909	7	7
41	5.5	3	2	4.5	1	0	5.0	2	2.5	1	0	5.5	6	3	6	4	5	2	4	6	5	2	4	4	6	5	4	5	4	4	6	6	no	1.57e-01	3.00e-01	1.00e-01	f	introvert	0	3	1	4	5	25.300	4	3
26	4.0	2	2	6.0	2	0	5.5	3	6.0	3	2	6.0	6	6	4	6	6	2	5	5	4	2	1	4	6	4	6	5	6	4	4	5	no	8.25e-03	0	1.00e-01	f	introvert	1	3	2	3	5	20.800	3	5
27	7.0	1	1	6.0	3	0	6.0	3	2.0	2	2	6.0	6	2	5	6	6	6	6	7	5	1	3	5	2	2	4	6	6	4	1	2	no	1.36e-01	2.00e-01	2.67e-01	m	introvert	1	3	2	5	6	21.267	6	7
41	6.0	1	3	6.5	1	0	7.0	1	5.5	0	3	4.0	3	7	5	6	5	1	7	7	6	2	3	5	2	1	4	3	7	3	4	4	yes	1.20e-01	2.67e-01	0	f	extrovert	0	3	1	6	6	26.667	3	7
20	3.5	1	4	4.5	5	0	3.5	2	6.0	1	1	5.0	4	6	6	7	6	6	6	7	7	1	1	5	5	3	5	3	5	3	5	3	no	6.93e-02	2.00e-01	0	f	introvert	1	3	3	2	6	27.867	5	3
22	4.5	3	4	3.5	0	1	4.5	1	3.5	1	7	5.5	5	2	5	4	6	6	5	4	3	1	2	4	4	3	7	3	6	5	3	3	no	8.87e-02	3.33e-01	2.22e-02	f	extrovert	3	3	2	9	10	25.556	6	5

age	agreeableness	behaviour	coercive	conscientiousness	defence	drop	emotional_stability	extrairitating	extraversion	familiarity	neg	openness	p10	p2	p3	p4	p5	p6	p7	p8	p9	persuasion	pos	ps10	ps2	ps3	ps4	ps5	ps6	ps7	ps8	ps9	hasrepeat	repetition2	repetition2Cnt	repetitionCnt	sex	sysperso	tau_after	tau_before	trust	usr	woz	repetition	p1	ps1
40	4.0	0	4	3.0	4	2	5.0	2	6.0	4	3	5.5	4	2	6	3	7	1	7	1	6	3	1	4	5	5	2	3	6	7	3	7	no	5.97e-02	1.33e-01	2.00e-01	m	extrovert	3	3	3	4	6	22.867	6	6
17	4.0	1	4	5.5	0	1	3.5	2	4.5	3	9	7.0	7	2	5	5	7	6	6	1	5	0	1	4	1	1	4	4	4	4	4	4	no	1.32e-01	3.27e-01	5.45e-02	f	extrovert	3	3	0	10	11	26.655	6	7
55	5.5	1	3	3.5	2	1	5.5	2	4.5	2	3	6.0	6	3	6	3	6	4	5	5	4	1	1	3	3	2	5	2	4	2	4	5	no	5.62e-02	2.00e-01	0	m	introvert	3	3	1	4	5	29.700	5	5
38	3.5	0	4	7.0	1	1	6.5	3	6.0	1	6	5.5	5	4	5	5	6	4	7	2	6	2	1	2	1	3	2	4	6	1	2	1	no	8.71e-02	2.50e-01	7.14e-02	f	extrovert	2	3	1	7	8	27.893	5	7
24	2.5	1	4	5.0	1	0	6.5	4	2.5	3	4	6.5	6	3	7	7	7	6	4	7	6	0	2	2	5	1	4	4	5	1	1	3	yes	1.26e-01	1.90e-01	1.43e-01	m	extrovert	3	3	0	6	7	21.476	6	5
23	4.5	1	1	2.5	5	1	1.5	4	1.5	4	1	6.5	6	2	6	7	7	4	3	4	6	0	4	4	2	2	3	4	7	1	3	1	no	4.43e-02	1.43e-01	0	f	introvert	3	3	0	5	7	27.619	1	1
33	4.0	0	4	6.5	0	0	3.5	4	3.0	3	7	5.5	6	3	3	1	5	1	6	2	2	1	1	3	5	4	6	5	6	5	4	5	no	7.72e-02	2.50e-01	8.33e-02	m	extrovert	2	3	2	8	9	25.583	2	6
54	6.5	0	3	6.0	1	0	4.0	4	6.0	4	0	5.5	5	3	6	3	6	3	5	7	4	0	2	3	3	4	4	4	5	3	2	1	yes	9.87e-02	2.00e-01	1.33e-01	f	extrovert	1	3	1	2	6	25.333	3	5
17	4.5	1	3	6.0	2	0	3.5	3	6.0	3	1	6.0	5	6	6	3	7	6	7	6	5	1	1	4	3	1	7	5	7	4	4	4	no	8.56e-02	9.52e-02	3.81e-01	f	introvert	1	3	1	2	7	20.143	6	7
28	4.5	1	3	4.0	4	0	4.5	4	4.5	4	3	6.5	6	3	6	3	7	5	6	6	4	2	3	3	2	3	4	3	5	2	4	3	no	8.87e-02	2.00e-01	0	f	extrovert	1	3	1	6	6	29.933	7	5
32	4.5	0	3	5.5	3	0	4.0	4	1.5	4	5	5.0	4	3	4	5	6	3	6	4	4	0	2	4	1	4	6	2	6	2	4	4	yes	8.02e-02	1.43e-01	1.43e-01	f	extrovert	2	3	1	7	8	21	6	6
33	3.5	2	1	6.0	2	0	4.0	3	6.5	1	4	6.0	6	6	5	3	6	1	3	6	5	1	1	4	4	3	6	4	5	5	3	3	no	6.27e-02	1.94e-01	1.11e-01	m	extrovert	1	3	1	5	9	24.972	2	6
34	4.0	1	4	5.5	0	1	3.0	2	6.0	0	3	7.0	7	2	6	2	7	6	5	6	6	3	1	6	2	2	4	2	6	3	5	3	no	3.60e-02	1.43e-01	1.43e-01	m	extrovert	2	3	1	4	8	22.964	7	5

Table E.1: Answers to the Restaurant Questionnaire

Appendix F

Desert Survival Scenario – Argumentation Model

This appendix provides an overview of the argumentation hierarchy developed for the Desert Survival Scenario. Figure F.1 shows the whole desert survival model extracted from the logical description provided to the planner. Figure F.2 and F.3 are excerpts extracted from this model.

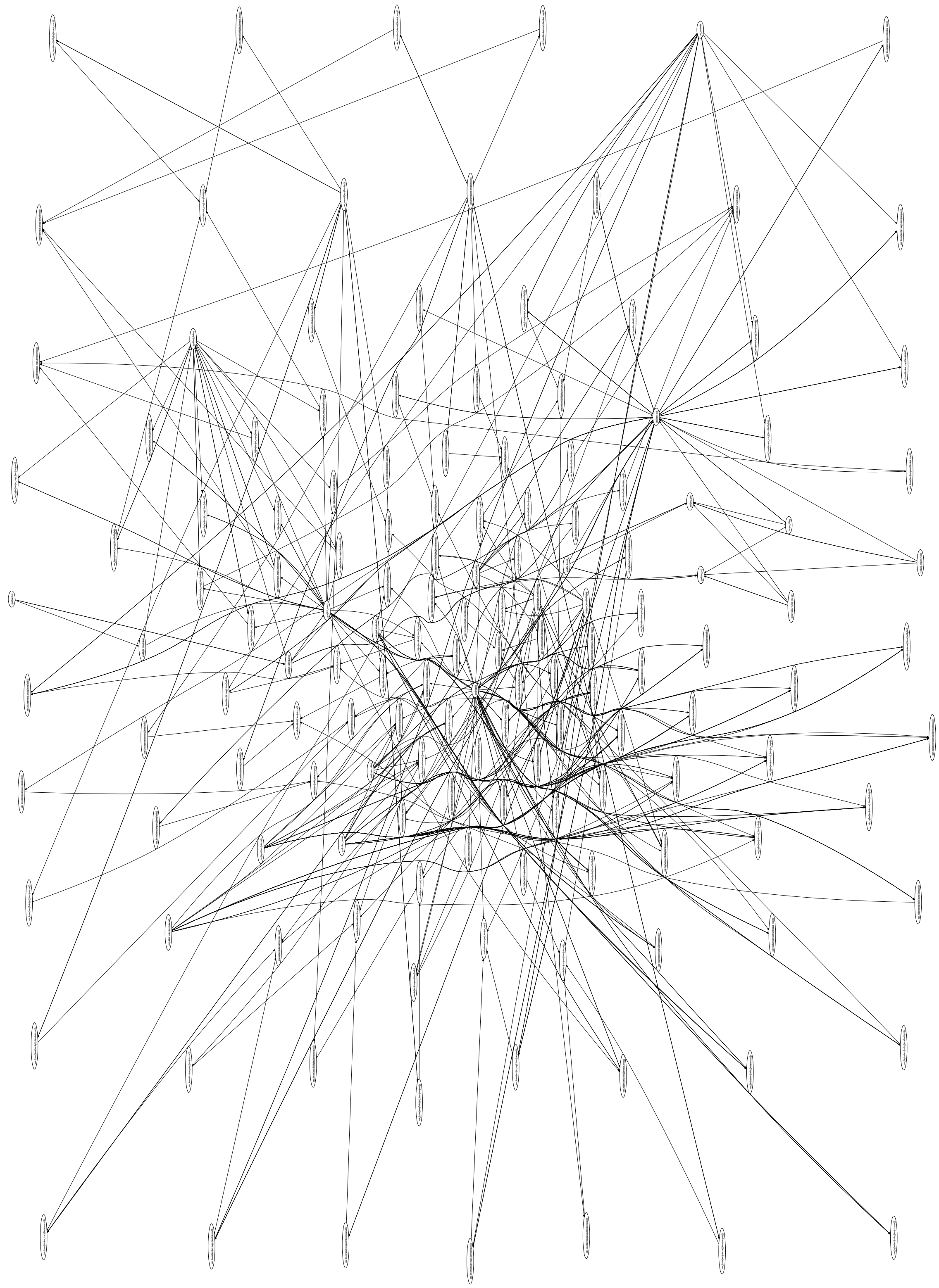


Figure F.1: Desert Survival Scenario Full Argumentation Model

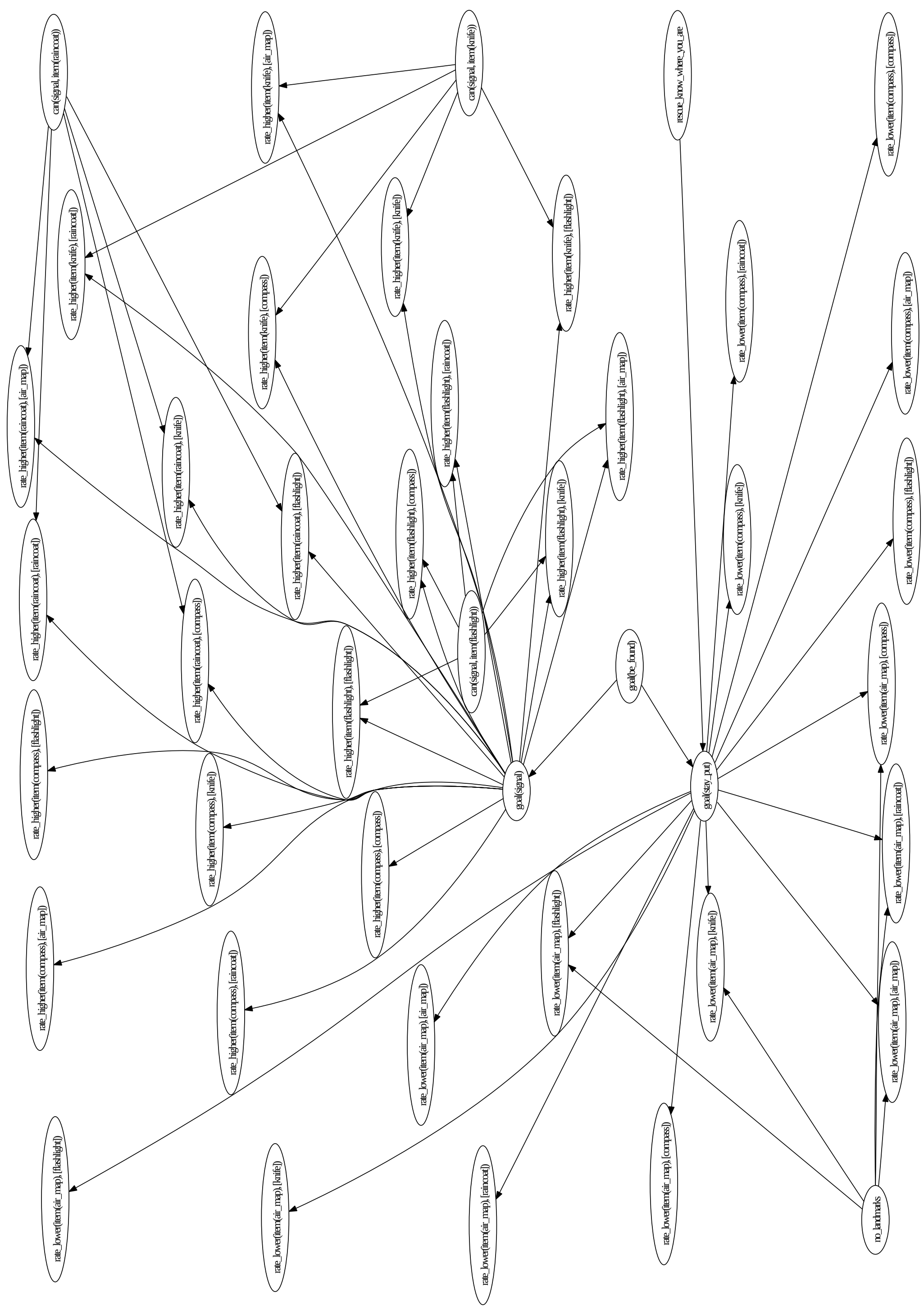


Figure F.2: Excerpt of the Desert Survival Argumentation Model

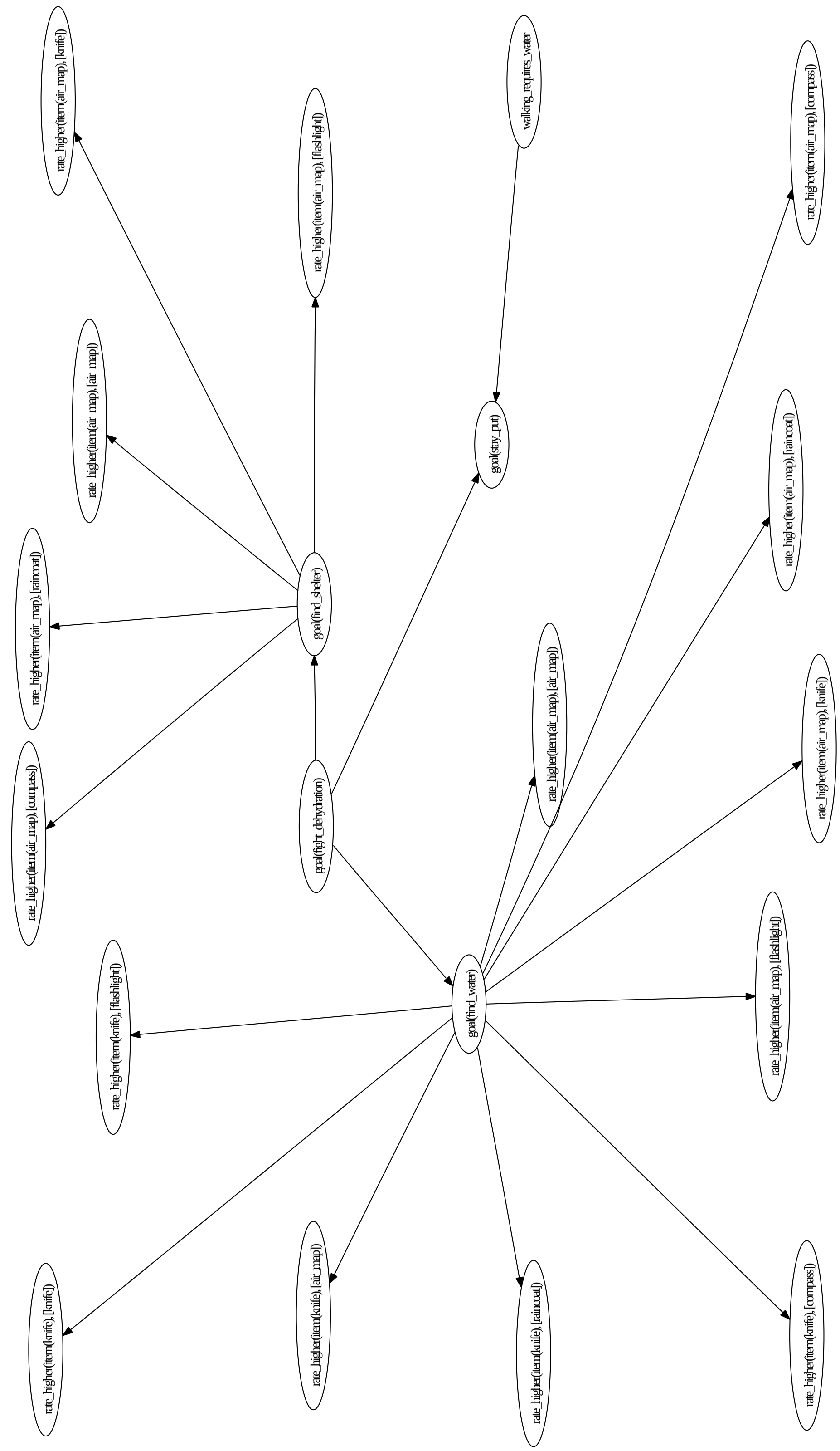


Figure F.3: Excerpt of the Desert Survival Argumentation Model

List of References

Aberdein, A. (2005). The uses of argument in mathematics. *Argumentation*, *19*(3), 287–301.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behaviour*. Englewood Cliffs, New Jersey: Prentice-Hall.

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001a). Toward conversational human-computer interaction. *AI Magazine*, *22*(4), 27–37.

Allen, J. F., Ferguson, G., Miller, B. W., Ringger, E. K., & Sikorski, T. (2000). *Dialogue Systems: From Theory to Practice in TRAINS-96*, chap. 14, (pp. 347–376). Marcel Dekker.

Allen, J. F., Ferguson, G., & Stent, A. (2001b). An architecture for more realistic conversational systems. In *IUI '01: Proceedings of the 6th international conference on Intelligent user interfaces*, (pp. 1–8). New York, NY, USA: ACM Press.

Aronson, E. (1968). Dissonance theory: Progress and problems. In R. P. Abelson, E. Aronson, W. J. Mcguire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.) *Theories of cognitive consistency: A sourcebook*, (pp. 5–27). Chicago: Rand McNally.

Barnes, J. (1998). *The Complete Works of Aristotle: Revised Oxford*. Princeton University Press.

Basili, R., De Cao, D., Giannone, C., & Marocco, P. (2007). Data-driven dialogue for interactive question answering. In *Porceedings of the 10th congress if Italian Association for Artificial Intelligence (AI*IA'07): Artificial Intelligence and Human-Oriented Computing*, (pp. 326–338).

Bench-Capon, T. J. M. (2003a). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, *13*(3), 429–448.

Bench-Capon, T. J. M. (2003b). Try to see it my way: Modelling persuasion in legal discourse. *Artificial Intelligence and Law*, *11*(4), 271–287.

Bench-Capon, T. J. M., Atkinson, K., & Chorley, A. (2005). Persuasion and value in legal argument. *Journal of Logic and Computation*, *15*(6), 1075–1097.

Bhatt, K., Argamon, S., & Evens, M. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Chicago.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transaction of Human-Computer Interaction*, *12*(2), 293–327.

Blum, A., & Furst, M. (1995). Fast planning through planning graph analysis. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, (pp. 1636–1642).

Brehm, J., & Cohen, A. (1962). *Explorations in cognitive dissonance*. New York: John Wiley & Sons Inc.

Breivik, E., & Supphellen, M. (2003). Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques. *Journal of Economic Psychology*, *24*(1), 77–98.

Brennan, S. E., & Ohaeri, J. O. (1994). Effects of message style on users' attributions toward agents. In *CHI '94: Conference companion on Human factors in computing systems*, (pp. 281–282). New York, NY, USA: ACM.

Carenini, G., & Moore, J. (2000a). A strategy for generating evaluative arguments. In *Proceedings of the International Conference on Natural Language Generation*, (pp. 47–54).

Carenini, G., & Moore, J. D. (2000b). An empirical study of the influence of argument conciseness on argument effectiveness. In H. Iida (Ed.) *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, (pp. 150–157). Hong Kong.

Carenini, G., & Moore, J. D. (2001). An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *IJCAI*, (pp. 1307–1314).

Carlson, R. (1996). The dialog component in the waxholm system. In Veldhuijzen, S. Luperfoy, & A. Nijholt (Eds.) *Proceedings of the twente workshop on language technology. Dialogue management in natural language systems*, (pp. 209–218).

Carofiglio, V., & de Rosis, F. d. (2003). Combining logical with emotional reasoning in natural argumentation. In C. Conati, E. Hudlika, & C. Lisetti (Eds.) *proceedings of the 9th International Conference in User Modelling; Workshop on Affect*. Pittsburgh.

Cassell, J., & Bickmore, T. W. (2002). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and Adaptive Interfaces*, *13*(1-2), 89–132.

Cohen, W. W. (2004). Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. web page `http://minorthird.sourceforge.net`.

Colby, K. M. (1975). *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. New York, NY, USA: Elsevier Science Inc.

Corston-Oliver, S. H. (1998a). Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. Tech. rep., Microsoft Research.

Corston-Oliver, S. H. (1998b). *Computing representations of the structure of written discourse*. Ph.D. thesis, University of California, Santa Barbara.

Crosswhite, J. (2000). Rhetoric and computation. In Reed & Norman (2000).

Dahlback, N., Jonsson, A., & Ahrenberg, L. (1993). Wizard of oz-studies – why and how. In *Workshop on Intelligent User Interfaces*. Orlando, FL.

De Boni, M., Richardson, A., & Hurling, R. (2008). Humour, relationship maintenance and personality matching in automated dialogue: A controlled study. *Interacting with Computers*, *20*(3), 342–353.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*(2), 321–357.

Edwards, W. (1994). Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes*, *60*(3), 306–325.

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, (pp. 28–36). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. California, USA: Stanford University Press.

Fogg, B. J. (2003). *Persuasive technology: using computers to change what we think and do*. Interactive Technologies. Morgan Kaufmann, San Francisco, CA.

Fox, M., & Long, D. (1995). Hierarchical planning using abstraction. *Control Theory and Applications, IEE Proceedings*, *142*(3), 197–210.

Galley, M., Mckeown, K., Hirschberg, J., & Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics.

Garssen, B. (2001). Argument schemes. In F. H. van Eemeren (Ed.) *Crucial concepts in argumentation theory*, chap. 4, (pp. 81–99). Amsterdam University Press.

Gilbert, M. A., Grasso, F., Groarke, L., Gurr, C., & Gerlofs, J. M. (2003). The persuasion machine. In Norman & Reed (2003), chap. 5.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Grasso, F. (2002). Towards a framework for rhetorical argumentation. In J. Bos, M. Foster, & C. Matheson (Eds.) *EDILOG 02: Proceedings of the 6th workshop on the semantics and pragmatics of dialogue*, (pp. 53–60). Edinburgh, UK.

Grasso, F. (2003). Rhetorical coding of health promotion dialogues. In E. Barahona (Ed.) *Artificial Intelligence in Medicine - 9th Conference on Artificial Intelligence in Medicine in Europe, AIME*, vol. 2780 of *LNAI*, (pp. 179–188). Springer-Verlag.

Grasso, F., Cawsey, A., & Jones, R. (2000). Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies, 53*(6), 1077–1115.

Greenwood, K., Bench-Capon, T. J. M., & Mcburney, P. (2003). Towards a computational account of persuasion in law. In *The Ninth International Conference on Artificial Intelligence and Law (ICAIL-2003).*.

Guerini, M., & Stock, O. (2005). Toward ethical persuasive agents. In *Proceedings of the International Joint Conference of Artificial Itelligence Workshop on Computational Models of Natural Argument*.

Guerini, M., Stock, O., & Zancanaro, M. (2003). Persuasion models for intelligent interfaces. In C. Reed, F. Grasso, & G. Carenini (Eds.) *Proceedings of the International Joint Conference on Artificial Intelligence Workshop on Computational Models of Natural Argument*. Acapulco, Mexico.

Guerini, M., Stock, O., & Zancanaro, M. (2004). Persuasive strategies and rhetorical relation selection. In *Proceedings of the European Conference on Artificial Intelligence Workshop on Computational Models of Natural Argument*. Valencia, Spain.

Hastings, A. C. (1963). *A Reformulation of the Modes of Reasoning in Argumentation*. Ph.D. thesis, Evanston, Illinois.

Henkemans, A. F. S. (2001). Argumentation structure. In F. H. van Eemeren (Ed.) *Crucial concepts in argumentation theory*, chap. 5, (pp. 101–134). Amsterdam University Press.

Hillard, D., Ostendorf, M., & Shriberg, E. (2003). Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proceedings of the 2003*

Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, (pp. 34–36). Morristown, NJ, USA: Association for Computational Linguistics.

Hulstijn, H., Steetskamp, R., Doest, H., Burgt, S., & Nijholt, A. (1996). Topics in schisma dialogues. In *Proceedings of the Twente Workshop on Language Technology: Dialogue Management in Natural Language Systems*, (pp. 89–99). The Netherlands: Enschede.

Hunter, J., & Boster, F. (1978). An empathy model of compliance-gaining message strategy selection. Paper presented at the annual meeting of the Speech Communication Association.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.*, vol. 1, (pp. 364–367).

Johnson, M. G. (2003). *The Ultimate Desert Handbook: A Complete Manual for Desert Hikers, Campers and Travelers*. McGraw-Hill Contemporary.

Katzav, J., & Reed, C. A. (2004). On argumentation schemes and the natural classification of arguments. *Argumentation*, *18*(2), 239–259.

Katzav, J., Reed, C. A., & Rowe, G. W. A. (2004). Argument research corpus. In B. Lewandowska-Tomaszczyk (Ed.) *Proceedings of the 2003 Conference*, Practical Applications in Language and Computers, (pp. 229–239). Frankfurt: Peter Lang.

Kendall, M., & Gibbons, J. D. (1990). *Rank Correlation Methods*. A Charles Griffin Title, fifth ed.

Lafferty, J. C., & Eady, P. M. (1974). *The desert survival problem*. Plymouth, Michigan: Experimental Learning Methods.

Lavoie, B., & Rambow, O. (1997). A fast and portable realizer for text generation systems. In *Proceedings of the fifth conference on Applied natural language processing*, (pp. 265–268). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Lemon, O., Cavedon, L., & Kelly, B. (2003). Managing dialogue interaction: A multi-layered approach. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, *10*, 707–710.

Levy, D., Catizone, R., Battacharia, B., Krotov, A., & Wilks, Y. (1997). Converse:a conversational companion. In *Proceedings of 1st International Workshop on Human-Computer Conversation*. Bellagio, Italy.

Lewin, I., Rupp, C. J., Hieronymus, J., Milward, D., Larsson, S., & Berman, A. (2000). Siridus system architecture and interface report (baseline). Tech. rep., Siridus project.

Lisowska, A., Rajman, M., & Bui, T. H. (2005). *ARCHIVUS: A System for Accessing the Content of Recorded Multimodal Meetings*, vol. 3361/2005 of *Lecture Notes in Computer Science*, (pp. 291–304). Berlin / Heidelberg: Springer.

Mairesse, F., & Walker, M. (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*. Prague.

Mann, W. C. (1999). An introduction to rhetorical structure theory (RST). web page `http://www.sil.org/~mannb/rst/rintro99.htm`.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, *26*(3), 395–448.

Marshall, C. C. (1989). Representing the structure of a legal argument. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Law*, (pp. 121–127). New York, NY, USA: ACM Press.

Marwell, G., & Schmitt, D. (1967). Dimensions of compliance-gaining behavior: An empirical analysis. *sociomety*, *30*, 350–364.

Mazzotta, I., de Rosis, F., & Carofiglio, V. (2007). Portia: A user-adapted persuasion system in the healthy-eating domain. *Intelligent Systems, IEEE*, *22*(6), 42–51.

Miller, G. (1980). On being persuaded: Some basic distinctions. In M. E. Roloff, & G. R. Miller (Eds.) *Persuasion: New directions in theory and research*, (pp. 11–28). SAGE Publications.

Moon, Y. (1998). The effects of distance in local versus remote human-computer interaction. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, (pp. 103–108). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.

Moschitti, A. (2006). Making tree kernels practical for natural language learning. In *11th conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: The Association for Computer Linguistics.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

Norman, T. J., & Reed, C. (2003). *Argumentation Machines : New Frontiers in Argument and Computation (Argumentation Library)*. Springer.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. *Journal of abnormal and social psychology*, *66*, 574–583.

Perelman, C., & Olbrechts-Tyteca, L. (1958). *Traité de l'argumentation: La nouvelle rhétorique*. Paris, France: Presses Universitaires de France.

Pervin, L. A., & John, O. P. (Eds.) (2001). *Handbook of Personality: Theory and Research, Second Edition*. The Guilford Press.

Prochaska, J. O., & Diclemente, C. (1992). Stages of change in the modification of problem behavior. *Progress in Behavior Modification*, *28*, 183–218.

Quarteroni, S. A., & Manandhar, S. (2008). Designing an interactive open domain question answering system. *Journal of Natural Language Engineering JNLE, Special issue on Interactive Question Answering (to appear)*.

Reed, C. (1998). *Generating Arguments in Natural Language*. Ph.D. thesis, University College London.

Reed, C., & Grasso, F. (2001). Computational models of natural language argument. In V. N. Alexandrov, J. Dongarra, B. A. Juliano, R. S. Renner, & Chih (Eds.) *Proceedings of the International Conference on Computational Sciences-Part I.*, vol. 2073 of *Lecture Notes in Computer Science*, (pp. 999–1008). London, UK: Springer-Verlag.

Reed, C., & Norman, T. (2000). Symposium on argument and computation: position papers.

Reeves, B., & Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(01), 57–87.

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: generating tailored smoking cessation letters. *Artificial Intelligence*, *144*(1-2), 41–58.

Rokeach, M. (1968). *Beliefs, Attitudes and Values: A Theory of Organization and Change*. San Francisco, CA, USA: Jossey-Bass, Inc.

Sandell, R. G. (1976). *Linguistic Style and Persuasion (European monographs in social psychology)*. Academic Press, Inc.

Schiffrin, D. (1988). *Discourse Markers (Studies in Interactional Sociolinguistics)*. Cambridge University Press.

Schmitt, C., Dengler, D., & Bauer, M. (2003). Multivariate preference models and decision making with the maut machine. In *User Modeling 2003*, (p. 148). Springer Berlin / Heidelberg.

Shechtman, N., & Horowitz, L. M. (2003). Media inequality in conversation: how people behave differently when interacting with computers and people. In G. Cockton, & P. Korhonen (Eds.) *Proceedings of the SIGCHI conference on Human factors in computing systems*, (pp. 281–288). Ft. Lauderdale, Florida, USA: ACM Press.

Sillars, A. (1980). The stranger and the spouse as target persons for compliance-gaining strategies: a subjective expected utility model. *Human Communication Research*, *6*, 265–279.

Smith, M., Garigliano, R., & Morgan, R. (1994). Generation in the lolita system: An engineering approach. In D. Mcdonald, & M. Meteer (Eds.) *7th Int. Workshop on Natural Language Generation*, (pp. 241–244).

Smith, R. W. (1992). Integration of domain problem solving with natural language dialog: the missing axiom theory. vol. 1707, (pp. 270–278). SPIE.

Soricut, R., & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*.

Stent, A. (2002). A conversation acts model for generating spoken dialogue contributions. *Computer Speech & Language*, *16*(3-4), 313–352.

Stiff, J. B., & Mongeau, P. A. (2002). *Persuasive Communication*. The Guilford Press, second ed.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press, updated ed.

van Mulken, S., André, E., & Müller, J. (1999). An empirical study on the trustworthiness of life-like interface agents. In *Proceedings of the 8th International Conference on Human-Computer Interaction*, (pp. 152–156). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer.

Verheij, B. (1999). Automated argument assistance for lawyers. In *The International Conference on Artificial Intelligence and Law*, (pp. 43–52).

Vrajitoru, D. (2003). Evolutionary sentence building for chatterbots. In *Proceedings of the Genetic and Evolutionary Computation Conference; Late Breaking Papers*, (pp. 315–321).

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). Paradise: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, (pp. 271–280). Morristown, NJ, USA: Association for Computational Linguistics.

Walker, M. A., Whittaker, S. J., Stent, A., Maloor, P., Moore, J., Johnston, M., & Vasireddy, G. (2004). Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science: A Multidisciplinary Journal*, *28*(5), 811–840.

Wallace, R. (2004). *The elements of AIML style.*. the A.L.I.C.E. Artificial Intelligence Foundation.

Walton, D. (1996). *Argument Structure: A Pragmatic Theory (Toronto Studies in Philosophy)*. University of Toronto Press.

Walton, D., & Reed, C. (2002). Argumentation schemes and defeasible inferences. In G. Carenini, F. Grasso, & C. Reed (Eds.) *Working Notes of the ECAI'2002 Workshop on Computational Models of Natural Argument*, (pp. 45–55).

Weizenbaum, J. (1966). Eliza-a computer program for the study of natural language communication between man and machine. *Commununications of the ACM*, *9*(1), 36–45.

Wiseman, R., & Schenck-Hamlin, W. (1981). A multidimensional scaling validation of an inductively-derived set of compliance-gaining strategies. *Communication Monographs*, *59*, 329–349.

Zinn, C., Moore, J. D., & Core, M. G. (2002). A 3-tier planning architecture for managing tutorial dialogue. In *ITS '02: Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, (pp. 574–584). London, UK: Springer-Verlag.

Index

— Q —

— R —

Citation Index