Multimodal interface design for multimedia meeting content retrieval

Lisowska, Agnès

Multimodal Interface Design for Multimedia Meeting Content Retrieval

PhD Thesis

Agnes Lisowska

N. thèse ETI 03

Thesis Co-directed by:
Susan Armstrong, ISSCO/TIM, ETI, University of Geneva
Mireille Betrancourt, TECFA, University of Geneva

Thesis Jury Members:
Barbara Moser, Interpretation Department, ETI, University of Geneva (President)
Martin Rajman, CGC, Ecole Polytechnique Fédéral de Lausanne
Sharon Oviatt, CHCC, Oregon Health and Science University

Geneva, Switzerland
September, 2007

This thesis is dedicated to my parents,
who taught me to ask questions,
inspired me to explore and
have always been there for me.

Thank you

Abstract

Multimodal Interface Design for Multimodal Meeting Content Retrieval

The goal of this thesis is to assess whether multimodal input brings added value to interaction for the multimedia meeting browsing and retrieval domain, and if it does, what the nature of that interaction is. In our work we define 'added value' in terms of increased efficiency when compared to standard mouse and keyboard input, the usefulness of multiple modalities, and overall subjective user satisfaction when interacting multimodally in an interface.

In particular, we are interested in the benefits and drawbacks that novel input modalities such as voice and pen bring to interaction, especially in the presence of more familiar modalities such as the mouse and keyboard. Our work focuses on six central questions: 1) how often are different modalities used, alone and in combination, for meeting browsing and retrieval tasks, 2) do certain modalities or modality combinations lead to an increase in efficiency, 3) does modality use change when a user encounters a problem during interaction, 4) how do users perceive different modalities, 5) does learning to use a system with a particular set of modalities influence how those modalities are used when other modalities also become available and 6) are some modalities more suited to finding certain types of information than others?

We answer these questions through the analysis of results from a large-scale user-centered study we conducted using Archivus, a multimodal system for multimedia meeting browsing and retrieval, which was specifically developed for this type of research. We also discuss the development of the Archivus system itself, as well as the difficulties encountered when designing an experimental protocol for the types of experiments necessary to answer the above questions, and the solutions we found and adopted.

Résumé

La Conception d'Interfaces Multimodales pour l'Accès aux Enregistrements de Réunions (Multimédia)

Le but de cette recherche est d'évaluer si l'accès multimodal à une interface apporte une « valeur ajoutée » à l'interaction avec les systèmes qui permettent de faire des recherches et de naviguer dans les enregistrements de réunion et de stocker les résultats dans un format multimédia. Si c'est le cas, nous cherchons à découvrir quelles sont les caractéristiques de cette interaction. Dans notre travail, nous définissons le terme « valeur ajoutée » en termes d'accroissement de l'efficacité comparée à une interaction avec souris et clavier, d'utilité de disposer de multiples modalités, et de satisfaction générale des utilisateurs utilisant une interface d'une manière multimodale.

En particulier, nous nous intéressons aux effets sur l'interaction de nouvelles modalités comme l'interaction à l'aide de voix ou l'utilisation d'un stylo électronique, particulièrement lorsqu'elles sont ajoutées à l'utilisation de la souris et du clavier. Notre travail se concentre sur six questions principales: 1) à quelle fréquence les modalités sont-elles utilisées, seule ou en combinaison, pour la navigation et la recherche dans les réunions? 2) Certaines modalités ou combinaisons de modalités permettent-elles une amélioration de performance? 3) L'utilisation des modalités varie-t-elle quand l'utilisateur se retrouve dans une situation problématique pendant l'interaction? 4) Comment les utilisateurs perçoivent-ils les différentes modalités? 5) Le fait d'apprendre à utiliser une système avec une modalité en particulier influence-t-il la manière d'utiliser cette modalité quand d'autres modalités deviennent aussi disponibles dans un second temps? et 6) Certaines modalités sont-elles plus adaptées pour trouver certains types d'information plutôt que d'autres ?

Pour répondre à ces questions nous avons mené une étude expérimentale qui nous a permis d'observer de façon systématique l'interaction de 80 utilisateurs avec une interface multimodale pour le stockage et la recherche d'enregistrements de réunions et des documents multimédia afférentes, Archivus. Ce logiciel a été conçu spécifiquement pour la recherche sur l'interaction multimodale. Nos résultats apportent également des éléments de discussion sur la conception du système Archivus lui-même ainsi que sur les difficultés rencontrées dans l'élaboration du protocole d'expérience et des solution adoptées.

Acknowledgements

.

I would like to express my heartfelt gratitude to…

… Susan Armstrong for getting me involved in the IM2 project, making it possible to pursue a thesis in a research area that I was passionate about, for giving me the freedom and flexibility to explore new ideas, for all the helpful comments and suggestions, for always being supportive of my work, and for facilitating the many opportunities that I had to enrich my knowledge.

… Mireille Betrancourt for suggesting new avenues of research, advice on running statistically significant experiments, and for the insightful comments.

… Martin Rajman, Mirek Melichar, Marita Ailomaa and the rest of the Archivus development team for their collaboration in bringing the Archivus system to life, for the long scientific discussions, and for making the hours of experiments much more fun.

… my colleagues, past and present, both at ISSCO and in the IM2 project for creating a fruitful work environment, teaching me about the scientific community and the exchange of ideas, and taking an interest in my work.

… the anonymous reviewers of the papers published on this work for their valuable comments and suggestions.

… all of those who volunteered as users for our experiments with Archivus.

… my friends and loved ones on both sides of the Atlantic, but in particular to Olivier, Marcin, Michele, Ania T., Carrie, Paula, the two Mariannes, Marita and Bruno for helping to keep me motivated, for always being there to listen and advise, and most importantly for making sure that I found the time to have fun once in a while.

… my parents for their love and encouragement, for teaching me to always do my best, to not be afraid to take chances and for supporting my decisions - even if they involved my moving to another continent.

This thesis would not be what it is without all of you.

Abstract
Resume
Acknowledgements
1. Introduction 15
2. State of the Art 21
2.1 Human-computer interaction 21
2.2 Input devices 24
2.3 GUI interfaces 26
2.3.1 Advantages 26
2.3.2 Disadvantages 26
2.3.3 Conclusions 27
2.4 Natural language interfaces 28
2.4.1 Advantages 29
2.4.2 Disadvantages 30
2.4.3 Dialogue types and system architectures 32
2.4.4 Novice vs. expert users and system control 33
2.4.5 Development and evaluation of NL systems 34
2.4.6 Conclusions 35
2.5 Multimodal interfaces 36
2.5.1 Multimedia vs. multimodal 36
2.5.2 Advantages 37
2.5.3 Disadvantages 38
2.5.4 Multimodal architectures 39
2.5.5 Existing systems 40
2.5.6 The past, present and future of multimodal interfaces 42
2.6 Natural interactivity 43
2.7 Modality theory 44
2.8 User modelling and sociological considerations 45
2.9 Multimodal meeting domain 47
2.9.1 Existing projects 47
2.9.1.1 NIST Meeting Recognition Project and the SmartSpace Laboratory 47
2.9.1.2 ICSI meeting corpus 48
2.9.1.3 M4 48
2.9.1.4 Interactive Multimodal Information Management (IM2) 48
2.9.1.5 Augmented Multiparty Interaction (AMI) 48
2.9.1.6 Carnegie Mellon University Meeting Room Project 49
2.9.2 Existing browsers for the multimodal meeting domain 49

2.10 Discussion 50
3. Research Goals 52
3.1 Use of modalities 53
3.2 Modalities and task types 54
3.3 Task completion 54
3.4 Problems and modality choice 55
3.5 User's perception of modalities 56
3.6 Learning effect 56
4. The Archivus System 58
4.1 Intended users and use cases 58
4.1.1 Range of users 58
4.1.2 Scenarios of use 59
4.2 User requirements study 61
4.3 Archivus backend database 62
4.3.1 Controlling for variables in the data 62
4.3.2 Recording scenarios 63
4.3.3 The data set 64
4.4 Design rationale 65
4.4.1 Modality choice 67
4.4.2 Flexible multimodality 67
4.4.3 Rapid Dialogue Prototyping Methodology (RDPM) 68
4.4.4 Archivus metaphor 68
4.4.5 Graphical components and layout 69
4.4.6 System feedback 70
4.5 Description of the system 72
4.5.1 What can be done 72
4.5.2 How it can be done 73
4.6 Conclusions 76
5. The Experiments 77
5.1 Wizard of Oz methodology 77
5.1.1 Extending the Wizard of Oz methodology for multimodality and Archivus 79
5.2 Archivus Wizard of Oz environment 80
5.2.1 Pilot experiment environment 80
5.2.3 Final experiment environment 81
5.3 Modality combinations 82
5.4 Experimental protocol 82
5.4.1 Choosing a protocol 84

5.4.2 The pilot experiment 85
5.4.3 The challenge of an unbiased tutorial 86
5.4.4 The evaluation protocol with a tutorial 87
5.4.5 Overview of experiment documents 87
5.4.6 Questions used 88
5.5 The users 90
5.6 Types of data gathered 91
5.7 Conclusions 92
6. Experiment Results 93
6.1 Introductory comments 93
6.2 Subjective user opinion of interaction with Archivus 96
6.2.1 General impressions 96
6.2.2 Perceived usefulness of modalities 97
6.2.3 Conclusions 98
6.3 Learning effects 99
6.3.1 Conclusions 100
6.4 Proportions of modality use 101
6.4.1 Conclusions 104
6.5 Evolution of modality use 104
6.5.1 Evolution of a single modality 104
6.5.2 Evolution within a condition 108
6.5.3 Conclusions 117
6.6 Modality switching 118
6.6.1 Error production 119
6.6.2 Nature of the errors 119
6.6.3 Proportion of errors 120
6.6.4 Proportion of modality switches 122
6.6.5 Nature of modality switches 123
6.6.6 Conclusions 125
6.7 Functional equivalence 125
6.7.1 Mouse vs. pen 126
6.7.2 Voice vs. keyboard 127
6.7.3 Conclusions 128
6.8 Task completion 129
6.8.1 Number of questions completed 129
6.8.2 Success scores 132
6.8.3 Correctness scores 136

6.8.4 Distribution of success among modalities 137
6.8.5 Influence of user background 139
6.8.6 Conclusions 141
6.9 Modality use and task type 142
6.10 Conclusions 144
7. Conclusions 145
7.1 Use of modalities 146
7.2 Task completion 147
7.3 Problems and modality choice 147
7.4 User perception of modalities 148
7.5 Learning effect 149
7.6 Conclusions 149
8. Future Work and Possible Extensions 154
8.1 Further analysis on existing data 154
8.2 Analysis requiring a change in experiment design 155
8.3 Analysis requiring a change in the Archivus system 158
8.4 Concluding remarks 159
Bibliography 160
Appendices 170
Appendix A: User requirements questionnaire 170
Appendix B: Archivus experiment consent form 172
Appendix C: Archivus evaluation description 173
Appendix D: Archivus tutorial examples – P, V and MVK conditions 174
Appendix D1: The P condition 174
Appendix D2: The V condition 177
Appendix D3: The MVK condition 180
Appendix E: Archivus experiment questionnaires 183
Appendix E1: Archivus pre-experiment (demographic) questionnaire 183
Appendix E2: Archivus post-experiment questionnaire 186
Appendix F: Example of question sheet from the Archivus pilot experiments 194
Appendix G: Questions from the Archivus final experiment 197
Appendix H Classification of Archivus task questions 199
List of Figures 201
List of Tables 203

1. Introduction

The problem we intend to investigate in the scope of this thesis is which modalities and in which combination are best suited for use in a multimodal interface that allows users to retrieve the content of recorded and annotated multimodal meetings. This problem involves the fusion of two emerging fields; multimodal meeting recording and storage, and multimodal interfaces.

The storage and processing of meetings is becoming a popular research area as businesses are realizing the benefits of storing such information. Several projects[1] involve, or have involved in the past, the collection of multimedia meeting data. In several of these, special SmartRooms have been designed in which meetings are recorded in such a way that the data can be easily synchronized, processed and stored. For example, in the IM2 project (in which the work presented here is grounded) meetings are recorded at the IDIAP SmartRoom [1] and the resulting data is stored in databases which contain video and audio tracks from a meeting, a text transcription of the meeting, as well as various levels of annotation, including linguistic (dialogue acts, topic segments, keywords) and meta-levels (meeting actions). Additionally, the meeting data contains electronic versions of all documents used in the meetings, copies of all notes taken by meeting participants, and what was written on the electronic whiteboard available in the room.

However, while research is being done into how the meetings are to be processed and stored, little has been said about how users interested in the content of those recordings should be able to access that information. Standard database access techniques such as SQL queries remain an option, but as technology improves, computer users are becoming increasingly demanding of the capabilities of the systems they use and such interfaces are likely to prove insufficient to meet their demands, particularly as the information stored in recorded multimodal meetings is richer and more complex than the types of information stored in conventional databases. Consequently, innovative interface design is necessary.

[1] The *IM2* project http://www.im2.ch, the *AMI* project www.amiproject.org, *The Meeting Room Project* at Carnegie Mellon University, http://www.is.cs.cmu.edu/mie/, *Rich transcription of natural and impromptu meetings* at ICSI, Berkeley, http://www.icsi.berkeley.edu/icsi-ro.html

One of the central questions that needs to be addressed before design can take place is how a real-world user such as an employee of a company where SmartRoom meeting data has been recorded can best exploit the data generated. Tucker and Whittaker [2] provide a good overview of the types of meeting browsers that have been developed in various projects and suggest a 4-category taxonomy for meeting browsers – audio, video, artifact and discourse (this topic is discussed in more detail in section 2.9.2). However, it appears that most, if not all, browsers that are described for the meeting domain rely on standard mouse and keyboard input. Little has been said about the possible benefits of incorporating multimodal input to a meeting browsing and retrieval system.

Due to the proliferation of window-based platforms, technologies such as the internet and the commonality of input devices such as the mouse and keyboard, certain interaction paradigms seem to have asserted themselves in western computer culture. For example, the use of a mouse for direct manipulation of graphical objects on the screen (point-and-click browsing) or the use of the keyboard for targeted web-searching. Moreover, if and when language is used to seek out information, it is often via 'intelligent' keyword-driven searches, in which obtaining desired results quickly and efficiently requires a certain degree of skill and knowledge on the part of the user.

While we agree that in most office-type environments and for almost all office-type applications the mouse-keyboard paradigm will be strongly preferred and users will be reluctant to stray from it, we believe that meeting browsing and retrieval such as outlined above is a sufficiently new and different domain of interaction that users can be encouraged to try out and consistently use novel input modalities such as voice (including more complex natural language interaction), touchscreen or pen input. The difference in the multimedia meeting browsing and retrieval domain is not found at the level of the actual media artifacts that are stored in the database. The web contains examples of the same types of media (video, audio and text files) and users are perfectly content to use the mouse and keyboard to access them. Rather, the distinction is made at the underlying level – in the direct, though not necessarily explicit, relationships between the information contained *across* that media, and the elements of that information that a person would want to access. Our assumption is that the results of the fuzzy underlying difference can best be exploited by providing the user with a multimodal interface.

Recent advances in various areas of technology such as voice and gesture recognition and language processing are propagating the trend of designing multimodal interfaces, interfaces that incorporate more than the standard modalities of keyboard and mouse input and text, sound and graphical output. Multimodal interfaces offer the user increased

flexibility in the ways in which they can interact with a computer, which in turn leads to smoother and more natural communication, and in many cases a higher level of satisfaction for the user. However, because multimodal interface design is a new field, there are no concrete methodologies [3] or ascertained truths that designers can rely on [4], and most of the potential technologies to be incorporated into an interface have only been investigated in a pair-wise (rather than a truly multimodal) manner for a subset of very specific tasks.

To our knowledge, no multimodal interfaces have been designed for the multimedia meeting browsing and retrieval domain. Several authors [5-7], have argued that determining which modalities are useful for a given interface will depend highly on the design and functionalities of the interface itself, on its context of use, and on the domain of the application. In this thesis, we aim to investigate which modalities users prefer to use, alone or in combination, to retrieve and browse multimedia meeting data as recorded, processed and stored within the IM2 project (as described above). This work entails both practical and theoretical aspects, outlined in the following sections.

As has already been mentioned, no multimodal interfaces exist for the multimedia meeting browsing and retrieval domain, and more importantly, none existed within the IM2 project at the outset of 2000. Therefore, along with colleagues at the Artificial Intelligence Laboratory at the Ecole Polytechnique Fédéral de Lausanne, we decided to design and develop just such a system, which we called Archivus [8], and which is described in greater detail in Chapter 4.

While there exist many traditional software development methodologies, and in particular the software development lifecycle as described in Dix et al. [9] to which we tried to adhere are much as possible, working on developing an interface for an entirely new domain proved more challenging than first expected. It is generally accepted in the field of interface design that to design a successful system, one must have a well defined set of user requirements, a clear vision of the context in which the system will be used (both the tasks that will be performed using the system, and the actual environment in which it will be used), and a good idea of who will be using the system. However, with a completely new interface for a new domain, and in particular one that is being developed in a research rather than in an industrial context, meeting these requirements is significantly more difficult [10]. First, there is no defined set of users other than an idealized hypothetical one. Second, the user requirements can only be intuited. There are no users who have had experience with similar systems/interfaces or in accessing the types of data available at the various levels of abstraction possible, so traditional methodologies for

gathering user requirements proved to be insufficient. Finally, the eventual context(s) of use can also only be guessed at since the system is not being designed with any particular 'client' in mind.

Moreover, Dix et al. [9] point out that one cannot determine all system requirements right from the start, that '*the tasks a user will perform are often only known by the user after he is familiar with the system on which he performs them*' but that in order to observe users in this context, an application that simulates the intended final application as closely as possible needs to be used, since even the slightest detail can influence its usability. Otherwise, the results of the observations may not be applicable to the real system once it is developed.

In our case a partial solution to this somewhat circular problem was that since the Archivus system was meant to be designed specifically for the data generated within the IM2 project, the overall goals and milestones of the project imposed restrictions on the design of both the database of multimedia meetings and importantly on what technology can and should be considered in the interface design. The issue of gathering user requirements was handled through an informal questionnaire-based study [11] which we ran ourselves. In this study, participants were given the context of the application, told what types of data would be available, and asked to list the types of information that they thought would be most useful to them if they had access to that data. A simultaneously top-down (from the user requirements) and bottom-up (from the technologies available in the IM2 project) approach was then chosen for the design process in order to achieve an equilibrium that we felt would be satisfactory to the user.

The Archivus system lets users access a database of recorded and annotated multimedia meetings using any one (or a combination) of four possible input modalities: voice (in freeform natural language), keyboard, mouse or pen (used as a pointing device). In order to investigate specific user preferences for modality use without influencing their choices, it was necessary to impose as few *a priori* assumptions as possible about which modalities the user would or should use at any given point. As a result, the Archivus interface in its current implementation is *flexibly multimodal*. This means that it allows users to perform any action using any of the possible input modalities at any time. The user is free to choose the modality that they feel is the most suitable or comfortable for any particular action that they want to execute.

It is also important to note that in fact, the Archivus system is not a fully functioning system, but rather a high-fidelity prototype. The motivation behind this choice, in

addition to the time required to fully implement such a system, is related to the fact that two other PhD students with goals different to those addressed in this thesis were also using the system and the experiments described in this work for their own purposes. One of the students was investigating the dialogue strategies that users adopt with the system in the case of natural language use, and the other was looking at the lexico-semantic data that was generated by natural language interaction with the system and the resulting implications for the development of natural language processing modules for a system such as Archivus. Building a high-fidelity prototype satisfied all three sets of goals and allowed us maximum flexibility in gathering data while reducing the time and effort for implementation of the system. In Archivus the speech recognition and language processing modules have not been implemented since we wanted to give as much freedom in language use as possible in order to investigate how language was used, and how voice was used as an input modality if the quality of the language based interaction was quite good (though not perfect).

The experiments used to gather data for this thesis were executed in a Wizard of Oz environment, which is commonly used to acquire data for and evaluate natural language interfaces. In Wizard of Oz experiments users interact with a system that they believe to be fully functional, while in fact certain components (in our case the speech recognition, natural language processing and dialogue management modules) are simulated by a human being in another room. The Wizard of Oz methodology and the extensions that were necessary in order to apply it to the Archivus system are described in detail in Chapter 5.

The theoretical aspect of this thesis is grounded first in the design of the experiments and then in the analysis and interpretation of the data gathered during the user-based experiments with the Archivus system. The aim was to determine if in fact multimodal input provides an added value to interaction for the multimedia meeting browsing and retrieval domain, and if it does, what the nature of that interaction is. It is important to note here that we define *added value* in terms of increased performance when compared to standard mouse and keyboard input, the usefulness of multiple modalities, and overall subjective user satisfaction when interacting multimodally in an interface. In particular, we are interested in the benefits and drawbacks that novel input modalities such as voice and pen bring to interaction, especially in the presence of more familiar modalities such as the mouse and keyboard.

Our work focuses on 6 central questions:

1. How often are different modalities used, alone and in combination, for meeting browsing and retrieval tasks?
2. Are some modalities more suited to finding certain types of information than others?
3. Do certain modalities or modality combinations make the system easier to learn, leading to an increase in performance in the long term?
4. Does modality use change when a user encounters a problem during interaction?
5. How do users perceive different modalities?
6. Does learning to use a system with a particular set of modalities influence how those modalities are used when other modalities also become available?

The remainder of this thesis will be laid out as follows. We will begin with an overview of the state of the art in Chapter 2. Chapter 3 will outline our research goals in more detail. Chapters 4 and 5 will give details of the Archivus system and the experimental methodology that we chose to use, respectively. Chapter 6 will describe the experiments themselves, while in Chapter 7 we discuss the implications of the results. We conclude by highlighting possible extensions to the work in Chapter 8.

2. State of the Art

Work on multimodal interfaces is a relatively new field, and as such, there are few general guidelines for how to go about creating and evaluating multimodal systems such as the one described in this thesis. Consequently, this chapter will begin by focusing on the general aspects of the more established fields of human-computer interaction, graphical user interfaces, input devices and natural language interfaces as a foundation, and will then move on to discuss existing work in multimodal interfaces, preliminary work on modality theory, and various psycho-sociological factors that we feel are relevant to the design and development of multimodal interfaces.

2.1 Human-computer interaction

Human computer interaction (HCI) is a vast and well established field that integrates knowledge about technological limitations and theories from cognitive science about the processing capabilities of human beings. Its goals are to enhance existing interaction capabilities between humans and computers, to create new interaction paradigms, and to develop concrete and standardized design principles for the devices and applications used in human-computer communication. The design principles and guidelines that have been developed can be grouped into three general categories - 1) those that pertain to requirements gathering and design processes, 2) those that help ensure the usability of the system, and 3) those that deal with evaluating a system. These three categories are discussed in the sections below, but only at the general level and in relation to the work in this thesis. Detailed information can be found in HCI textbooks such as *Human Computer Interactio*n [9] and *The Human-Computer Interaction Handbook* [12].

User requirements gathering and design processes

Various methods have been developed to gather user requirements for system design. These include questionnaires, interviews, task analysis and task modelling. The choice of an appropriate method depends on the tool or software being designed, and the situation and resources available to the designers. Similarly, there is no one established design process that can meet the design needs of all types of systems [13]. Design processes can and do vary, in particular as concerns the point at which end-users are involved. In some cases, the end-users are included right from the conception of the system and are active participants/partners throughout the development cycle [10]. In others, they may only be involved in the requirements gathering and final evaluation stages.

Ensuring usability

There are several inter-related factors that contribute to the overall usability of a system. At the general level there are what Dix et al. [9] call *usability principles* which include:

1. learnability – how easy the system is to learn, which is often measured by how quickly a user can begin to effectively use the system. This concept is broken down further into the sub-principles of predictability (how easily a user can predict the effect of a new action based on their experience with previous actions), synthesizability (how the user can know what affect their past actions have had and how that has manifested itself in the current state of the system), familiarity (how much the user's existing knowledge and experience with the world around them can help them use the new system), generalizability (how easily a user's existing knowledge can be carried across between different applications), and consistency (the likelihood that a given behaviour will be the same given the same situation).

2. flexibility – the variety of different ways in which the user and the system can communicate. This can for example take the form of different levels of dialogue initiative which allow the user differing degrees of control over the system, or input/out flexibility, which allows the user to use different devices to communicate with the system.

3. robustness - how the system helps the user to determine whether their goals have been achieved. Here, the notions of observability (how easily the user can see the effects of their actions in the system), recoverability (how easily a user can recover if they encounter a problem during their interaction with the system) and responsiveness (how the user perceives the speed at which the system responds to their commands) play the key roles.

In addition to these general rules, there are lower-level factors that are important including the choice of input device, choice of output media, the look and placement of interface elements, and consideration of the interaction strategies that users are likely to adopt.

Performance of hardware in relation to its use by humans for a particular application or task has been studied to a large degree. For example, the comparative performance of various input devices, both in terms of inducing or reducing physical strain on the user, or

for improving their speed and accuracy. When trying to determine the appropriateness of different input devices for a particular system, such factors must be taken into consideration as they can greatly influence the usability and efficiency of a system

Another important step in system design is choosing the appropriate media to express information to the user. The choice of media most often depends on the technical resources available when developing the system, on the types of users that are expected to use it, and on the types of information that the system is trying to convey.

Closely related to the question of appropriate media are the communication mechanisms that are incorporated and how they are used. For example, early research in HCI discovered that providing the user with well timed feedback as to whether the system had received and was processing a command was crucial to ensuring smooth interaction [10]. This meant for example the inclusion of progress bars, or the well known Apple hour-glass, which told the user that the system was working/thinking. Choosing how and when to incorporate feedback can have a significant effect on the user's perception of the system.

The layout of interface elements on the screen is also an important issue. Designers need to blend aesthetic characteristics of the elements (which have been shown to have a subconscious psychological impact on how users perceive a system) with their need and utility in the interface, to find just the right balance to promote usability. Dix et al. [9] for example, cite three different ways of organizing controls and displays on the screen: (1) functional, where elements with related functionalities are kept together, (2) sequential, for cases where the order in which elements are used is more important to the overall interaction, and (3) frequency, where elements are placed together depending on how often they are used.

A large amount of research has also gone into studying the principles of graphical design of interface elements, such as the use of colour, fonts, text size, and the use of icons and other imagery, as well as cognitive factors in user perception. Some of these include studies in the field of vision, the optimal speed and volume at which to perceive sound and light, response times to visual and auditory stimuli, attention and fatigue. All of these factors must be taken into account at various levels and to different degrees when designing HCI systems.

Finally, numerous studies both in the psychological and HCI literature have shown that novice and expert users react differently to problem solving and consequently in their

encounters and interactions with computer systems. Novice users tend to require more guidance and assurance, whereas expert users require quick access to functionalities which they know exist, and tend to pay more attention to the more cosmetic aspects of a system [13]. Providing the system with capabilities to satisfy the needs of both groups of users is an important factor in creating a system that is usable by a larger public.

Evaluation methodologies

Finally, various methodologies have been developed to evaluate HCI systems. These range from expert and heuristic evaluations to different types of end-user evaluations such as questionnaires, walkthroughs, and think-aloud protocols to more specific or targeted types of evaluations such as the Wizard of Oz methodology, which will be discussed in detail in Chapter 5.

It is important to note that many of these design principles have been established for commonly used means of input (such as keyboard, mouse, and joystick) and output (graphics, text and sound). There has been relatively little work on similar principles for more sophisticated modalities such as voice and gesture, and even less on principles that guide the integration of several complex modalities in a single system [14].

2.2 Input devices

> *'The appropriate choice and design of input devices and interaction techniques can help to structure the interface such that there is a more direct match between the users' tasks and the low-level syntax of the individual actions that must be performed to achieve those tasks'*
>
> – HINCKLEY [15]

There exist many different types of input devices on the market today, some more conducive to performing certain types of tasks than others. As Hinckley [15] stresses, it is therefore important to choose the right one(s) in order to maximize the usability of an application. He suggests the following general classification of the various types of input devices: keyboards, mice, trackballs, joysticks, tablets, touchpads, touchscreens, pen input devices and alternative input devices such as software aids, feet for input, head tracking, eye tracking and direct brain interfacing. In addition to these, Dix et al. [9], who propose two categories of input devices - pointing and text entry - add handwriting and speech recognition.

Each type of input device has specific characteristics or properties that give it an advantage or disadvantage over others. For example, Hinckley [15] cites the following as being key properties of pointing devices:

- resolution and accuracy
- sampling rate and latency
- noise, aliasing and non-linearity
- absolute vs. relative positioning
- control-to-display ration
- physical property sensed
- number of dimensions
- direct vs. indirect input
- metrics of effectiveness

Many of these properties describe input devices at a much more technical level than we are concerned with in this thesis, so they will not be discussed further. The characteristics that are of interest to us are more abstract. In particular, we are interested in whether the device is familiar to users, whether it is integrated into the computer (as software or hardware – for example a touchpad on a laptop computer), its degree of accuracy (or robustness in the case of input devices that deal with natural language), whether it is compact, and whether it allows for direct or indirect access. To highlight which of these characteristics apply to the various possible input devices, we created Table 1, below. We did not include the alternative input devices proposed by Hinckley as we felt that they were not appropriate for the domain being addressed in our work.

Pointing Devices					
	Familiar	Integrated	Accurate	Compact	Direct
Mouse	●		●	●	
Trackball			●	●	
Joystick	●		●	●	
Tablet			●		●
Touchpad	●	●		●	
Touchscreen	●				●
Pen			●	●	●
Language Based Devices					
	Familiar	Integrated	Robust	Compact	Direct
Keyboard	●		●		●
Speech recognition		●		●	
Handwriting recognition		●		●	●

Table 1: Characterization of input devices

2.3 GUI interfaces

Interfaces always have been, and always will be, constrained by the technologies that are available at the time at which they are developed. Early graphical interfaces were no exception. Before the emergence of true graphical interfaces there were command-line interfaces, since text was the only type of 'graphics' that technology used in computer monitors was capable of producing. As monitor technology improved, simple graphics, at first using alphanumeric characters, were developed. Eventually, as screen resolution increased and colour became available, increasingly complex images could be produced, which lead to the types of graphical interfaces that we are now familiar with, first developed by Apple Computers. But, such advances also required the design and development of new input technologies, since manipulating graphical images in a screen via keyboard while possible, was extremely cumbersome. To this end, the mouse, along with the direct manipulation interaction paradigm was developed. *Direct manipulation*, a term coined by Ben Schneiderman [9], is the idea of being able to move objects on a screen using an input device such as a mouse and immediately (directly) seeing the effect of the action. This is in fact the interaction paradigm that we are all most familiar with given the widespread use of Microsoft Windows and the Apple operating system.

2.3.1 Advantages

Among the advantages of GUI interfaces are that all relevant objects are visible on the screen so that the user knows what they can and cannot do, there is immediate feedback when an action is made, actions can be undone easily, and most importantly for novice users, only '*syntactically correct*' actions can be made [9, 16], which reduces the chances that the user will provide the system with an erroneous command. Additionally, a consistent '*look and feel*' in a GUI allows users to quickly learn new software by analogy, as does the use of direct manipulation coupled with familiar metaphors, which means that users can use a system without having to know the details of the underlying computational concepts [16].

2.3.2 Disadvantages

While GUI interfaces might be well established because they have been in existence for some time now, they are not without their drawbacks. These include the fact that it takes time and effort to navigate through multi-level menus, that it is difficult to know which objects might be available in a system but are simply not visible on the screen at a given point in time (e.g. hidden in a sub-menu) and that you cannot exploit temporal relationships between objects, and cannot really exploit the context of interaction except at the most basic and hard-coded level [16].

GUIs and the direct manipulation paradigm are now so familiar to users, especially to the generation of younger users who have grown up with an already mature form of the technology, that they might seem particularly natural and easy to use. However, from the perspective of someone who is unfamiliar with the technologies, or has a handicap that does not allow for their full exploitation, they are in fact not so easy to use, and by no means are they natural. The concept of manipulating non-tangible objects in a two dimensional space is not obvious to a lot of people. Once a word is written on a piece of paper, that same word, in the physical sense, cannot be moved elsewhere, or altered, without changing its physical manifestation. But, word processing systems allow you to do just that. You can easily cut and paste a piece of text, seeming to physically change the location of those same letters that you typed. This is a concept, and a skill, that has to be learned, much like many of the basic functionalities of standard GUI interfaces.

The desktop metaphor, which has been around for some time now, was meant to help solve this problem, but it too is not always completely transparent and even the common input devices such as the mouse and keyboard are not 'natural' to use. One might think that the keyboard, due to its enormous similarity to the typewriter, would be easy to learn to use. In a sense it is, but in fact, computer keyboards are much more powerful. Many keys and key combinations (hot-keys) can be used to perform otherwise hidden actions on an object, actions that physically alter its form and structure, which is something that a typewriter could not do to a page – other than to transfer text onto it. Mice are also more complicated to use than most people realize because they require a specific type of hand-eye coordination and refinement of movements that takes time and practice to acquire. Of course, once these skills are acquired, they make the use of standard GUIs simple since they can be applied to many current software interfaces, but until they are acquired, the first steps of computer use can be clumsy and awkward.

2.3.3 Conclusions

In an early article, Bill Buxton [17] strongly suggested that HCI designers need to look beyond the direct manipulation interface paradigm into other, perhaps more effective interaction modes, believing that the direct manipulation paradigm should be seen as a '*point of departure*' in interaction research, rather than a '*point of arrival*'. But, despite the range of input devices that have been developed since, and the ever increasing screen resolution and graphical capabilities of computer systems, the way we interact with GUIs has not changed very much in recent times. One of the questions that has to be asked is whether this apparent success is due to the ease of use of the paradigm itself, or rather because people have simply become accustomed to it since it was the only widely used

paradigm at the time that seemed appropriate for a non-specialized public. One of the issues that we hope to investigate in this thesis is whether the dominance of this paradigm still holds given that the general public has become increasingly exposed to other technologies and other types of interaction with computers and voice-enabled technologies – for example, the widespread use of mobile phones and text-messaging, instant messaging and VoIP protocols such as Skype.

Even though these standards for interaction are widely accepted, as technology improves and as computer users start to come from ever different populations and with broader backgrounds, interfaces will have to be developed that are more natural and require minimal effort to learn. One school of thought believes that since language is a natural, highly expressive and commonly used form of communication, computer interfaces that incorporate language by allowing users to talk to them, and by responding through language, will make the human-computer interaction experience more natural. This issue is discussed further in the next section.

2.4 Natural language interfaces

> *'Most users today are familiar enough with computers not to anthropomorphize them, and unfamiliar enough with natural language query systems to expect them to be similar to formal language systems.'* - KARLGREN [18]

The idea of using natural language as a means of human-computer communication dates back over thirty years. One of the main premises behind natural language interfaces is that they should make interaction with a computer easier for a user, in particular for new users, because they already have the basic communication skills to be able to interact with the system [18], whereas GUI systems impose a learning curve for new users in terms of how they can communicate with them.

Natural language systems started off as voice-only over the phone interaction systems, where the user would speak into the telephone, and a system would respond. Such systems were most often used for information seeking or travel reservation applications, where the domain was fairly small, the number of options for what a user was expected to say was limited and interaction could easily be constrained where necessary. Over time, graphical interfaces were added to such systems, giving rise to more complex interaction possibilities. Both types of system will be discussed here, although much of what follows

pertains largely to more complex natural language interfaces such as those incorporating a graphical component[2].

2.4.1 Advantages

Despite the existence of applications such as SMALTO [19], which is designed to help developers determine whether or not to include natural language use in a system, there is little consensus, other than at the very general level, as to when the use of a language-driven interface would be the most appropriate. Cohen and Oviatt [16], and Bradford [20] agree that it is likely to be in situations such as when the hands and/or the eyes are occupied with other tasks, when the availability of a screen or keyboard are limited, when the user is disabled, and when the tasks being undertaken necessitate mobility. All of these examples are related to physical or cognitive constraints on the user during interaction. However, another situation in which the use of language could be advantageous is when an expressive power is needed that cannot be achieved through more traditional GUI interfaces, in particular when it comes to question answering systems, systems that require the user to provide complex descriptions as input [21, 22], for example in cases such as accessing databases [23], or cases where some objects that the user wants to access are not visible on the screen [21, 24]. In such cases, natural language interfaces provide expressive power in terms of allowing the user to use definite descriptions, discourse reference, temporal information, quantification, coordination, negation, comparatives, sorting expressions, conditionals, causal relationships, and navigational shortcuts [7].

Moreover, Bretan and Karlgren [25] and Lefebvre, Duncan and Poirier [26] argue that cognitive load is shifted from the visual and manual channels, to the auditory channel, making it 'easier' to use the system. And, Rudnicky [27] states that there may be a reduction in the amount of effort needed to create input. There are however arguments by those such as Halverson [28], that in fact, at this stage in the development of technology, people are so familiar with keyboards, that using language may in fact not be more 'natural' when it comes to interacting with computers.

[2] Unless otherwise stated, the discussion in this chapter focuses on oral (rather than written) natural language interaction, although many of the problems cited apply to both oral and written language communication since they concern linguistic processing of natural language, which is independent of the input means.

2.4.2 Disadvantages

Despite all of the apparent advantages, there are disadvantages to using natural language interfaces which need to be carefully considered when choosing to use language in an interface, and in designing the interface itself.

Lack of explicit interaction history

Notably, with language, there is often no obvious trace of a communicative action once it has taken place. This can pose problems for inexperienced users of natural language systems [9]. Unless the system shows the user what it has recognized, the only trace of the input that was given is in the user's own memory, and as interaction progresses, the stream of exact input is forgotten. If the user encounters a problem in their interaction they may not be sure if it was caused by erroneous syntax, out of bounds vocabulary, or some particular stream of input.

Speech recognition and linguistic errors

This leads to the next crucial issue, which is the robustness of the speech recognition system that is available in the interface. Speech recognition systems in natural language interfaces generally show three types of errors: *rejection errors,* where the recognizer fails to find the term in question, *substitution errors*, where the recognizer finds a term, but it is not the term that the user intended, and *insertion errors*, where noise is recognized as a legal utterance [29]. Such errors, as Lai and Yankelovich [29] point out, can cause the user to create a false model of the linguistic capabilities of the system. In addition to the robustness of the recognizer, the noisiness of the environment also plays a role [30]. If the user's surroundings are too noisy, it will be harder for the recognizer to distinguish between sounds, and will increase the chance of substitution and insertion errors. Noisiness also plays a role in the opposite direction - the user of a natural language system could disrupt those around them [26]. Finally, there seems to be an underlying assumption that the speech recognition will take place in the user's native language. Work by Karat, Vergo and Nahamoo [31] has shown that native-language speakers tend to do better with language recognition systems than non-native speakers. For applications that are used in public places or by multiple users with varying accents, performance would be reduced, or production requirements increased, if the system were to accommodate a maximum number of users.

In addition to errors in the speech recognition itself, there are more general linguistic errors such as those outlined by Ogden and Bernick [32] and Walker and Whittaker [33] that need to be considered in the development of natural language systems. These include spelling/typing errors (in the case of natural language input via keyboard), lexical errors

(where the word used is outside of the domain or lexicon), syntactical errors (where the formulation that a user chooses cannot be handled by the system) and functional errors (where the user tries to make a command that is not available in the system).

Opaqueness

Finally, there is the problem that, as Bretan and Karlgren [25] put it, '*natural language systems are opaque*'. It is hard to tell what language capabilities a system has, and which are the legal and illegal utterances [23, 25, 29]. The problem then, is how best to teach a user the functional and linguistic limitations of a natural language system. An interface, and particularly a natural language interface, needs to be able to guide the user, to allow them to focus on their task, and be easy to learn and use – how best to accomplish this remains one of the biggest challenges of designing multimodal interfaces [18, 31, 32]. Bradford [20] for example suggests that the system should have a dialogue structure that gives the '*structural advantages*' of menus, but that at the same time allows free form natural language use.

System responses

Another suggestion, which is based on research from cognitive science and psychology, claims that people tend to alter their behaviour depending on whom they are interacting with [34, 35]. This claim relies to a large extent on the notion of linguistic convergence [18, 25, 36]. Linguistic convergence is a concept in which two interlocutors start off with different registers of language use, and over the course of a conversation, naturally adjust so that the linguistic level at which they are communicating (in terms of vocabulary, speed, complexity etc.) is as closely matched as naturally possible. In the case of human-computer interaction, the idea is that the user would linguistically converge on the level of the computer, and in doing so, would passively learn the linguistic level which the system can handle. One such way to do this is through the system prompts, and [32, 37-39] have shown that in fact, both prompts and presentation of information have a significant impact on how a user interacts with a system. Although Pirker, Loderer and Trost [35] do not believe that the notion of convergence would work in all cases, there seems to be sufficient evidence from other work to suggest that it is a viable option to explore.

However, care needs to be taken in how the system responds. Not only must the responses be correct, relevant, unambiguous, consistent and provide just the right amount of information [19], they must also not be too 'human'. Several authors [19, 21, 29, 40] have found that since users tend to have an instinctive reaction to language, they also tend to attribute more complexity and processing capabilities to computers that incorporate

natural language in an interface. As Oviatt [38] points out, the use of language can only be constrained up to a certain point, after which instinct takes over. Thus, natural language systems need to have sufficient linguistic capabilities, in terms of vocabulary, semantics, and dialogue structures in order to provide sufficient support for infrequent users [16].

Free-form vs. constrained language

One of the long-running discussions when it comes to language-based interfaces is that of using free-form or constrained language during the interaction. Oviatt, Cohen and Wang [37], and Ogden and Bernick [32] show that humans can and are willing to adapt to linguistic constraints imposed by a system, and as Ogden and Bernick [32] believe, in most cases users constrain themselves of their own accord. So, while natural language interaction can be fairly free-form, designers do have a margin in which they can attempt to constrain the scope of that interaction. In [32] Ogden and Bernick discuss four areas in natural language interfaces that need to be addressed: 1) lexical – where users need to be able to discover the expected vocabulary, 2) syntactic – where paraphrases should be possible and easy to find, 3) functional – where users should be able to figure out how things can be expressed, and 4) conceptual – where users need to be able to know what can be expressed. But, in order to be truly successful, interface techniques will need to be developed for interaction with natural language interfaces that can help users find the right way in which to communicate with the system [37]. However, as Ogden and Bernick [32] point out, it is not clear just how much a system will need to 'understand' in order for it to be natural to use.

2.4.3 Dialogue types and system architectures

Dybkjaer, Bernsen, and Minker [41] suggest that there are three general types of dialogue that the user can have with the system: goal oriented, practical and conversational. In goal oriented dialogues the user simply provides the information that the system needs in order to fulfill some task. These are often based on finite state systems, which take the user through a set of predetermined steps in order to find some piece of information [31, 42]. Practical dialogues are more complex than goal-oriented dialogues because the dialogue becomes more conversational, giving the user more freedom in how they interact with the system. Such systems are most often based on a frame-based architecture, which involves the system and user working together to fill a sufficient number of slots (pieces of information/search criteria) for the system to be able to find some information. In conversational dialogues, the user can directly state in richer and more complex terms exactly what they are looking for, and it is up to the system to interpret this information appropriately by '*defining and discussing tasks, rather than by*

executing a series of commands' [43]. In order to accommodate such complexity, agent-based systems should have the capability to reason and have beliefs about the world. Many goal-oriented dialogue systems can still be found today, particularly in over-the-phone information-seeking applications. Most more complex systems are practical though, according to the classification above, with only a few such as the Hans Christian Andersen system [44] attempting to go beyond that to the conversational level.

Dialogue systems, independent of the architecture, tend to have the following components [23]: speech recognition, language understanding (semantic processing of user input), dialogue management, communication with the internal system (i.e. retrieval of items from a database), response generation (preparing/retrieving the appropriate response), and speech output (providing the response to the user). The dialogue management component involves the dialogue model which, particularly in the case of practical and conversational dialogues, can be further broken down into components such as the dialogue history, the task record, a world knowledge model, a domain model, a generic conversational competence, and a user model.

2.4.4 Novice vs. expert users and system control

An issue that plays an important role in any type of system, but seems to be particularly problematic for natural language or speech based systems, is the difference in interaction between novice and expert users. Novice users tend to need much more guidance, since they are less familiar with the capabilities of the system and the range and register of the language that they can use to communicate successfully. Such problems can be rectified for example by giving users less control over the system, or providing an increased number of confirmation prompts. However, these are the very features which are likely to inconvenience expert users who tend to know what they want and how to get it without help from the system [45].

There are three general ways in which the dialogue between the system and the user can proceed [23, 29]. The first is system-driven, where the system is in total control of the interaction and the user's role is simply to answer the questions that the system poses. Such systems are ideal for novice users. The second is user-driven, which tends to be preferred by expert users since they give the user full control of the interaction and it is the user who is responsible for specifying sufficient amounts of information so that the system can find what they need. The third possibility, mixed-initiative systems, is a compromise between these two extremes which, not surprisingly, is the option that is taken most often. Mixed-initiative systems rely on the user and system working in a collaborative manner to find the information. However, Bretan and Karlgren [25] point

out that in many cases, users themselves, independent of their competence, expect to control the discourse management of a system, most likely because they find it difficult to believe that the system will be capable of handling this on its own. This suggests that in fact, many mixed-initiative systems are mixed, but that the mix is not necessarily equally distributed between the user and the system, but rather that the user has slightly more control over the system at all times. Another way to facilitate use of a dialogue for expert users is to allow for barge-in, where the user can interrupt the system at any time, for example if they wish to ignore the prompt, or take the interaction in a different direction [29, 31].

2.4.5 Development and evaluation of NL systems

The literature on natural language interfaces is quite large and takes several different approaches. At the higher level researchers such as Jönsson [46, 47] and Dahlbäck and Jönsson [22] focus on the dialogue and discourse representations that are necessary for successful natural language communication between a human and a computer. At the lower level research focuses on the specific linguistic issues that need to be resolved when implementing natural language interfaces. Androutsopoulos, Ritchie, and Thanisch [48] provide a thorough overview of the use of natural language interfaces for database access, focusing on the benefits and drawbacks, and the linguistic problems faced during implementation while Nerbonne [49] describes the core and secondary requirements in general terms. Similarly, Ogden and Bernick [32] provide a broad overview of the field in general, including design and evaluation methodologies. Such general design guidelines are useful, but as Dahlbäck, Jönsson and Ahrenberg [34] point out, the language that is used in interaction with existing technologies may differ when the underlying technology changes, and thus analysis based on existing technologies may not be appropriate. Moore and Morris [50] note that human-machine communications seem to be shorter and more goal oriented than human-human communications.

There is also quite a large body of work on the development processes involved [23] and on system requirements gathering methods such as user studies, speech corpora and Wizard of Oz experiments [23, 25, 32]. However, there is little in terms of explicit guidelines. Dybkjaer, Bernsen and Minker [41] and Le Bigot et al. [51] argue that a lot of additional information is needed in order to build ‘usable’ spoken language dialogues. This includes more detailed investigations into general user behaviour, both at the linguistic and non-linguistic levels. But, as Dybkjaer and Bernsen [19] point out, little research has been done in these areas. It should be mentioned that some best practice guidelines for designing spoken language dialogue systems do exist, such as those presented by Dybkjaer and Bernsen in [19].

2.4.6 Conclusions

Despite the fact that natural language systems seem to offer a myriad of advantages over GUIs, they too have their problems. Nerbonne [49] suggests that although natural language interfaces are meant to free the user from such problems as knowing the underlying structure of the data that they are trying to access or the language that is available to them to do it in, they seldom fulfill these goals. Sturm, Wang and Cranen [45] note that even in very limited domains, speech recognition will not be perfect since there will always be a chance of users using words that are outside the scope of the domain or its predefined lexicon. They go on to point out that quite often users have problems building correct mental models of speech-only systems simply because they do not get sufficient feedback on their actions. Moreover, environmental noise is also a factor that limits widespread use of NL systems [30]. In fact, some of the important problems cited in early work on natural language keyboard input to databases done by Androutsopoulos, Richie and Thanisch [48] have not yet been well resolved. These include the fact that system capabilities and the status of the system aren't always clear to the user, and that the user has no way of knowing whether an error is due to a problem with the linguistic or conceptual coverage. Moreover, Bretan and Karlgren [25] propose that one of the key methodologies for gathering data for natural language systems, Wizard of Oz studies, may be flawed. They suggest that in fact, the data gathered may not be linguistically accurate since the user will always find a way to communicate with the system at the right register, in part because the system is in fact a wizard, who may unconsciously relax restrictions.

Cohen, McGee and Clow [52] note that, '*…in order to affect mainstream computing, spoken interaction would at a minimum need to be found to be superior to graphical user interfaces (GUIs) for a variety of tasks*.' But, at this point in time, various studies have shown that it is in fact not clear that such an advantage exists, except in very particular situations [27, 53]. For example Christian et al. [54] found that formulating a voice command takes more time than doing the equivalent action through clicking, but this assumed that the object being clicked on was visible on the screen. They also point out that the preference for voice interaction may depend on the amount of time that a user has spent with the system – as they spend more time, they learn all of the functionalities that are available in the system, and thus know exactly which functionalities to access and how, without them having to be visible on the screen. For the time being at least, it is still not clear in which situation and for which types of tasks spoken interaction will truly be advantageous [40, 55].

2.5 Multimodal interfaces

One of the solutions proposed to resolve the problems inherent to both GUI interfaces and natural language interfaces, is to blend the two together into multimodal interfaces, which provide '*greater expressive power, naturalness, flexibility and portability*' [56]. Advances in technology are making powerful multimodal interfaces an increasingly viable option for solving certain usability problems that are faced by other types of interfaces.

2.5.1 Multimedia vs. multimodal

When it comes to defining what constitutes a multimodal interface, several authors have their own subtly different definitions. One point of contention seems to be the delimitation between multimedia and multimodal systems. Maybury and Lee [57] refer to *mode* or *modality* as '*the human senses employed to process incoming information*', such as vision, touch and hearing, whereas *medium* refers '*both to the material object (e.g. paper, video) as well as the means by which information is conveyed (e.g. a sheet of paper with text on it)*'. Anastopoulou, Baber and Sharples [24], as well as Sutcliffe [58] agree with this general definition. Sutcliffe puts it perhaps the most succinctly when he says: '*The message is conveyed by a medium and received through a modality*' [58]. These definitions however, only seem to support the notion of multimodality to process computer output, and not from both the input and output perspectives. Sutcliffe [58] begins to address the issue by granting that a modality is '*a sense by which a message is sent or received by people or machines*'. Coutaz and Caelen [59] take a more input based perspective and define a multimodal system as being one that '*is able to support human modalities such as gesture, written or spoken natural language*'. However, what is really needed is a definition that takes into account both multimodal input and output.

Coutaz, Salber and Balbo [4] present a more generalized definition, where they discuss multimodality in a system as depending on the availability and use of several different input and output channels, and most importantly on the system being able to process the incoming information at different levels of abstraction. This last point seems to be the common thread in several definitions [14, 24, 30, 59-61] – that multimodal systems are capable of interpreting and producing semantically driven information and can handle input and output from several channels (sequentially or simultaneously), whereas multimedia systems do not perform any semantic interpretation and can only give *output* along multiple channels. As Coutaz, Salber and Balbo [4] and Coutaz and Caelen [59] also point out, a system can be both multimedia and multimodal, where for example, user input can be multimodal and is interpreted as such, but system output is uniquely

multimedia. Anastopoulou, Baber and Sharples [24] take this a step further when they claim that '*While in multimedia systems the user has to adapt to the system's perceptual capabilities, in multimodal systems the system adapts to the preferences and needs of the user*'. Finally, Maybury and Lee [57] also note that '*the notions of medium and mode are somehow always defined relative to the interests and purposes of a particular kind of application*'. Thus, it seems that no completely clear cut definition can be made. For the purpose of this work, we will consider an interface to be multimodal if it can accept and process input from different types of input modalities including pointing and language.

2.5.2 Advantages

Some authors have argued that for many types of tasks, there is no proof of a gain in efficiency with a multimodal interface. However, authors such as Oviatt [56, 62] believe that efficiency is not necessarily the best indicator of advantage. One of the major advantages that multimodal interfaces have over uni-modal ones is their flexibility in giving the user a choice in which modality to use [5, 14, 61-64]. Oviatt [56, 62] notes that users seem to have a '*strong and nearly universal preference to interact multimodally*' even if they don't issue every command in a multimodal manner. In many cases this includes the use of voice, which as was pointed out in section 2.4, is thought to make the system more *natural* to use. Nass and Gong, [65], Jokinen [14] and Anastoploulou et al. [24] argue that what might be natural, or human-like, is not necessarily what will make a system easy to use or effective and that it is up to designers to find the right balance between making an interface natural and easy to use/efficient. However, such design is also thought to be application-specific and cannot be easily generalized or abstracted. But, many authors agree on the fact that giving users a choice as to which modalities to use leads to greater stability during use [30, 42], and increased usability [25, 42]. Bretan and Karlgren [25] point out that overall, the usability of a multimodal system seems to be greater than that of a uni-modal system, and Oviatt et al. [30, 66, 67] claim that multimodal systems are '*better in exactly the cases where uni-modal systems fail*'.

The choice of modalities by any particular user can be influenced by a variety of factors such as the environment in which the system is being used, the task for which it is being used, and even the particular preference of a user for a particular modality, which can be influenced by factors such as their age, skill level, native language, cognitive style and physical impairments [62, 64, 66, 67]. It may seem strange to think that a system that gives a user more choices in how they interact with it, especially one that includes error-prone modalities, may actually improve robustness, but several authors have found this to be the case. The robustness of speech recognition can be improved through mutual disambiguation via input from other modalities [14, 30, 36, 42, 56, 62, 64, 68-70]. Error

handling can also be improved, and recovery from errors can be done in a manner that is more efficient and less frustrating for users [30]. Moreover, studies have shown that language that is used in multimodal interactions is simplified in comparison with that used in natural language-only interfaces [30, 34, 42, 56, 62], which in turn can reduce the complexity needed in the language processing components of a multimodal system. For example, human-computer dialogues tend to have simpler structure than those between humans, are generally shorter, have fewer disfluencies, and use fewer co-occurring expressions, all of which are elements that normally pose problems for speech recognition systems [34, 56, 62].

Bell et al. [36] and Oviatt [30, 38, 42, 56] noticed that modality switching seems to occur most frequently when there is a problem. So, in order to be able to do the type of leveraging needed to improve robustness, Oviatt [30, 42, 56, 67], Grasso [40] and Reeves et al. [63] believe that the input modalities available must have '*semantically rich information sources'*, must be complimentary, and should be able to duplicate functionalities [16, 30, 40, 42, 56, 62, 67]. Moreover, Oviatt [56, 62] points out that when several modalities are available, voice is not always the dominant modality – it is not the modality that carries the most information and it is not the modality that is used first. Furthermore, some modalities are better suited to transmitting certain types of content than others, and a user's choice of modality will depend on the nature of the content that they want to transmit. She also found that speech and pointing are not the dominant integration pattern and warns that systems should not be designed with that in mind.

2.5.3 Disadvantages

Two known disadvantages for multimodal interfaces is that their architectures are much more complex [14, 42], and there is a higher risk of cognitive overload for the user [14]. Work in [71] and [70], has shown that '*people are more efficient when tasks are distributed among several input channels*'. But, several authors note that various lacunae are present in multimodal interface design and development. Whittaker and Walker [72] discuss the fact that there is little theory about the interaction between graphics and text, and how and when graphics should be used with other media. In [29] Lai and Yankelovich stress the need to understand how, where and why users interact with an application which in turn should drive the design. Oviatt et al. [42, 56] stress the need for more work on how different modes are combined and organized in human-computer interaction [42, 56]. In some of her work [56, 62], she found that there is significant variability in how individual users integrate multimodal patterns. She also stresses the need for longitudinal studies on the subject, as well as the development of better tools to

help develop successful multimodal systems. In [62] Oviatt states that design of multimodal interfaces will depend on multidisciplinary information and in particular on:

- properties of modes and their information content
- characteristics of multimodal language
- integration and synchronization characteristics of input
- predictions of how users will act multimodally
- how alike different users are in their integration patterns

Even though some work has been done in these areas, notably by Oviatt et al. [42] for speech and pen input and linguistic features in map-based tasks, and by Grasso [40] for speech and pen input in the medical domain, as Salber and Coutaz [60] summarize, there is simply not enough understanding of multimodal HCI.

2.5.4 Multimodal architectures

What does exist however is some insight into the nature of interactions, such as whether multimodal interaction tends to be sequential or concurrent, the level at which modality fusion should be done, and the nature of multimodality in the interaction. In the first case, Oviatt et al. [73] found that most multimodal constructions are performed sequentially rather than concurrently, with the manual actions done first, followed by vocal cues. Coutaz and Caelen [59] put forth the idea that multimodal interaction can be of two types – exclusive or synergic. In exclusively multimodal interaction, multiple modalities are used, but they are used independently of one another. For example, a voice command will be issued and executed, to be followed by the next command which is issued through pointing. In synergic multimodal interaction, one command is issued via several modalities. But, as Oviatt [56, 62] points out, even in cases where a single command is issued using multiple modalities, the modalities are often not executed simultaneously, with for example, pen input preceding speech in 99% of the cases. Such differences then need to be taken into account when deciding how and on which level to perform modality fusion. Fusion can occur on several levels, which in turn influences the type of architecture that needs to be adopted. Feature level fusion architectures perform fusion at the lowest level, so much of the later information retrieval and processing that is done, is done with only one overall ‘feature’ in view. Semantic fusion allows for the interpretation of individual modalities first, and fusion occurs on the interpretation.

In [67], Oviatt and Cohen describe multimodal architectures in greater detail, and the Quickset system architecture in particular. And in [62] Oviatt gives an excellent comparison of GUI and multimodal architectures – Table 2.

GUI Architecture	Multimodal Architecture
Single event stream	Continuous and simultaneous input from parallel streams
Interactions are atomic and unambiguous	Interactions are ambiguous and result from combined modes
Separate from application and locally installed	processing is often distributed
	Time stamping of input and temporal constraints in modality fusion.

Table 2: Comparison of GUI and multimodal interface architectures (from Oviatt [62])

Coutaz et al. [4] suggest the following classification for multimodal systems:

- Exclusively multimodal – '*if input (or output) expressions are built up from one channel only and no parallelism is permitted at the interface*'
- Alternately multimodal – '*if input (or output) expressions are built up from multiple channels but no parallelism is supported*'
- Concurrently multimodal – '*if input (or output) expressions are built up from one channel only and parallelism is permitted*'
- Synergistically multimodal – '*if input (or output) expressions are built up from multiple channels and parallelism is permitted*'

In [74] Coutaz, Nigay and Salber propose a method, CARE, for reasoning about multimodality in terms of modality **c**omplimentarity, **a**ssignment, **r**edundancy and **e**quivalence, in a system.

Finally, both Lai and Yankelovich [29] and Reeves et al. [63] suggest that since prompts in multimodal interfaces can be both spoken or presented as text, the content and context will play a large role in deciding the appropriate strategies for their presentation.

2.5.5 Existing systems

Oviatt [62] and Ogden and Bernick [32] provide extensive lists of multimodal and natural language interface systems respectively. However, it is important to note two things in particular about the systems described therein. In the case of the multimodal systems, as Oviatt herself states, the systems and most existing systems in the field to date, are actually bi-modal rather than multimodal. In the case of natural language interfaces, where the work in the field is relatively extensive, little is ever mentioned about the incorporation of additional modalities with natural language, and particularly the incorporation of more than one modality. Other systems, such as SmartKom [75], MERIT [76] and COR [76] allow for multimodality, but the research focus is primarily on the dialogue aspect of multimodal interaction rather then the choice of modalities and their influence on the task and the nature of the interaction. In the rest of this section, we outline in general terms some of the most commonly mentioned multimodal systems and platforms.

- MATIS - Coutaz, Nigay and Salber [74]
  MATIS (Multimodal Airline Travel Information System) lets users use speech, direct manipulation, keyboard and mouse or any combination of them to retrieve travel information. The user is free to choose any modality they wish for any of the tasks that are possible in the system.

- EMBASSI - Elting et al. [77]
  Uses speech, gestures and GUI manipulation as input, and chooses appropriate modalities from speech, an avatar and/or GUI as output.

- LARRI - Bohus and Rudnicky [78]
  (Languge-based Agent for Retrieval of Repair Information) uses GUI combined with a spoken dialogue system for information-access and task guidance for the '*support of maintenance and repair activities for aircraft mechanics*'.

- Voice-paint and Notebook - Goudrol et al. [79]
  Allows for mouse, keyboard and voice to manipulate standard drawing and text-editing applications.

- SmartKom - Wahlster [75]
  A '*mixed initiative multimodal dialogue system that combines speech, gesture, and facial expressions for input and output*'. The system output is provided through an interface agent (avatar).

- QuickSet - Oviatt et al. [42]
  Allows users to create and place entities on a map using speech, pen and direct manipulation as input modalities.

- HCWP - Oviatt et al. [42]
  The Human-Centric Word Processor, from IBM, which allows for speech and pen input to correct and format text that has been dictated.

- Boeing's VR Aircraft Maintenance Training prototype - Oviatt et al. [42]
  Lets users use speech or gestures in a VR environment to '*asses the maintainability of new aircraft designs and train mechanics in maintenance procedures*.'

- Field Medic Information System - Oviatt et al. [42]
  Allows medics to see and modify patient records through voice or pen, and make free-form notes in cases where their hands and/or eyes are busy with the patient.

- Portable voice assistant - Oviatt et al. [42]
  Pen/voice interface that lets users access or enter data on the web. Users can choose either of the modalities.

In [80] Waibel et al. discuss multimodal systems involving technologies such as automatic lip-reading to enhance speech recognition, gesture recognition and on-line cursive handwriting recognition. However, since these technologies are not used in the work in this thesis, they will not be discussed in any more detail.

2.5.6 The past, present and future of multimodal interfaces

As Oviatt [62] points out, one of the first multimodal interfaces was put forth by Bolt with his 'Put that there' system, which allowed users to blend voice and pointing actions in order to rearrange blocks on a screen. Early multimodal systems usually involved the combination of voice and mouse/touchpad and since then, research has gone into speech and pen interfaces, as well as speech and lip movement, which according to Oviatt are the two most mature modality combinations. It is also important to note that much of the early work involved users working with multimedia maps in either tourist information seeking, or military planning contexts.

Some of the results from this early work showed that multimodal interfaces did indeed give users advantages, in particular for map-based tasks [52], but work also highlighted the fact that the advantages might very well be context and task dependent. For example, Oviatt [56, 62] found that voice input was particularly useful for descriptive and temporally oriented tasks, but pen was used more for denoting digits, symbols and graphical content [42] or in public environments where background noise and privacy issues became constraining factors. Moreover, findings in [42] showed that speech and pen were particularly useful in mobile tasks and visio-spatial applications. In the case of mobility, Oviatt points out that this also expands the number of different contexts in which interfaces could be used, such as use in field environments [64, 67]. In such cases, the architecture could be designed in such a way as to adapt the weighting put on the acceptance or inclusion of various input modes depending on the external environment [63, 64].

The future of multimodal interfaces, according to Oviatt at least, will involve adaptive interfaces, which will allow a much wider variety of user groups [30], as well as what she calls '*perceptual user interfaces*' which incorporate technologies such as gesture recognition and eye tracking to help improve the user experience. In this case, interfaces could be divided into two general groups: passive interfaces, which include gesture recognition and eye-tracking data that is processed on the backend and used to help the user experience in a passive manner and active interfaces which have a more direct role in interacting with the user. The combination of passive and active input modes thus leads to what she calls blended interfaces, which take the best of these two worlds in order to maximize the user experience [42, 62, 67].

2.6 Natural interactivity

Nils Ole Bernsen, along with colleagues Laila and Hans Dybkjaer, puts forth the idea of *natural interactivity* in relation to multimodal systems. For them a natural system is '*one which allows users to use free and unconstrained spontaneous speech in efficiently achieving their goals*' [81]. But, he is also careful to point out that '*Natural interactivity is multimodal most of the time, of course. But a multimodal system is not necessarily a natural interactive system. Multimodality in a system merely signifies that users may, or must, exchange information with the system using several different input and/or output modalities*' [82].

In [82] Bernsen takes the idea of naturally interactive systems further by suggesting that ideally, a naturally interactive system should approximate the '*role of an extremely capable assistant or servant*', but that the technologies to create such a system are either not yet available, or not sufficiently robust for regular use. Moreover, he highlights that '*…naturalness is never a property of language in isolation, but rather a property of the relation between the language and the set of things it will be used to express*' [18], so in certain cases, natural interactivity might in fact be less suited to a particular interaction, in which case the desire to make a system naturally interactive should not take precedence over allowing a user to use that system to solve a particular task. In the same article, he does mention however four projects that were underway at the time which have begun to deal with the issues found in creating naturally interactive systems. These are:

1. the DARPA Communicator project whose '*goal is to foster the next generation of intelligent multiparty conversational interfaces to distributed information*'

2. the Oxygen project from MIT which '*focuses on the development of a global infrastructure for technology-mediated human-to-human communication*'
3. the SmartKom project which '*focuses on natural interactivity (starting from spoken dialogue) and multimodal interfaces*' and '*emphasizes individual adaptivity and cartoon-like presentation agents*'
4. and CLASS, a project from the European Human Language Technologies Project whose aim is to '*specify a reference platform and architecture for next-generation natural interactive systems as well as to develop a best practice development methodology for natural interactive systems*'

2.7 Modality theory

The notions of modality and multimodality have been explored in different contexts and with different goals in mind. Mark Maybury [83] focuses on the higher level communicative role that multimodality and multimedia play in interaction. He proposes ways in which multimedia and multimodal dialogues can be structured through the use of communicative acts, and argues that linguistic, dialogue and graphical acts all '*have specifically interpretable roles in multimodal communication*' and outlines what those roles are. Nils Ole Bernsen on the other hand approaches the problem from the perspective of the modalities themselves. In recent work [5, 83-85], he has proposed the notion of Modality Theory, in which he addresses the problem of determining which input/output modalities are best suited for the exchange of particular types of information in particular contexts. Another way to view this is that he attempts to provide a theoretical basis for getting from the requirements for an application to the '*selection of input/output modalities for the application which will optimise the usability and naturalness of interaction*'.

His approach, which takes the form of a generative taxonomy, is to decompose all modalities into their most basic, 'atomic', elements and then investigate the role of those elements (and elements in combination) in the representation and exchange of different types of information in various contexts. But, he points out that Modality Theory only addresses the media of graphics, acoustics and haptics (at least for the time being), that it is more focused on the choice of representational modalities (as distinguished from sensory modalities in psychological literature) and not on the devices which are used to manifest them, and has thus far been primarily developed for output modalities. While he

says that work on a similar taxonomy for input modalities has begun, it has been difficult to find any literature on it. One particular aspect of modality theory which is of interest here and is described by Bernsen in [86-88], is the '*speech functionality problem*' which is '*the question of what speech is good or bad for, or under which conditions to use, or not to use, speech for information representation and exchange*' and in [86] he presents an online system that uses modality theory to make this determination. However, the work primarily pertains to the use of speech as an output, rather than an input modality. But, Lee and Maybury [89] point out, it is the combination of these two approaches that is needed to fully understand and structure multimodal interaction.

2.8 User modelling and sociological considerations

> '*The user's actions are based on the user's needs in a certain moment, on the user's assessment of the situation (including the technology they perceive and the functionality they interpret to be available) and on the user's knowledge of procedures and expectations of the outcome*'
>
> -VAN DER VEER AND MELGUIZO [90]

The final two aspects that need to be taken into consideration in the scope of this thesis are how the user perceives the system, both from a technical, and a more socio-psychological point of view.

The notion of user models of a system is quite well known in HCI and pertains to the user's knowledge, or model, of the system. There are two levels at which the knowledge is important: structural and functional. The structural level refers to the user's understanding of the system - the functionalities that are available, the domain being treated, how the data is organized etc. The functional level refers to how to actually perform specific tasks using the system. According to van der Veer and Melguizo [90], it is the functional model that is more important for novice users since it is easier to learn than the structural model, and allows them to immediately get tasks done. Particularly, they point out, if there are fewer 'rules' for accessing the same functionality. In this case, even novice users can find solutions in novel situations, simply by developing a new mental model based on analogy. However, they also point out that for complex systems, the division between the functional and structural models is more blurred.

While the mental model considers the user's perception of a system from a more functional and technical point of view, it is equally important to consider their socio-psychological experience. At the most basic level, the user's perception of the system can change simply depending on how familiar the system looks, which in turn influences how

comfortable they are with the system. This is particularly important for novice computer users who require more support from the system itself than experienced users [91]. If the system is similar to one they have already used and feel happy working with, they will be more at ease working with a new system. But, on the other hand, if the system is too different, or in the particular case of multimodal systems if the user is intimidated by the potentially unfamiliar new technologies being incorporated, they are likely to be more anxious, which in turn can negatively impact their experience with the system. Anxiety, for example, can induce an effect known as tunnel vision [10], where a user becomes so concentrated on a particular approach or method for solving a task, that they are unable to see other potential solutions, even ones that are quite obvious. Conversely, those that are at ease with a system, or as Salber and Coutaz [10] put it, '*happy*' users, are much better at seeing alternative methods. A similar issue is the amount of control that the user has over a system. Novice users may prefer to be guided if they do not know exactly what they want to do with the system, but at the same time, not having a sufficient degree of control over a system can lead to frustration [91]. Finding the right balance based on the degree of user experience and the context in which the system is being used is a critical issue in human-computer interaction [91]. As Popescu et al. [91] put it, '*An application that users like can do no wrong, whereas one that users dislike does everything to anger them, regardless of the application's actual behaviour*.'

But, the user's perception of the system extends much deeper than simply to the aesthetic level. Nass and Moon [92] found that human beings 'mindlessly' attribute humanness to a computer. They found that user's gender-stereotype computers, attribute loyalty behaviour, accord politeness and reciprocity to computers, and behave differently if they are told that the computer is a specialist in the domain rather than a generalist, even though the underlying software is the same. These findings were the result of interaction with regular GUI systems, and their consequences for multimodal voice-enabled systems will likely have an even greater impact on how users view and react to computers.

Nass and Gong [65] found that since humans react to speech instinctively, they also apply the same interaction heuristics when communicating with computers as they would with other humans. This is the case for instance when there is an apparent breakdown in communication. When a person feels that their interlocutor has not understood them, they will take one of several corrective actions such as hyper-articulating the phrase, or reformulating it, and decreasing disfluencies. Nass and Gong [65] found that similar actions are taken when the interlocutor is a voice-enabled system. Additionally, they found that users also unconsciously assign gender roles to synthesized voices, which has a direct impact on the impression that the choice of synthetic voice for a system will have

on the user. For example, male voices are found to be more authoritative, while female voices are more comforting.

Jussi Karlgren [18] suggests that work from discourse theorists regarding the knowledge that a person has of their discourse counterpart, how the counterpart's behaviour is monitored and how a person adjusts their own behaviour accordingly, in particular in their adaptation of linguistic behaviour, is an important consideration for designing systems that involve human-computer dialogue. This implies that users will have certain expectations of a system based on its apparent linguistic competences. If a system seems to use language very smoothly, the user might assume that its linguistic processing capabilities are high, and attempt to use it accordingly. If there is a mismatch between the capabilities that the system does have, and the apparent capabilities, this might cause problems. However, he also points out that as increasing numbers of voice-enabled systems appear on the market, users' expectations will become more realistic and their attitudes towards language-enabled systems might change, at least in some respects.

2.9 Multimodal meeting domain

Thus far we have discussed primarily the design and cognitive issues related to the field of multimodal interface design. In this last section we would like to address the work that had been done on the domain that is the subject of this thesis – multimedia/multimodal meeting processing and retrieval. In the following, we will discuss both past and current projects in the domain, as well as a taxonomy for meeting browsers that was developed at the University of Sheffield.

2.9.1 Existing projects

Several projects dealing with multimodal meeting recoding, browsing and retrieval have to various degrees, either finished or are currently under way. Most of these projects, described in the sections below, deal with the development of technologies for meeting recording and processing. For those few that have developed browsers for users to access the processed data, we provide a brief description of the browser.

2.9.1.1 NIST Meeting Recognition Project and the SmartSpace Laboratory

Work in the NIST Meeting Recognition (www.nist.gov/speech/test_beds/mr_proj/index.html) and SmartSpace Laboratory (www.nist.gov/smartspace/) projects focuses on development of technologies to record meetings in audio and video form and to perform analyses on the gathered data.

2.9.1.2 ICSI meeting corpus

This project (www.icsi.berkeley.edu/Speech/mr/), which finished in 2006, focused on gathering audio corpora of meetings and performing linguistically driven analysis on the speech to, for example, make dialogue act annotations and determine meeting hot-spots.

2.9.1.3 M4

The M4 (www.dcs.shef.ac.uk/spandh/projects/m4/index.html) project, which finished in 2005, dealt with creating a system to structure, browse and query recorded and automatically analyzed meetings. The artifacts that resulted from the recordings were audio, video, text and interaction information. The browser that was developed within the project (and tested on a corpus of Dutch parliamentary meetings), aligned video and audio with a text transcript, and provided browsing and searching facilities.

2.9.1.4 Interactive Multimodal Information Management (IM2)

The IM2 project (www.im2.ch) tackles the issue of multimodal meeting recording, data analysis and access. They have developed a Smart Meeting Room in which meetings are recorded in audio and video form and meeting artifacts such as participants' notes, whiteboard data, slides used in the meetings and any documents brought to the meeting are stored in electronic form. A text transcript of the meeting is then produced, and is annotated with dialogue acts and topic segmentation. They have also developed methods for aligning documents used in meetings with the relevant part of a meeting. Within this project, several interfaces with different foci have been developed. These include the document-centric browsers FriDoc and FaericWorld, JFerret - a modular architecture for developing personalizable meeting browsers, TQB, a form-based browser which lets users retrieve meeting data based on dialogue act and topic annotations on the transcript, and the multimodal Archivus interface, which is used in this thesis. Detailed descriptions of these interfaces can be found in [93].

2.9.1.5 Augmented Multiparty Interaction (AMI)

The AMI project (www.amiproject.org/) is a European project concerned with developing technologies to facilitate the recording, processing, storage and browsing of multimodal meetings. Little literature is available

on the browsers that have been developed within the project other than the JFerret browser already mentioned in the related IM2 project.

2.9.1.6 Carnegie Mellon University Meeting Room Project

The Interactive Systems Lab at Carnegie Mellon University has developed a meeting room (http://penance.is.cs.cmu.edu/meeting_room/) which is composed of a multimodal people identifier (people segmentation, colour appearance identification, speaker identification and sound source position, face identification, multimodal input fusion), a speech recognizer and a meeting browser [94]. The main view of the browser shows a display of the meeting over time, a meeting transcript and either a video from the meeting or a dialogue summary. The transcript, which also includes annotations of discourse features and emotions, is time-aligned with the audio and video files, and highlights text as it is said in the media files. The browser can create audio, video and text summaries, and store meeting transcripts as they are being created. There is also a 'meeting archive', which presents a meeting in a tree format and allows for searching on the meetings using a variety of predefined criteria. Their browser is primarily aimed for use during the meeting.

2.9.2 Existing browsers for the multimodal meeting domain

In an overview paper describing the state of the art in meeting browsers (where their definition of a *browser* covers '*any post-hoc examination of meeting data*'), Tucker and Whittaker [2] begin by proving a browser taxonomy driven by the focus of navigation of the browsers and to a secondary degree by the unique properties of that focus. They define the focus of a browser as '*either the main device for navigating the data, or the primary mode of presenting the meeting data to the user*.' Their taxonomy includes

- **Audio browsers** – These are browsers that are based on audio data presented with or without visual feedback. Audio browsers without video feedback could for example present the information in a sped-up form, or present only salient information derived from pauses in speech, or intonation. Those with visual feedback use annotations on the audio track such as speaker turns to present a visual overview of the meeting.

- **Video browsers** – These are browsers that use features of video such as keyframes to show overview information of meetings which can then be used to access video from the meeting itself. The authors argue that due to

the fact that video is usually augmented with other information it does not in and of itself contain the right types of information that are needed for browsing meetings.

- **Artifact browsers** – These are browsers which are based on physical items recorded during the meeting such as slides or personal notes, and which aren't audio or video. Such browsers often use the artifacts as indexing mechanisms, which allow users to pinpoint parts of a meeting that they are interested in and jump directly to them.

- **Discourse browsers** – These are similar to artifact browsers, but are also searchable, relying on annotations on the transcript for the indexing mechanism. Such annotations could be dialogue acts, topic segments, keywords, and metadata such as named entities, dates and locations.

The authors also cite several examples of meeting browsers which fit the different categories of the above taxonomy to varying degrees. However, none of the browsers mentioned by the authors, nor those described in the various projects in the previous section (with the exception of Archivus), are claimed to be multimodal. They seem to rely on fairly standard input mechanisms such as mouse and keyboard for computer-based interfaces, or standard knobs and buttons on physical devices for audio and video browsers. Thus, from the current literature we can draw conclusions about the types of technologies that might be useful to include in a meeting browser from a technical perspective. However, nothing concrete can be said about which of those technologies would be most useful from the user's perspective, nor how the introduction of various input modalities might alter how a user perceives and interacts with a meeting browser.

2.10 Discussion

As can be seen from the work discussed in this chapter, there are numerous factors that need to be taken into account when designing any type of interface. There is an established body of knowledge about graphical user interfaces and the interaction paradigms we apply to them. Natural language interfaces, which attempt to make human-computer interaction more natural, also have their problems. The hybrid between these two types - multimodal interfaces - seems to be the optimal solution, in particular given the rapidly advancing state of the various technologies involved. However, the field of multimodal interfaces is still quite new, and the body of knowledge needed in order to

develop such interfaces successfully is not yet complete. Consequently, the work in this thesis will try to build as much as possible on the established knowledge as outlined in this chapter, while trying to find new solutions to the problems that will inevitably arise, and through these solutions, attempt to enrich the field in general.

3. Research Goals

The research goal of this thesis is to assess whether multimodal input brings added value to interaction for the multimedia meeting browsing and retrieval domain, and if it does, what the nature of that interaction is. Following lines similar to those of Larsen [95], we define *added value* in terms of increased performance when compared to standard mouse and keyboard input, the usefulness of multiple modalities, and overall subjective user satisfaction when interacting multimodally with an interface. In particular, we are interested in the benefits and drawbacks that novel input modalities such as voice and pen bring to interaction, especially in the presence of more familiar modalities such as the mouse and keyboard.

Our work focuses on 6 central questions:

1. How often are different modalities used, alone and in combination, for meeting browsing and retrieval tasks?
2. Are some modalities more suited to finding certain types of information than others?
3. Do certain modalities or modality combinations make the system easier to learn, leading to an increase in performance in the long term?
4. Does modality use change when a user encounters a problem during interaction?
5. How do users perceive different modalities?
6. Does learning to use a system with a particular set of modalities influence how those modalities are used when other modalities also become available?

Before going into further detail about how we plan to answer these questions, it is important to define three concepts that will be frequently referred to in this work. The first is the notion of *familiar* vs. *novel* modalities. We consider familiar modalities to be those which most (if not all) users are accustomed to using when interacting with a computer interface – namely the mouse and keyboard. Novel modalities on the other hand are any input modalities with which the average computer user has had limited or no experience in the past. In the case of the study presented in this thesis, the novel modalities are voice and pen input.

The next concept, which is directly related to that of familiar modalities, is what we call the traditional interaction paradigm (TIP). The TIP refers to the use of mouse and

keyboard together to interact with an interface. As has already been mentioned, these modalities are the most frequently used in human-computer interaction, and we hypothesize that their entrenchment in current computer culture will play a role in how various modalities are used in a multimodal environment.

The third and final concept is that of *functional equivalence*. Two modalities are considered to be functionally equivalent if they can provide input that has exactly the same semantic and functional content, and if the input is processed in exactly the same way by the system, resulting in exactly the same system output. For example, pen and mouse are functionally equivalent to one another, as are voice and keyboard, since the user can type and say exactly the same words, the input is handled by the natural language processing components in the same manner, and results in the same output.

We now move to a more detailed discussion of how we plan to address each of the six questions posed above.

3.1 Use of modalities

How often are different modalities used, alone and in combination, for meeting browsing and retrieval tasks?

Understanding how different modalities are used in a multimodal input situation plays an integral role is assessing their usefulness. We intend to examine how often different modalities are used during interaction (as proportions of all interaction over a fixed period of time), whether their use is dependent on co-occurring modalities, and how their use evolves over time.

An important first step will be to observe how users use the traditional interaction paradigm of mouse and keyboard with a system, in order to establish a baseline. Next, we will investigate (in all modalities and modality combinations) if there are any trends in use that appear, such as consistently higher proportions of use of some modalities over others. This work will then be extended to cover trends that apply to specific pairs of modalities. For example, if there is a tendency to use specific pairs of modalities more than others. In particular, we would like to contrast the use of functionally equivalent modalities such as mouse/pen and keyboard/voice, to see if there are differences in how much they are used, and which other modalities might be influencing those differences.

Once general interaction trends have been established, we will focus on how the use of various modalities and modality combinations evolves over time. This is important in

order to determine whether experience with both modalities and the system itself affects how modalities are used. This evolution will be investigated by looking at two aspects. The first is whether the use of a single modality changes over time, and if it does, whether it does so in a similar manner across all relevant modality combinations. If this is the case, it suggests that the use of that modality was not dependent on other modalities. The second aspect involves comparing changes in parallel modality combinations. By parallel modality combinations we mean those where the difference between two modality combinations is a single modality – for example, mouse-voice and pen-voice are parallel combinations, since only the mouse and pen differ while voice is common between the two, while pen-keyboard and mouse-voice are not parallel since all of the modalities are different. We look specifically at the changes in relationships between modalities within a single modality combination and those that occur in parallel combinations.

Finally, we will look at whether there are any marked influences introduced by novel vs. familiar modality pairs. For example, are novel modalities used more often when they are combined with another novel modality than with a familiar modality?

3.2 Modalities and task types

Are some modalities more suited to finding certain types of information than others?

We are also interested in looking at whether there are any correlations between the use of a particular modality, or modality combination, and the type of task that the user is trying to solve. By the type of task we specifically refer to the type of information that the user is being asked to find while using the system. For example, are certain modalities more efficient for finding multimedia data as opposed to textual data. The presence of such correlations will have implications on future system design since it suggests that the choice of input modalities to include in a system will strongly depend on the types of data that the system is providing access to.

3.3 Task completion

Do certain modalities or modality combinations make the system easier to learn, leading to an increase in performance in the long term?

How well users perform a task using a system is an important measure of how useful the system is, and the input devices that are used play a significant role. Consequently, we

intend to determine whether certain modality combinations lead to an increase in performance when using a multimodal meeting browsing and retrieval system. We will use 3 different factors to measure performance. The first is what we call the *success score*. This is a normalized measure of how well questions are answered using different modality combinations. The second factor is how many questions are answered over a fixed period of time. This gives an indication of whether some modalities are 'faster' than others. The final factor is the *correctness score* which is a measure of how correct, on average, users were when answering questions. The results from these three measures, when taken together give an indication of which modality combinations, if any, increased performance.

Additionally, we will look at which modalities or modality combinations had the highest numbers of users who performed particularly well or poorly, both in terms of the number of questions answered and their success score. Data from questionnaires will be used to help control for external factors that might impact user performance such as the user's background.

Given the widespread use of the mouse and keyboard and the fairly short amount of time with which users had to familiarize themselves with the system, we expect that the TIP combination (mouse and keyboard) will be the most effective for solving tasks, at least during early stages of interaction with a new system.

3.4 Problems and modality choice

Does modality use change when a user encounters a problem during interaction?

During multimodal interaction, it is natural that users will change between modalities. During smooth interaction (when no problems are encountered), these changes, which we call switches, are likely to be driven by the nature of the particular system components that are being accessed at any given point in time. There has been quite a lot of work done by other authors as to which modalities are most likely to be used for certain types of interactions during smooth use – for example that pointing modalities are more likely to be used to select elements that are immediately visible on a screen, while voice is more likely to be used for those that are not, such as elements in submenus. We are less interested in this type of analysis, although we will use subjective user opinion to determine if modality use in a multimedia meeting browsing and retrieval system follows the trends established in the literature. We are more interested in whether users switch modalities when they encounter a problem during interaction.

We will first look at which modalities tend to produce the highest number of problems during interaction, and what the nature of those problems is. Then, we will look at the proportion of problematic interactions as compared to smooth interactions in different modality combinations before examining the proportions of switches that occur for problematic and smooth interaction for each combination. In the case of switches that result from problems in the interaction, we will also investigate the nature of the switch by looking at which modality is switched to. For example, in the case of failure using voice, does the user switch to keyboard input, maintaining natural language interaction, or do they switch to a pointing device.

3.5 User's perception of modalities

How do users perceive different modalities?

How users perceive and contrast individual modalities is an important factor in determining which modalities would be useful in a multimodal interface. We examine this issue using both subjective and objective data. The subjective data will be gathered in a post-experiment questionnaire, asking users to rank the usefulness of different modalities for accessing various functionalities of the system. This will help determine what users thought of the usefulness of the modalities, and whether their responses correspond to those found in the literature. The objective data will be gathered from interactions during the experiments themselves, and we will focus on how users perceive functionally equivalent modalities during interaction. We assume that if, for example, users who used the mouse used it in a similar way to those who used the pen, then the modalities are perceived to be equivalent, and therefore the choice of which to include in a multimodal interface is random. If the modalities are not perceived as functionally equivalent however, then the choice between them has much more significant implications on both modality choice and system design.

3.6 Learning effect

Does learning to use a system with a particular set of modalities influence how those modalities are used when other modalities become available?

It is reasonable to assume that users who are familiar with the use of a certain modality for interacting with a specific system will be more likely to use that modality with that system if it is available. Meanwhile, users who are not familiar with using that modality

with that system will be less likely to use it. We are interested in knowing whether this assumption in fact holds, and what the implication is for modality choice during system design.

Specifically, does learning to use a new system with a specific set of modalities influence how those modalities are used throughout interaction with the system, and in particular does their use change if other modalities are also introduced? For example, we expected that users who only have language input (and in particular voice) available when learning to use a system will be more inclined to continue to use language input even when other input modalities were made available. We feel that this is particularly important in determining whether users can be encouraged from the outset to use unfamiliar modalities and modality combinations, or whether their choice of modalities is entirely personal and not dependant on training.

We make two hypotheses regarding the appearance of a learning effect. The first is that there will be a learning effect in general for novel modalities, but that this effect will not be stronger than the influence of the familiarity of the TIP modalities. In other words, use of novel modalities will be higher than average in cases where the novel modalities are used to learn to interact with the system, but it will not be higher than the use of mouse and keyboard if those two modalities are available. The second is that users who answer a higher number of questions during early stages of interaction using a particular modality combination (or single modality) will continue to use that modality combination throughout the interaction while those who answer fewer questions will try to use additional modalities that are made available later in the interaction in an attempt to increase the number of questions they answer.

4. The Archivus System

> '*Where real improvement can be achieved by making major changes, the interface designer must balance the legitimate use of familiar paradigms, which ease the learning process, against the enhanced usability that can be attained by abandoning them.*' - RASKIN [96]

In order to be able to investigate the research questions on the use of modalities as discussed in the previous chapter, we needed to design and develop a multimodal interface for the meeting browsing and retrieval domain. The result was the Archivus system which was developed in collaboration with colleagues from the Artifical Intelligence Laboratory at the Ecole Polytechnique Fédérale de Lausanne[3]. In this chapter we discuss the various aspects that motivated its design such as the intended users and scenarios of use, user requirements, and the content of the multimedia meeting database. We will also provide a detailed description of the interface itself and its various functionalities.

4.1 Intended users and use cases

Normally, complex interfaces are designed with specific users and use cases in mind. Typically the users are either the direct clients of the developers or they are a group of people who the developer's clients represent. In either case, they are a fairly well understood group. Furthermore, there are usually specific scenarios in which the users will be using the interface, and specific tasks that they will be trying to accomplish with it. The design of the interface is thus geared to suiting those needs. Contrary to commercial development, research prototypes such as Archivus face the problem of having neither a well-defined set of users (clients) nor scenarios, which makes their development much more complicated.

4.1.1 Range of users

Meeting browsing and retrieval systems can be used to access any type of meeting (business or social) on any subject (research issues, business issues, decision making, discussion etc.); so the range of possible users is very broad. When developing the Archivus system however, we chose to concentrate on users in a business-oriented work environment who attend or are expected to attend meetings on a fairly regular basis.

[3] The author of this thesis was heavily involved in gathering use cases and user requirements for the design, deciding on the content of the database, determining the functionalities to include in the system and arranging the layout of the graphical components, but was not involved in the actual coding of the system.

These people can range in function from new employees of a company to managers and presidents, and can vary in ages from those just finishing school to those nearing retirement. Such a broad range of age and experience carries with it a potential for differing levels of familiarity and comfort with computers and technology, which poses challenges to the design of the interface and the degree of incorporation of novel modalities. Moreover, none of the users, no matter what their function, will be familiar with the domain, at least in the early stages of system use. The fact that the domain is entirely new to users, as well as the interface that will be designed (since not only is it a new application but also multimodal), places our intended user population in the *system novice/domain novice* category according to the user-type classification proposed by Dybkjaer and Bernsen [19].

4.1.2 Scenarios of use

Defining possible scenarios of use for a multimedia meeting browsing and retrieval system posed a number of problems due largely to the novelty of the domain. Methods such as task analysis [97] are not applicable because people simply have no experience with either the type of data available, nor the contexts in which it could be used. While activity analysis using techniques such as story-telling could have been used, we felt that the process would be too costly in terms of time in comparison to the types and amounts of information gained. Consequently, we were obliged to intuit the possible scenarios of use. This was done in the framework of the early stages of the IM2 project by Sire and Lalanne [98]. The result was a working assumption that a vast majority of the tasks that the system is expected to account for will be covered by the following five scenarios of use, where the first potentially overlaps with the remaining four.

Fact checking

Sometimes, a person remembers something from a meeting, but isn't certain whether or not they have remembered the fact correctly. Using a system such as Archivus, they are able to quickly find the relevant piece of information based on perhaps only vaguely remembered criteria, and double check that what they have recalled is in fact correct.

A manager tracking employee performance

Managers rarely have sufficient time to follow the performance of all of their employees on a regular basis, particularly if there are a large number of them. Never the less, managers are required to make decisions and suggestions based on the performance of their employees. One aspect of such decision-making can be seen as interaction with others and

participation in important decision making processes. Both of the preceding factors can partially be gleaned from interaction during meetings. And while a manager may be unable to attend all meetings, much less recall the actions and reactions of a particular participant, they can use a system such as Archivus to 1) pinpoint all interactions with and by a particular person, and 2) do this across a large number of meetings at once. Archivus would save the manager both time and effort as their required criteria would be entered into the system and relevant sections of meetings found and presented for viewing. The manager would then be free to view these meeting sections and gather the information they need.

A manager tracking project progress

Similarly to the case above, the same can be said of a manager tracking the progress of a project. For example, the system can be queried about all of the decisions made or discussions pertaining to a particular project, without having to view hours of unrelated or unimportant data from the remainder of the meetings, or perhaps more importantly, avoid having to review meetings that seemed relevant but in fact were not.

A current employee who has missed a meeting

A common problem these days is that someone misses a meeting, but needs to find out what happened in that meeting. Their main recourses are to read the minutes (which in many cases aren't sufficiently detailed), read documents that were used during the meeting (which lack context and information about what in particular was discussed) or ask a colleague who had attended the meeting (which often results in a very subjective view of how the meeting went). Given a system such as Archivus, the said employee is in a position to view the information that is important to them, from an impartial source, and have access to all of the information from the meeting, including all of the nuances that are lost in traditional catch-up methods.

A new employee who needs to learn about a project

It is often difficult for a new employee to learn about a project that they are going to participate in. Documents are read out of context, and often it is difficult to quiz colleagues about all aspects of a project, particularly if the project has been going on for a long time. A system such as Archivus can allow a new employee to explore different aspects of a project and the

decision making processes that went into it in their own time, in an objective manner, and without bothering others.

Having narrowed down the possible set of users and use cases that we had in mind for Archivus, the next task was to define in more detail the specific tasks that users might want to perform, and the functionalities that would be needed to help accomplish them. These are discussed in the next section.

4.2 User requirements study

> '*Even if the device is predetermined, for example, if we know the solution has to be a software program on a particular platform, working from user goals is necessary.*' – REDISH AND WIXON [97]

Authors such as Holtzblatt [99] and Norman [100] argue that it is imperative to perform in-depth user requirements analysis before system design can begin. But, as has been previously mentioned in the introduction, determining the user requirements for the Archivus system proved to be quite difficult. First of all, the domain in which we are working is new. There are no users who are experienced with searching and browsing in data of the type that is available in the Archivus system, nor are they used to being able to search at the semantic levels available via the annotations on the data. For similar and related reasons, it is hard to define the specific tasks for which a system like Archivus would be used. Finally, Archivus is being developed as a research prototype rather than as a business-oriented application, which means that there are no end-users available who have a vested interest in it. In order to get around these problems, we decided to use some preliminary user requirements that were gathered as part of an in-house brainstorming session within the IM2 project [98], in addition to a questionnaire designed specifically for this task. Even though Holtzblatt [99] believes that such user requirements studies are not particularly useful since users themselves often do not know what they want or how they will actually use a system, we found that it was the only option we had under the circumstances.

Our questionnaire (see Appendix A) explained to participants the types of data that would be available in the system (video, audio, text transcript, documents from the meeting) and asked them to imagine themselves in one of four situations: 1) a project manager following the progress of a project, 2) a project manager following the progress of an employee, 3) a new employee who needs to catch up on a project, or 4) an existing employee who has missed a meeting and needs to catch up on what happened in it. The

participants were asked to list the types of questions that they would pose to the system, or the types of information that they would like to find. The study involved 20 users, which according to Redish and Wixon [97] is enough to disclose a large percentage of key issues. The users came from a variety of different backgrounds and fields including administrative assistants, researchers and medical practitioners.

The results from this study, which are described in detail in [11], as well as the results from the brainstorming session mentioned above, were used to help define the functional requirements for the Archivus system.

4.3 Archivus backend database

The Archivus system was developed with the IM2 multimedia database in mind as the backend data store. Consequently, both its external and internal designs are tightly coupled with the content and structure of that database. In this section, we describe the content of the database and the rationale for choosing that content in particular.

4.3.1 Controlling for variables in the data

When conducting an experiment that is meant to test the applicability and usefulness of a piece of software for a wide range of users, it is important that the users don't feel like they themselves and their abilities are being tested. This is a particular risk for those users who are not very experienced with computers, or who are not at ease with them. This problem is compounded when the data with which the users are interacting is too complex. If the user feels uncomfortable with the data that they are accessing, they could feel intimidated by the testing scenario, resulting in a less valid testing result. This is particularly relevant in our case since we knew that we would be testing with users who had no vested interest and would likely not fall directly into any of the foreseen use cases. Therefore, we needed to control for two variables in the test set of data, topic neutrality/accessibility and cognitive load, to make sure that our users, who would not be the real users of the system, would not have more difficulties than could reasonably be expected, confounding experimental results more than necessary.

1. **Topic neutrality and accessibility**
   The topic of the meeting should be such that the average user is able to follow it easily, and ideally be able to relate to it on a personal level. This should help to maintain the interest of the user in the data and motivate them to continue using the system. Were the data to be too technical in any particular field, there would be a risk of alienating certain users who may not feel comfortable

with the domain or be able to follow the discussions, and thus would be less inclined to use the system itself and in particular be less motivated to complete the testing tasks to the best of their abilities. There is also the risk that if the user does not understand the data, they will expend too much effort on trying to understand the data rather than using the system to solve the tasks. This would mean that the results of the evaluation would reflect user satisfaction with the data in addition to their interaction techniques.

2. **Cognitive load**
   In order to allow the user to easily follow what is going on in the meeting, the meeting should be relatively clean in the sense that the people in it should be easy to understand, and that there should not be too much happening at once. While it is important to maintain a sense of naturalness in the meetings since the data should be as realistic as possible, meetings should be chosen that are at the same time natural, but for example have participants who speak clearly, and who do not speak simultaneously throughout large parts of the meeting.

In addition to these two factors, the database that users are accessing should contain a cross-section of the various data-types that the system is intended to give access to, and they should appear with enough frequency that users aren't forced into un-natural testing tasks just to determine whether a particular piece of data is accessible in a particular manner. Of course, the data must remain natural so meeting participants can only be explicitly asked to use specific data-types and generate particular events to a certain extent. To these ends, we have chosen the meeting scenarios described in the next section for inclusion in the Archivus database.

4.3.2 Recording scenarios

Three different meeting scenarios were used for the data in the Archivus database: room furnishing, a movie club meeting, and a meeting to determine the design of a remote control. All of these meetings can be accessed through the AMI project hub (www.amiproject.org).

Remote control design scenario

This meeting is one of a set of 4 meetings available on the AMI project hub, which deals with the design of a remote control. The participants each had a different role to play (project manager, industrial engineer, user interface designer, marketing expert) and were introduced to the scenario on the day on which the recordings took place. They were given individual

instructions by email, which were quite general in nature, allowing for some degree of freedom in the flow of the meeting.

Room furnishing scenario
This scenario is in fact a set of 4 meetings involving 5 co-workers (who appear in the meetings 4 at a time) whose task is to select the furniture for a reading room/lounge in a university department. The first meeting is an introduction to the problem and a request that each participant prepare a presentation of their ideas for the following meetings. The next two meetings are used to present and discuss ideas, while the final meeting is used to make a decision.

Movie Club meeting scenario
The movie club meeting involved 4 people trying to decide which movie to show at the next Movie Club screening. It includes a short introduction of what was shown at previous screenings, proposals from the meeting participants as to possible movie selections, the choice of the movie, and the choice of an advertising poster.

These last two scenarios had been explained to the meeting participants before the meetings. The participants had been given time to prepare their presentations, although they had been asked not to discuss their ideas with other participants ahead of time to allow for the natural introduction of variables and unexpected elements. The participants were told to act naturally, as they would in any other meeting, but to avoid frequent cross-talking. The meetings were not moderated and participants were free to act and react as they wished.

We feel that these scenarios are ideal because they pose no difficulty for the meeting participants in terms of the roles they had to play and in understanding the topics at hand, and similarly, they are simple and familiar enough for anyone testing the Archivus software to understand and relate to on some level. Our hope was this would reduce problems such as understanding the vocabulary used in the meetings, and overall comprehension of the topics being discussed.

4.3.3 The data set

The data set on which the Archivus system was tested includes 6 meetings (192 minutes of video data) recorded in English by a total of 8 different participants in the Smart Meeting Room at IDIAP [1]. The recordings captured audio, video, electronic copies of

all documents used (paper artifacts, slides etc), activity on an electronic whiteboard, and electronic copies of notes taken by participants during the meeting. The video included 3 room views (2 cameras on two participants and one on the whole room) and 4 individual views (a personal camera for each participant). However, for the experiments described in this thesis, only the whole-room view was used for the video stream, and whiteboard data was not included as it was unavailable at the time of development. The collected data was manually transcribed and then annotated with dialogue acts, topic segmentation, argumentative annotation and keywords. The raw data and the annotations were stored in the Archivus database.

4.4 Design rationale

In this section, we motivate the general decisions that were made in the design of the Archivus system. In particular, we discuss why we chose the input modalities used, why we chose a *flexibly multimodal* system and what this entailed, our choice for the underlying system architecture, the reasoning behind the choice of graphical components and their layout, and the system feedback mechanisms that we implemented.

Archivus was designed and developed with colleagues at the Ecole Polytechnique Fédérale de Lausanne as part of the Interactive Multimodal Information Management (IM2) project [8]. It was developed in a research context, with three distinct sets of research goals in mind. The first, which is addressed in this thesis, was to determine which input modalities are the most useful and appropriate for meeting browsing and retrieval. This required users to have quite a large degree of freedom in using the different modalities available, including standard mouse and keyboard as well as pen and voice. Allowing for voice as one input modality requires both freedom to control all elements of the interface using voice, and the freedom of expression (choice of vocabulary and grammatical structure). This implies that the speech recognition and natural language processing models cannot be too constrained.

The second set of research goals, which partially overlaps with the first, was to determine how natural language is and can be used in such an interface - specifically, the type of vocabulary and grammatical structures that were preferred. However, the cost of implementing and incorporating language processing and speech recognition modules is high, both in terms of time and effort. The cost increases even more when, in cases such as ours, the developers don't have a well developed corpus on which to base their work. As a result, Archivus needed to act not only as a tool for investigating how language was used, but also as a way to gather data for a corpus so that targeted recognition and

language processing modules could be developed. The third set of research goals was to determine which dialogue strategies are the most appropriate for multimodal interaction in this domain, which meant that the Archivus system not only needed to be multimodal, but that it needed to be a multimodal *dialogue* system. Neither of these last two topics are discussed here as they are the subjects of two complementary theses currently underway. The fact that the needs of the investigators of all three topics needed to be taken into account in the design and implementation had significant effects on the resulting system.

In order to meet all of these needs and still keep implementation costs reasonable, we decided to take the approach of developing Archivus as an evolutionary, high-fidelity prototype. Evolutionary prototypes, as described by Dix et al. [9] and Norman [101], are prototypes that evolve into actual systems over time. In our particular case, Archivus was also a high-fidelity prototype in the sense that Archivus was (and is) not a fully functioning system. While many of its key components and functionalities are already in place so that preliminary evaluations could be carried out, other components, such as the speech recognition, natural language processing and dialogue management modules are missing. These modules were simulated during experiments with the system, using the Wizard of Oz testing methodology which is addressed in detail in Chapter 5. The modules will be developed and incorporated into the system over time as data gathered during early user-studies with the system is analyzed and used to define them.

A final general point to make about the Archivus system before we go into detail about specific topics is that it was conceived as an application that allows for both searching and browsing, or a blend of the two, which we think gives users a sufficient degree of flexibility in accomplishing their tasks. As Tricot explains in [102], access to large databases is underexploited because the structure of the interfaces to them encourages only linear search, whereas a hybrid approach that blends browsing and directed search would be more effective.

In designing and developing Archivus, we have tried to keep as much as possible to the software engineering and user-centered design principles outlined in [9, 96, 100, 103]. We will not go into details here, except where we feel a design choice could directly impact the experimental results in this thesis.

In the following sections we motivate the system design choices that were made. The first three focus on the input and architectural aspect while the final three focus on the graphical user interface. We conclude the chapter by presenting the system itself, and explaining the types of tasks that can be performed with it.

4.4.1 Modality choice

As previously discussed, the choice of which modalities to include in an interface depends on many factors including the intended user population and the tasks for which the application will be used. In addition to these criteria, we add that of realism. Working in a research environment affords us the opportunity to explore new technologies such as gesture and facial expression recognition, but these technologies are not sufficiently advanced yet to be included in interfaces that would be usable within the next few years. Consequently, we have chosen to root our input modality choices in what we thought would be most appropriate and realistic modalities given the envisioned context of interaction in the near future.

We imagine a system such as Archivus to be used primarily in a typical office environment on a standard desktop or laptop computer. Therefore, any input devices considered had to be suitable for those environments. We decided to use mouse and keyboard for the obvious reason that they were the input devices that users were most familiar with. Additionally, they help to establish a baseline for general evaluations of the system. Given the nature of the data in the database, we also wanted to give users a semantically based means for accessing the data. While this could be achieved via keyboard input, we felt that voice interaction, which is becoming increasingly common in various applications, would also be a viable option. In pilot studies, the fourth modality that was used was a desktop touchscreen monitor. This option was abandoned in subsequent studies because users found it too inconvenient and unnatural to use. The touchscreen was replaced by a tablet PC with pen-based input. In view of many work environments becoming increasingly mobile, we felt that this was a more appropriate choice and wanted to explore how pen-based input might affect interaction.

4.4.2 Flexible multimodality

The Archivus system was designed and developed to be flexibly multimodal. This means that the user can use any of the modalities available, at any time, to perform any action. The motivation behind this choice was that one of the primary goals of the system was to be able to explore how users use the various modalities available, and in particular if there are preferences for certain modalities while performing specific types of tasks. In order to accomplish this, we needed to design the system with as few *a priori* assumptions as possible about potential interaction patterns. Designing the system to be flexibly multimodal allowed us to do this. Lai and Yankelovich [29] warn that giving users more flexibility will increase the number of errors produced while interacting with the system. However, Oviatt [68] and Rudnicky [27] argue that this is in fact not the case,

and that users are quite good at making appropriate input choices to suit their interaction strategies. Moreover, Oviatt [38, 42] argues that in fact, multimodality can reduce the number of errors produced because more robust modalities can be used to compensate for the problems that might be introduced by weaker modalities.

4.4.3 Rapid Dialogue Prototyping Methodology (RDPM)

As Archivus was intended to be a dialogue-based system and was being developed in collaboration with the Artificial Intelligence Laboratory at the Ecole Polytechnique Fédéral de Lausanne, the Rapid Dialogue Prototyping Methodology (RDPM) [104], which was developed in the lab, was used as a backbone for the architecture of the Archivus system.

The RDPM is a generic platform that allows developers to quickly specify a dialogue strategy for human-machine interaction and tailor it to a desired domain. The resulting dialogue helps end-users of a system express relevant search criteria by guiding them in such a way that the type of information that the system needs in order to find the information that the user is seeking is gathered in an efficient manner. The methodology uses criteria specified as attribute-value pairs and a dialogue model composed of instances of two types of nodes. The first type of node is application-specific and is used specifically to help the user select relevant attribute-value pairs. The second type of node is application-independent and focuses on the dialogue flow and management of the system.

The RDPM was originally conceived for voice-only systems and had to be extended to allow for the processing of multimodal interaction. Since the Archivus system was intended to be flexibly multimodal, each vocal action needed to be translated into its corresponding graphical component (similarly to work done by Haddock in [105]) , and each graphical component needed to be manipulable using voice. The resulting extensions involved the addition of multimodal prompts (since output now needed to be not only vocal but graphical as well), more sophisticated grammars for natural language input processing (since a wider variety of actions was now possible), and graphical representations of each application-specific node had to be developed. A detailed account of how the RDPM was adapted for a multimodal environment and its application in the Archivus system can be found in [106].

4.4.4 Archivus metaphor

Due to the fact that the Archivus interface is intended to be used in a domain with which users are generally unfamiliar and is intended to include input modalities which users

may not have been exposed to before, we decided to apply an interaction metaphor in order to facilitate user understanding. Several authors [71, 107-109], suggest that the use of a metaphor can help users understand and build mental models of a system with which they are unfamiliar. This knowledge can then be used as a foundation on which they can build their understanding of new or less familiar functionalities. However, Constantine [108] also points out that the metaphor must be chosen carefully, and used only in as far as its elements are useful for facilitating understanding. Stretching a metaphor too far can actually interfere with a user's understanding if the implementation of the metaphor suddenly contradicts users' expectations.

To this end, we chose the library metaphor for the Archivus system. The notion of a meeting recorded and stored in multimedia form, including spoken or written content (body text) with visual elements (pictures) and attached documents (appendices), as well as its overall structure (meetings topics as book chapters, subtopics as subchapters, keywords as an index etc.) is conducive to representation as a book. Moreover, companies storing large numbers of meetings will require some sort of organizational system for them, in which case the analogy to books in a library, with constrained ways in which to arrange them, also suits the metaphor.

We are not the first to have used books as a metaphor. Card et al. [110] used the book metaphor in their work focusing on representing the World Wide Web as books, where each book could be a webpage and all of its associated links, or a group of several topically related pages. They propose methods similar to ours for manipulating books (going back and forth through pages), but their implementation is more graphically advanced than that of Archivus. Ozsoyoglu et al. [111] use the book metaphor in an application which constructs lessons from multimedia data. They include topic hierarchies in the form of keywords, much like we do in Archivus.

4.4.5 Graphical components and layout

Our choice of graphical elements and their layout on the screen was based both on user requirements specific to the Archivus system, and on general principles in interface design such as grouping together of functionally similar elements and visibility, described by Norman in [100].

Visibility, according to Norman, is a crucial element of design since objects that are visible on the screen help remind users what functionalities are available in the system. However, it is also important not to overcrowd a screen, as this might confuse users, or split their attention unnecessarily. In [110], Card, Robertson, and York note that some

information-based interfaces exhibit what they call a '*cost-structure of information*', where '*a small amount of information is organized to be available at a low cost, larger amounts are available at moderate costs, large amounts at high costs*.'

Following these guidelines, we designed the Archivus system in such as way that all of the static principle objects such as an overview of the database, the user's interaction history and the search criteria buttons are always on screen, as are elements such as the keyboard input bar, the text version of system prompts, and system control buttons. The only item that changes (in content but not location) is the principle pane in which the results of user searches and browsing are displayed. As the number of these elements was manageable, it was possible to make them all available directly on the screen at low cost. A positive side-effect of this was that it eliminated the need for menus.

We also tried to maintain some of the layout factors that we felt would be most engrained in users' implicit knowledge. For example, the overview of the database, as well as the user's interaction history, are kept on the left side of the screen, much as histories are in internet browsers, or menu frames on web-pages. Finally, Tricot [107] suggests that providing indexes or tables gives users an overview of the information content, which in particular helps casual users, or browsers who may not be sure exactly what they are looking for. We felt that with the nature of the Archivus database, where there are many meetings and where topics change often within a meeting, such a mechanism would be quite useful, so we included Tables of Content in the representations of the meeting, which were composed of a hierarchy of topics from that particular meeting.

A detailed description of the components of the system and their specific functionalities is presented in section 4.5.

4.4.6 System feedback

The final factor to consider was how and when the system should provide feedback to the user. Authors such as Dybkjaer and Bernsen [19] argue that appropriate feedback is crucial to successful human computer interaction and that not only does an interface need to provide feedback, it must do so in such a way that it is noticed by the user but does not interfere with their work.

Feedback in the Archivus system happens at two levels. The first level is purely visual. It shows state-changes when a button has been pressed, or signals that the system is processing input or searching for information. In Archivus, the former is handled by the

appearance of the button subtly changing, and the latter is handled either by an inactive (greyed-out) screen, or a progress bar. Additionally, any global changes to the system which are a direct result of the addition or deletion of search criteria are highlighted to the user. For example, meeting books change colour and move slightly when they first become relevant. Since the bookcase (the representation of the entire database) is off to the left side of the screen, a user might not notice a change if they are focusing on the central panel. If the books move slightly, the action is more likely to catch their attention.

However, highlighting information is not the only mechanism that is used in Archivus to provide feedback to the user. Brewster [112] notes that '*Users can choose not to look at something but it is more difficult to avoid hearing it. This makes sound useful for delivering important information*'. Since Archivus is a dialogue-based multimodal system, it also allows for spoken output, which we use in the form of dialogue prompts to either inform the user of a state-change or encourage an action. Studies such as those by Goodman cited in [78] show that using speech to convey information is helpful to users, but as Le Bigot, Jamet, and Rouet [53] point out, language use also increases cognitive load in a way in which text does not. Therefore, we designed Archivus to not only speak the prompt, but to display a text version of it as well. Since the processing of speech and hearing channels are independent of one another [112] including both forms of system feedback allows the user to choose the form that is least straining for them given the task that they are doing.

Having made the decision to use voice, another consideration is choosing the right voice. Moore and Morris [50] hypothesized that the quality of the output voice would make a difference in the users' perception of system capabilities. Low quality voices imply poor capabilities and vice versa. In pilot experiments, we used a text-to-speech system that resulted in a tinny, synthetic sounding voice. The result was that users in most cases ignored what it was saying entirely either by turning it off or interrupting the prompts. In the final experiments, we opted for higher quality text-to-speech, which was much more readily accepted by users.

We also had to choose the voice itself. Dahlbäck et al. [113] have shown that the interaction between a human and a computer changes when the vocal output produced by the computer matches that of the human interlocutor in accent. This means that for example a non-native English speaker, whose first language is French, will feel more comfortable interacting with a computer system whose vocal output synthesizes English spoken with a French accent than they would interacting with a system that synthesized English with a British accent. They attribute this factor to the similarity-attraction effect

which states that '*we prefer to interact with personalities that resemble our own*'. While the similarity-attraction phenomena may play an overall role in how users perceive interaction with Archivus, we believe that it will not play a crucial role in influencing the results in this work since almost all of the users will be non-native English speakers, and among the English speakers only a few will be North American English speakers. Consequently, almost all of the participants in the experiments will have a similar degree of dissociation with the Archivus vocal output language. Finally, we chose a female voice, since most on-the-market text-to-speech systems use a female voice.

4.5 Description of the system

In order to better understand the nature of the experiments described further on in this work, and to situate the context of this research in a more specific framework, we present here a detailed explanation of the Archivus interface and its various functionalities.

4.5.1 What can be done

The Archivus system can be thought of as a virtual librarian that helps users find information that is contained in meetings that have been recorded in specially equipped meeting rooms. Meetings that take place in these rooms are recorded in video and audio form, and are later transcribed so that a text form of the entire meeting also exists. This text is also analyzed and annotated with information such as who was speaking, the topics that were discussed, the parts of the meeting in which decisions were made, etc. The major tasks that the user can perform using Archivus are:

- find meetings, parts of a meeting, or specific information in a meeting based on criteria such as 1) the date, location or participants in a meeting, 2) the topics covered, keywords spoken, or documents used or referred to in a meeting, or 3) the dialogue acts (i.e. questions, statements, etc.) or argumentative sections (discussions, arguments, etc.) in a meeting
- get an overview of all meetings in the database that are relevant to a user's goals
- browse quickly and easily through only the meetings or meeting sections that are relevant to the user's goals
- view and browse through documents from a meeting
- watch video, listen to audio or read the text transcript of a meeting
- browse through any meeting without specifying search criteria

- customize the organization of the entire database of meetings based on one or two criteria, for example by date and speaker

Additionally, the Archivus system has been designed to be conversational, which means that when it can, the system will try to help the user determine which information is needed so that Archivus can find the information that the user is looking for.

4.5.2 How it can be done

In his section, we explain the graphical components of the Archivus system and the types of functionalities that they give access to. Figure 1 shows the Archivus interface as it appears after the search criteria '*Which article did Susan suggest at the meeting in Geneva?*' have been specified. The various parts of the interface are explained in more detail below.

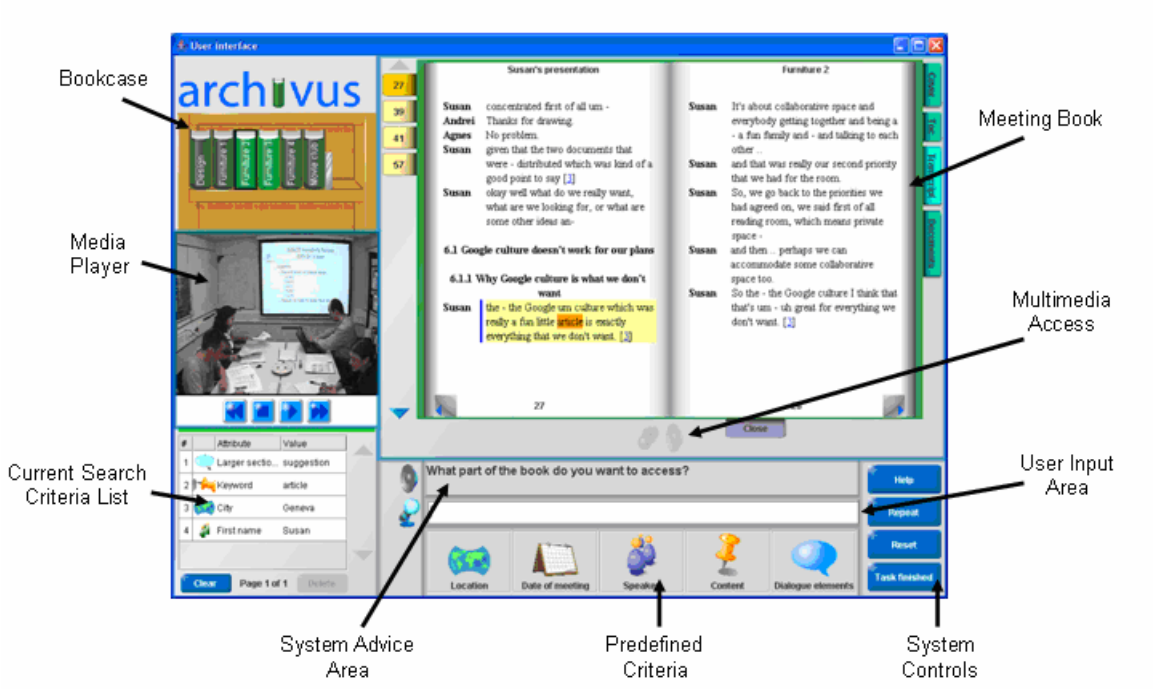


Figure 1: The Archivus interface

The bookcase

Each meeting in the Archivus system is represented as a book, and the bookcase, located in the upper left corner of the screen, contains all of the meetings that are available in the system. Books can be sorted in order to help the user browse them more easily. This can be done by changing the labels on the bookcase, a function accessible by clicking on the buttons that appear towards the bottom of the bookcase. There are two such buttons - one for specifying the label for the legs of the bookcase,

and another for the shelves. By default, no labels appear when the system is first started. If the books are spread across more than one bookcase, arrows will appear near the top and bottom of the bookcase which can be used to move between the bookcases. If the user has specified some search criteria, the books change colour – some become light green and others dark green. The light green books (called '*active*' books) are those that contain information that matches the search criteria. The dark green books (called '*inactive*' books) are those that do not contain any information relevant to the search.

System prompt and query input areas

A text version of the audio prompt appears in this area, for cases where the user has turned off the output sound, or wants to have a visual reminder of the last thing that the system said to them. Just below the system prompt area is the query input field, in which users specify their typed queries. Moreover, the user can turn off voice input to the system by selecting the microphone icon to the left of the query input bar. The icon appears with a red cross over it when voice input has been turned off.

Interactive browsing area

The interactive browsing area is the central pane of the system, and can contain different items depending on what the user is currently doing with the system. Most commonly, it displays selection options resulting from the activation of a criteria selection button, or a book that has been opened.

The book (Figure 2) has several components that help the user browse the meeting and see the results of their search. The main part of the book shows the content of the meeting. The name of the person who spoke a particular phrase appears in the margin, and what they said appears in the main body of the page next to their name. The section of the page that is relevant to the search criteria specified is highlighted in yellow.

The hit tabs shown in yellow indicate the pages where Archivus found results that meet the search criteria that have been provided. The up and down arrows above and below the tabs can be used to move between them, but the tabs themselves are not clickable. The currently visible hit tab is a darker yellow.

The content tabs take the user to various salient sections of the book. The cover tab gives access to the cover page, which contains information such as the date of the meeting and the participants. The content tab takes the user to the table of contents,

i.e. the topics of the meeting. The documents tab takes them to a list of all of the documents that were used and/or discussed during the meeting. The tab that has been selected will be a slightly darker colour than the others.

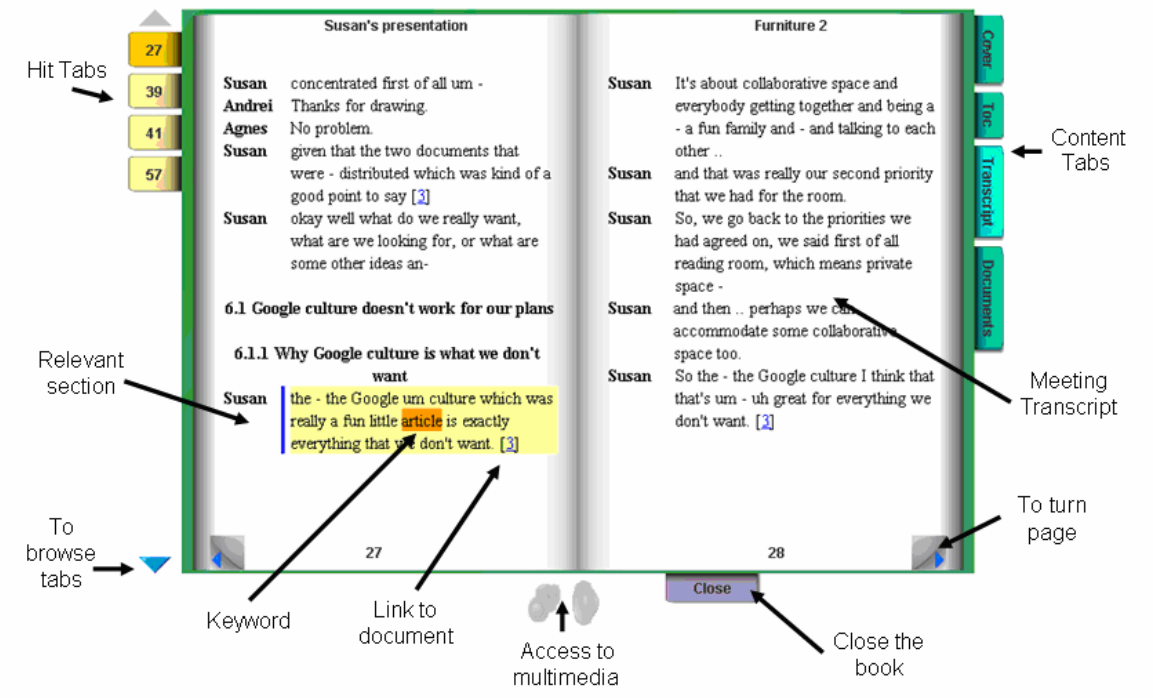


Figure 2: The Archivus book

Multimedia such as audio and video can be accessed directly from the book by selecting either the  icon to play the video or the  icon to play the audio. The media will appear in a media player in the middle left side of the screen and can be controlled using the standard video-type control buttons that appear in the player. Selecting the stop button will close the media player. Moreover, the media player is tuned to start the media at the page from which it was selected, giving users easy and quick access to very specific points in the meeting.

The user can also browse through the book by paging through it using the arrows located on the bottom outside page corners. The book can be closed by selecting the close tab, or will close automatically when criteria that don't match it are specified.

Interactive history

The interactive history helps users keep track of the search criteria that they have specified during a particular interaction. These can either be viewed by scrolling (if there are many), or removed from the list, which redefines the search.

Criteria refinement buttons

The criteria refinement buttons serve as shortcuts to various categories of information that user requirements studies have shown to be useful for meeting browsing and retrieval, such as specifying the location of a meeting  , the date on which it was held  , or the participants of that meeting  . The content  button provides direct access to content related information such as topics, keywords or documents while the dialogue elements button  , provides access to more detailed linguistic annotations on the data such as dialogue acts and argumentative segmentation.

System buttons

The system buttons provide general control over the system such as access to help, repeating a prompt that the system has just played, resetting the system (which effectively clears all search criteria), and the task finished button which was added solely for the purpose of the experiments described in further sections of this thesis.

Finally it is important to note that following the suggestions of Sutcliffe [58] regarding the design of multimedia interfaces and how to influence what users look at, several features have been added to the interface whose intention is to draw the user's attention to areas of the book that have changed or that the system feels are directly relevant to the user, such as relevant books briefly moving in the bookshelf, or green-blue boxes appearing as borders around the areas of the interface that are directly important for the user's current search.

4.6 Conclusions

In this chapter we presented the Archivus system. We began with a discussion of potential users and their requirements, and the data that was available for the Archivus backend database. We then went on to discuss the motivations and reasoning behind the design choices that were made, finishing with an overview of the possible tasks for which Archivus can be used, and a detailed description of the system itself. In the next chapter, we discuss the user-driven experiments that were performed using Archivus.

5. The Experiments

In this section we describe the experiments that were developed and executed in order to gather data for the work in this thesis. We begin with a discussion of the methodology used, and how it was extended and applied to the Archivus experiments, then discuss the particular environment that we used for the experiments, give a detailed account of the experimental protocol, and finally describe the types of data that resulted from the experiments.

5.1 Wizard of Oz methodology

Since the field of multimodal interaction is quite new, there are few established techniques for evaluations involving multimodal interfaces. In fact, only one methodology for experiments and evaluations of multimodal interfaces involving language use has been consistently mentioned [114] - high fidelity simulation (HFS), more commonly known as Wizard of Oz (WOz) experimentation [4, 34, 39, 61]. WOz is a technique that has primarily been used for the investigation of natural language interfaces [60, 61], but is becoming an increasingly popular tool for the investigation and development of multimodal interfaces [62].

Wizard of Oz experiments are designed in such a way that a user interacts with a system which has not been fully implemented. However, the user himself is never aware of the incompleteness of the system while they are interacting with it, as missing functionalities or components are simulated by a human wizard who monitors the interaction from a remote location [39, 60-62, 115]. The key benefit of this type of experimentation is that it allows for the investigation of the use of different modalities, modality combinations, technologies and functionalities in practice before significant amounts of time and effort have been expended on their full implementation in an interface or a system.

Moreover, the Wizard of Oz technique can be used as both a requirements gathering [31, 38, 61], and evaluative tool [19, 38, 39, 61, 62, 116], which is ideal for the needs of the Archivus experiments. In the first case, a Wizard of Oz experiment can be used to determine the types of expectations that users have of a system by developing only a basic version of it and having the wizard simulate as many functionalities as is necessary. Analysis of user interaction with such a system would indicate which functionalities are expected and in which particular situations. Using the WOz methodology, these functionalities can be elicited with minimum frustration for the user since unexpected actions that would be fatal to a fully implemented but underspecified system can be

compensated for by the wizard. In addition to interaction functionalities and modality combinations, requirements for a specific modality can also be gathered using Wizard of Oz experiments. For example, the vocabulary and language model for a natural language processing component can be continuously developed and refined as testing progresses, since each test provides a potentially new set of spoken interactions [117]. In the evaluative case, the Wizard of Oz technique can be used to validate or test hypotheses about particular aspects of design before significant amounts of time are expended on their development [38], and can provide insight into human behaviour when faced with a given system [35, 60].

Many informal guidelines and platforms for the design of Wizard of Oz studies exist [4, 34, 38, 60, 61, 115, 116, 118, 119], in particular for speech-only systems. From these, there are three key points that all authors agree are crucial to successfully executing Wizard of Oz studies. These are that:

1) wizard reaction times must be quick, and actions consistent - this is necessary in order to convince the user that they are interacting with a real system. Users are accustomed to computers reacting very quickly to their input, particularly in the case of simple actions such as button clicks, but also in situations where searching might be involved, such as internet and database searches. If the wizard responds too slowly, the user might become suspicious or even worse, believe that the system is simply performing poorly. Wizard actions need to be consistent for similar reasons. We are used to computers reacting in exactly the same way if the same action is specified under the same circumstances multiple times. If there is a difference in how the system reacts, again, the user is likely to become suspicious of the system.
2) the prototype must be fairly solid to begin with – even though it may only be a prototype, the system must behave as if it were a real system. If there are too many 'bugs', it becomes hard to determine which user actions and reactions are problematic due to the system being unstable, and which are due to legitimate problems with the design.
3) pilot experiments need to be run – the complexity of multimodal systems and the resulting environments in which they are tested are such that several runs of pilot experiments are likely to be necessary before a sufficiently complete environment and a strong evaluation protocol can be established.

5.1.1 Extending the Wizard of Oz methodology for multimodality and Archivus

The complex and diverse nature of multimodal interaction means that WOz experiments for multimodal applications need to be tuned to the specific needs of the application being studied and the modalities that are being tested [61]. Since part of the research involving Archivus was intended to explore the use of language, it was known from the early stages that Wizard of Oz experiments would be used to simultaneously evaluate the interface and modality use, and gather data for the development of speech recognition, natural language processing and dialogue model modules. Consequently, facilities to perform WOz experiments were built directly into the system. The interfaces which the wizard used were based specifically on the Archivus screen elements, the structure of the database and the actions that the wizards would have to perform. In fact, for the Archivus system, we developed two wizard interfaces – one to process user input, and one to determine the appropriate conversational prompt to provide as output.

Pilot experiments with our Wizard of Oz environment and the Archivus system revealed two important points. The first was that users were highly unsatisfied with the generic and highly repetitive dialogue prompts that the system was providing. In post-experiment questionnaires and interviews users told us that in most cases, the prompts were not helpful because they did not take into account enough of the context of interaction. This implied the need for the wizard to be able to control and dynamically change system prompts based on context, since it was unclear at the time which types of prompts would be needed in which situations, making automation of the process impossible. We handled this using an approach similar to that of Pirker, Loderer and Trost [35], where system output is controlled by the wizard through a set of predefined but editable prompts. The second point was the wizard who simulated the natural language processing, called the input wizard, was working under a high cognitive load, and that the interface they were using to do the simulation needed to be improved. This was accomplished by increasing the usability of their interface via layout changes and faster database access techniques.

However, it was clear that even with an improved interface, the input wizard would not be able to handle the additional task of dynamically changing the system prompts. So, we decided to split the tasks between two wizards. The input wizard worked as before, processing user input and retrieving information from the database accordingly. The system prompts however, were handled by a second wizard, called the output wizard, who worked in sequence with the input wizard, basing their prompt selection on the decisions made by the input wizard. Work by Salber and Coutaz [60, 61] has also shown

the need for a multi-wizard environment, particularly in experiments with multimodal interfaces where the complexity of the tasks being accomplished and the need to fuse the use of various modalities causes a higher cognitive load for the wizard.

A more detailed description of the extensions to the Wizard of Oz methodology and details of the wizard's interfaces can be found in [120].

5.2 Archivus Wizard of Oz environment

In this section, we describe the physical Wizard of Oz environments that were used both in the pilot studies, and in the final experiments.

5.2.1 Pilot experiment environment

As Cheng et al. [121] did, we initiated a lengthy pilot experiment which served to fine tune both the Archivus software and the Wizard of Oz environment and to provide training for the wizards. The experiment involved 24 users in 8 different modality combinations. The environment in which those experiments took place was as follows:

The user's room (Figure 3) contained a desk and chair, as well as a standard desktop computer with a 17 inch 3M touchscreen. The user had access to a wireless keyboard and

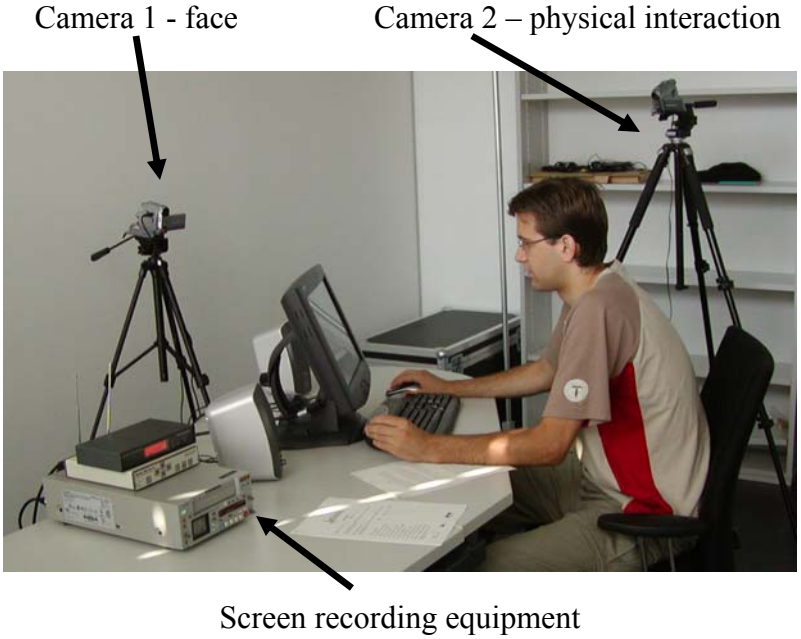


Figure 3: User's room during Wizard of Oz experiments

mouse, as well as a lapel microphone. Two cameras were placed in the room. The first, which was located directly behind the computer screen, recorded the user's face as they interacted with the system. The second, located behind and to one side of the user, recorded their physical interaction with the input devices. In addition to the cameras, the user's screen was also recorded, and all user actions and system reactions were logged.

In the wizard's room, we had two computers with monitors, plus an additional monitor. One of the computers streamed the video from the face camera so that the wizard could see how the user was reacting to the system. The other computer was used by the wizard to simulate interactions. The additional monitor was used to stream a view of the user's screen, which gave the wizard a more complete picture of what the user was doing, which in turn allowed them to react in a more appropriate manner.

5.2.3 Final experiment environment

In the final set of experiments, much of the environment remained the same, with two notable exceptions. The first was that we removed the touchscreen, replacing it instead with a 13 inch tablet PC. The motivation for this was that tablet PC was necessary to investigate modality combinations with pen input, and since we wanted to keep as many variables as possible constant, we chose to use the same screen for all other modality combinations, removing the possibility to use pen input (and other input modalities) as appropriate. The second change involved the addition of a post for the second wizard. The new wizarding room, shown in Figure 4, now contained 3 computers and an additional monitor. Two of those computers and the additional monitor served the same

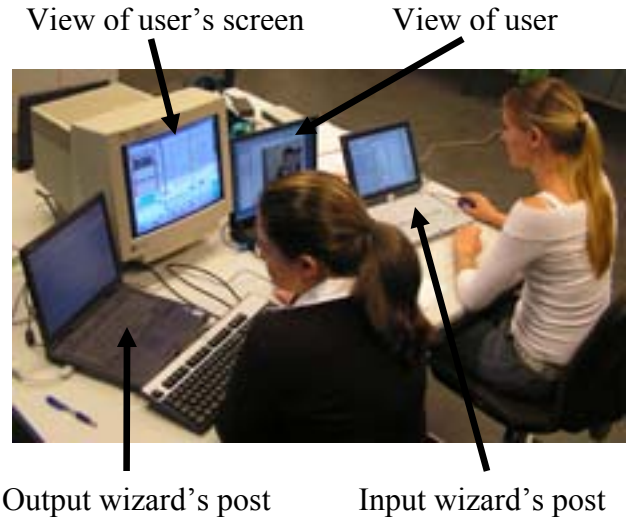


Figure 4: Wizard's room during Wizard of Oz experiments

purpose as those in the pilot experiment, and the third computer served as the output wizard's post.

5.3 Modality combinations

The Archivus system is designed to accommodate a number of possible input modalities – voice (V), mouse (M), keyboard (K), and pen (P). In our experiments, the pen was used strictly as a pointing device, and not as a tool for natural language input. In order to determine which functionalities and in which combinations give the most added value to the system, we wanted to systematically test all possible modality combinations and compare their performance on the same task. However, we felt that of the 15 possible combinations (listed in Table 3), 5 were not interesting or not feasible (in grey).

MVK	PVK	MVPK	MP	M
MV	PV	MPK	VK	P
MK	PK	MPV	K	V

Table 3: All possible modality combinations for WOz Archivus experiments

For the experiments, we chose to eliminate:

- K - keyboard alone is too archaic
- any combination involving MP as the mouse and pen functionalities are too similar and redundant

One might argue that voice and keyboard are similarly redundant to mouse and pen, but we feel that the novelty of voice interaction and the tendency for users to use the keyboard more as a keyword input tool rather than a full-scale natural language input tool outweighs the similarity.

Finally, we decided that the experiments would be conducted between subjects [115], with each subject using a system that tested only one modality combination. The reason for this was that it would be more difficult to control across users how much of the interface the user had learned or been exposed to with a previous modality combination.

5.4 Experimental protocol

We were faced with several challenges when selecting and developing the evaluation protocol. The first was choosing which scenario to use in the evaluation. Dahlbäck, Jönsson and Ahrenberg [34], Dybkjaer, Bernsen and Dybkjaer [39], and Dumas [115] argue for the importance of choosing an appropriate evaluation scenario, since the

phrasing and topics of a scenario can have a significant impact on how a user later interacts with a system. A scenario that is too specific runs the risk of users mirroring behaviour or tasks presented in the scenario in their own interaction, which limits the amount of useful information that experimenters can then extrapolate. Conversely, a scenario that is too broad runs the risk of not giving experimenters enough statistically significant data about particular behaviour since user actions can be too divergent. Consequently, the scenario must be carefully chosen to pinpoint the types of information that the experimenters want to investigate, while being broad enough to give users some degree of freedom of expression when interacting with the system. We remind the reader that the goal of the experiments was to investigate how users would use modalities to interact with a multimedia meeting browsing and retrieval system, and asses the added value that multimodal interaction would bring.

Our primary concern was that given the novelty of the domain, it was not clear what the most important tasks would be, nor the most appropriate way to present them. The Archivus system had been designed from the outset with 5 general scenarios of use in mind (section 4.1.2)

1. fact checking
2. a manager tracking employee performance
3. a manager tracking project performance
4. an employee catching up on a meeting that they should have attended but missed
5. an employee learning what has been going on in a project that they are previously not aware of

Since the volunteers for our experiments would not be 'real' users of the system, we decided that situation 1, fact-checking, was the most appropriate context for our evaluations as it required the least emotional investment from the users and allowed us to test interaction with a larger number of functionalities over a short period of time. Fact-checking implied verifying statements about the data, or answering simple questions about its content such as '*The budget for the room furnishing was 1000CH.*'or '*Who was leading the design meeting?*'.

The second challenge was that the results had to meet the needs of several interdependent research goals. One of these was looking at the types of dialogue strategies that would be needed. Another was to examine the nature of the natural language use (vocabulary, grammar, etc.) in order to develop robust natural language processing capabilities. Finally, we wanted to examine which of the proposed input modalities users would find most useful, both overall and for specific tasks.

We also had several practical constraints to take into consideration – specifically, limited time and manpower for running the experiments with large numbers of users. This meant that the experiments had to be as short as possible (2 hours per user) and would not require users to come back for several sessions over a period of time.

Finally, and perhaps most importantly, since the system was developed to have as few *a priori* assumptions as possible about interaction, we wanted to leverage that freedom in the evaluation. Consequently, we wanted to minimize the amount of influence that we had over how users would use the input modalities available and the interaction patterns that they would develop to find information in the system. This turned out to be the most challenging and difficult task to accomplish.

5.4.1 Choosing a protocol

We found that existing descriptions of step-by-step experimental procedures such as those described by Oviatt et al. [122] and Strum et al. [123] were very useful as a point of departure, but did not satisfy many of the particular needs that arose in the case of Archivus. Instead, we considered and eliminated several possible protocols, some of which are discussed in the following paragraphs, before settling on the one that we believed best suited our needs.

Free exploration, where the user 'plays' with the system on their own and without any instruction for a fixed amount of time, was eliminated since we believed that users would not discover all of the functionalities of the system, nor would they try sufficient degrees of freedom with the various novel input modalities to recognize the benefits that they could provide.

Structured exploration, where the user would be given a manual for the system and would be able to interact with it at the same time was also rejected. While in this case we could ensure that the user was made aware of all of the possible functionalities of the system, we were still left with the risk of them not being willing to explore using the novel modalities on their own.

A protocol that seemed plausible and would provide a solution to the two problems presented above was to have the experimenter do a guided tutorial with the user, where they show the user both the different functionalities of the system that are available and how to access those functionalities using different modalities. This would ensure that the user sees all of the functionalities of the system and also a variety of ways in which to

access them. Moreover, there would be no paper trace of the interaction patterns presented by the experimenter, so the user would find it more difficult to emulate the patterns in their own interactions.

However, we did find one significant drawback to this protocol – it seemed to be impractical for large-scale evaluations. Since we were planning to run an evaluation with over 80 users, it would be difficult for the experimenter to ensure that each user was given the same amount of information in the same manner. Moreover, variables such as social factors and experimenter fatigue would also have to be controlled for.

Finally for the pilot experiments we opted to give the user a detailed manual which highlighted all of the possible functionalities of the system but without providing explicit examples of how they could be used, as we might have with the structured exploration protocol. However, the user would not be allowed to interact with the system until it came time to do the actual evaluations. This was a conscious choice motivated by the fact that this way, there would be no 'priming' or favoritism for specific modalities before the experiments began. All users would have the exact same level of knowledge about the system and there would be no established interaction patterns that the user could follow. The detailed protocol is given in the following section.

5.4.2 The pilot experiment

The pilot experiment involved 24 users who were first given a demographic questionnaire to fill in and a consent form to sign. Then they were given an explanation of what the evaluation was about – specifically, they were asked to pretend that they had just been hired by a company and had been asked by their manager to do some fact finding and checking using Archivus. Finally, they were given the detailed manual to read but were not allowed to interact with the system itself.

Once they had filled out the forms and read the documents, they began the two evaluation sessions, during which they had 20 minutes per session to answer a series of true/false and short answer questions using Archivus. We henceforth refer to these sessions as phase 1 (P1) and phase 2 (P2). In the first phase, users had access to a subset of all available modalities. Results from P1 would allow for comparison of the performance of specific modalities and modality combinations to one another. In P2, users had access to all available modalities, which provided data on whether users had a preference for certain modalities over others. The final part of the protocol was a paper-based questionnaire that was meant to elicit their subjective impressions of using the system.

The experimenter was not in the room during the sessions except when the user was filling out the second questionnaire.

Unfortunately, we found that the detailed manual alone was insufficient for the user to be able to understand the system functionalities and be comfortable enough with the system to perform the evaluation tasks. This was evidenced by the low number of answered questions and infrequent use of novel modalities. In the case of voice, many users stated in the post-experiment questionnaire that they simply didn't feel confident that the vocal control would work. Moreover, a significant number of users said that they would have liked to have had a step-by-step tutorial for using the system. All of these factors were taken into account in the design of the final round of experiments, described in the following sections.

5.4.3 The challenge of an unbiased tutorial

We had been initially reluctant to provide users with a hands-on tutorial since we thought that it would unduly bias both the interaction paths that they chose and the modalities used for specific interactions. However, the obvious difficulties that the users had in using the system to answer the questions made it clear that one would be necessary. Ideally, users would have been trained in the use of the novel modalities such as voice and pen input on other interfaces so that when the time came to evaluate the Archivus system, they would be comfortable with them and would only have to learn the functionalities of the system. However, both time and manpower constraints made this impossible. As a result, the tutorial had to not only teach users about the different functionalities and data types available in the system, but also to prove to users that the novel technologies worked well and could be beneficial to the users' interaction with the interface.

We created a separate tutorial for each of the possible modality combinations under examination (ten in all). The example questions, which took users step-by-step through how to find the answers, were the same in each of the ten tutorials, as were the interaction paths used to reach the answers, except where the nature of a given modality forced a change. For example, when using mouse and pen, the path taken to find a keyword involves more steps than in the case of using voice, where the user would immediately be able to specify the keyword. Moreover, we tried to keep the language as modality-neutral as possible, using words such as 'select' instead of 'click'. Each user had only one of the ten conditions during the experiment, and consequently only read one of the ten tutorials.

5.4.4 The evaluation protocol with a tutorial

As in the case of the pilot experiments, users were given the questionnaire and consent form, as well as the explanation of the evaluation. But, in the final set of experiments, instead of the detailed manual, users were given the tutorial and allowed to interact with the system as they worked their way through it. The first page of the tutorial briefly explained in general terms the Archivus system and the parts of the interface, while the rest took users step-by-step through 3 sample questions. Users were also given a manual (a refined version of the explanation document from the pilot experiments), which explained in more detail the various parts of the system. The users were told that the manual was there as a reference and that they did not have to read it. As in the pilot experiments, the next two steps were the actual evaluation sessions, followed by the questionnaire.

We found that with the tutorials, users were much more comfortable with the system, exhibited by a higher success rate in answering questions, and were also more open to using novel modalities such as voice and pen. We did however, notice that while users were now more willing to try out and continue to use novel input modalities, and in particular voice, they still tended to follow interaction patterns established in the tutorial to various degrees. This was of course not the case for all users – some quickly established their own ways of using the system.

5.4.5 Overview of experiment documents

In this section we give an overview of all of the documents that the users were given to read and fill in during the experiments. These documents helped put the experiments in context, familiarize the user with the system, and gather demographic and subjective data about the users themselves.

Introductory documents

The first two documents that the user saw were a consent form (Appendix B), which explained that they would be recorded during their interaction and that only members of the project would have access to the recordings, and a document explaining the evaluation scenario – the user was a new employee in a company and their boss had asked them to do some fact checking (Appendix C).

Archivus manual and tutorial

In [124] Mehlenbacher explains that users can essentially be given two types of information about a system, procedural (how to use the system), and declarative (how the system works). To our mind, this is the difference between giving the users a system

manual (declarative) and a tutorial (procedural). He goes on to say that in general, users are less satisfied and less productive when given a manual rather than a tutorial. Our motivations for providing both types of documents, as well as a description of their content, has already been given in the previous section and will not be repeated here. Examples of the tutorials can be found in Appendix D.

Questionnaires

The final set of documents were questionnaires used to elicit information about the users themselves and their subjective impression of the system. Questionnaires are commonly used to gather this type of information [91], as are interviews. We decided to use questionnaires rather than interviews because we believe that the questionnaire format gives users more time to consider their answers, puts less pressure on them to give answers that please the experimenter (some people find it harder to criticize when they are faced with a real person – possibly the developer of the system) and will control for important variables introduced by the presence of an interviewer such as exactly what is said and how they respond to an interviewees answers. The former factors are extremely difficult to control for even within a single interviewer meeting with a large number of users, and become even more so when several interviewers are involved, as we expected to be the case in our experiments.

We decided to administer two questionnaires – one before and one immediately after the experiment. The pre-experiment questionnaire gathered information about the user such as their gender, age, experience with computers and software, and how often they participate in meetings. The post experiment questionnaire asks the user to give their impression of the system in general, and on particular aspects of it, in a mix of Likert-like, check-box and open-ended questions. To ensure that the type of data used in the database was not a limiting factor in the interaction, questions about the users' impressions of the data are also included. Similarly, questions about the likeability of the graphical interface are included. Examples of the questionnaires can be found in Appendix E.

5.4.6 Questions used

Since fact checking was the scenario that was chosen for the evaluations, we chose to present questions in two formats – either as true/false questions or as short-answer questions. An experiment described by Erbach [125] claims that people tend to make significantly more errors when answering factoid type questions than definition

questions[4] and that the latter can be answered twice as fast as the former. However, we believe that the tasks in the meeting domain will entail more factoid type questions and have decided to use this type of question despite the time that is consumed answering them.

The questions themselves were chosen so as to give as much variety as possible in the type of content that users would have to access and to vary the amount of steps that users would need to take to find the information required. For example, some questions required only 2-3 steps, using any modality, to find the answer, while others required between 10 to 20 depending on which modalities were being used. In pilot experiments, the questions were presented on paper, in two groups of 5 questions each, alternating true/false and short answer questions. One additional question was added to the question set for the first phase of the experiments. This was a question that required the user to find a particular slide in the meetings based on an image of that slide on the question sheet. An example of the question sheet for the pilot experiment can be found in Appendix F. During the pilot experiments we also performed a cross-validation, alternating the question order and question groups, in order to determine whether there was any influence introduced by question order, since as Dumas [115] and Ogden and Bernick [32] point out, the wording of instructions, tasks and scenarios can influence users in subtle ways if care is not taken. We found no influence. However, we remarked that because users were presented with all of the questions that they had to answer at once on the question page, they did not adhere strictly to the question order. This made analysis of the recorded data difficult as the experimenters had to guess which question a user was trying to solve, when they had solved a question, and when they had moved on to a new question.

For the final round of experiments, we kept the questions but gave the same set question order for all users in all evaluation conditions. A document containing the questions and the question order can be found in Appendix G. Instead of presenting the questions on paper, the questions were typed on laminated question cards. The volunteers were asked to do the questions in the order in which they were given in the stack of cards, write their answer directly on the card with an erasable pen, and once finished, put the card into a box which made it difficult for them to retrieve the card again once it had been submitted. The aim of the box was to prevent users from going back and changing answers if they came across additional information later in the evaluation. Once they had submitted the

[4] Although no direct examples of factoid and definition questions are cited by Erbach, we infer from the paper that a factoid type question would be something like *What happened on May 21st?* while a definition question would be something like *What role does Mr. Smith have in the company?*.

card, they were asked to reset the system by saying *task finished* or selecting the *Task Finished* button. We found that this method worked quite well. Almost all users respected the order of the questions, and it was much easier for the experimenters to match the answers with the recorded interactions in post-experiment analysis. The only cases that posed difficulty were those where users did not reset the system before starting a new question. However, this impacted their interaction pattern analysis rather than the delineation of question start and end points.

5.5 The users

Dybkjaer, Bernsen, and Dybkjaer [39] argue that test subjects should be as similar as possible to the target population for the application being tested, since the background of subjects can significantly alter how they interact with it. Ideally, our users would be employees of a company, in a blend of managerial and lower-level positions. Moreover, since applications such as Archivus are meant to be for the mass market, the backgrounds of users could vary immensely in terms of nationality, language and computer skills, and cultural background. Therefore, users for the experiments should be taken from a relatively diverse pool of people with different degrees of familiarity with computers, databases and meetings, preferably coming from different types of institutions (universities, businesses, etc.), and of different ages and cultures. These users may be solicited, or volunteers, but all should have a relatively good command of the English language as both the interface and the database content are in English.

In a research environment and with limited resources and budget, meeting these requirements is particularly hard to achieve. Consequently, we decided to only filter volunteers based on two of the requirements - good command of the English language and a self-assessed comfort with computers. The other requirements could be taken into account post-experiment since the pre-experiment questionnaire asked users specific questions about their computer backgrounds, as well as the frequency with which they attend meetings in their everyday lives. Recruiting for our experiments was done on the University of Geneva and the Ecole Polytechnique Fédéral de Lausanne campuses.

We had a fairly even distribution of male (58.75%) and female (41.25%) volunteers, who ranged in ages from 18 to 55. Of these, just over half (53.75%) were aged between 18 and 24, 37.5% were between the ages of 25 and 35, 6.25% were between 36 and 45, and the other 2.5% were between the ages of 46 and 55. Of these, 85% were non-native speakers of English with a self-assessed good level of reading and speaking skills in English. Most of the volunteers (70%) were university level students. The others came from a variety of

different professions ranging from researchers, to engineers, to translators. Nevertheless, 38.75% of them attended meetings a few times a month, with 12.5% attending a few times a week, and 5% attending meetings every day. Only 7.5% said that they never attended meetings and 21.25% said that they attended them only a few times a year.

Computer literacy was self-assessed by the volunteers. One of the questions that they were asked was how many hours they spent using a computer each day. 38.75% said that they spend 2 to 4 hours with a computer, 23.75% said it was 5 to 7 hours, and 21.25% said it was 7 to 10 hours. A surprisingly large number (13.75%) said that they spend over 10 hours each day in front of a computer. Only 2.5% said that they spent less than an hour. An overwhelming number of our users (95%) used Windows systems on a daily basis and were familiar with browsers (98.75%). Slightly fewer used word processors (90%) and audio players (82.5%) regularly. Only 63.75% of the users used video players regularly. A surprisingly high number of users (33.75%) said that they had used automatic speech recognition programs in the past and 23.75% said that they had used voice to control their computer. However, only 18.75% had experience with database tools.

We feel that in general, this user population is fairly representative of a general computer user population, and is a suitable substitute in lieu of having access to actual potential users of the Archivus system.

5.6 Types of data gathered

Several types of data were gathered during the experiments. The first is personal information and the user's subjective opinion of their experience with the system as gathered in the pre- and post-experiment questionnaires. The second is audio/video recordings of the user as they are interacting with the system. This was done from two different angles. The first angle was a view of the user's face. The other was a view taken from behind, above and slightly to the side of the user. This angle allowed us to see in more detail the actions that the user was making. This was particularly useful in the case of interaction involving pen, where attempted interactions that were too light or performed at the wrong angle were not registered by the system, and therefore not logged, but were visible on this camera view, and could be manually added to the logs after the experiments had been completed. In addition to the two camera views, we also recorded a view of what was happening on the user's screen throughout the experiments, and logged all user and system actions in a text file. All of the data was time-stamped so that it could be easily synchronized in the post-processing stage.

5.7 Conclusions

In this chapter we have described the Wizard of Oz methodology and the extensions to it which were necessary in order to apply it to experimentation with a multimodal system, as well as the physical and technical environments in which the experiments took place. We also discussed the difficulties encountered in designing the protocol for the experiments and how they were resolved, as well as describing the protocol itself in detail, both for the pilot experiments and for the final set of experiments. The description included the two phases of the experiment, the various steps during each phase, the types of questions they had to answer, and the types of documents and information that they had access to throughout the experiment. Finally, we discussed the broad range of users that participated in the experiments, as well as the types of data that were gathered. In the next chapter, we present an analysis and discussion of the data that we gathered as a result of these experiments.

6. Experiment Results

In this section we discuss the results of the Wizard of Oz experiments with the Archivus system described in the previous chapter. We begin with some general comments in section 6.1, followed by a discussion of the users' subjective opinions of their interaction with the Archivus system is section 6.2. Here, we look at both their general impressions, and specifically at how they perceived the usefulness of the different input modalities available. In section 6.3 we discuss whether learning to use the system in a particular condition induced a learning effect such that in phase two modalities from that condition were preferred over the modalities that were added. In section 6.4 we draw some general conclusions about modality use based on the proportions with which the various modalities are used in the experiment conditions. We then look at how the use of modalities evolves over time in section 6.5. In section 6.6 we look at how use of modalities changes when users encounter a problem in their interaction with the system. Specifically, we look at which modalities are more error prone, the nature of those errors, the proportions of errors between the different modalities, how often switches between modalities are made, and the nature of those switches. In section 6.7 we discuss whether users perceived pen and mouse, and voice and keyboard to be functionally equivalent modalities. In section 6.8 we look at task completion measures including the number of questions answered in each condition, how correctly users answered those questions, and whether external factors from the users' background influenced performance. Finally, in section 6.9 we look at whether there are any correlations between modality preferences and the types of tasks being solved

6.1 Introductory comments

The data that we discuss in this chapter came from a set of experiments with 80 users in 10 conditions, who interacted with the Archivus system for a total of 40 minutes. As described in the previous chapter, each user worked with the system for two 20 minute phases. In the first phase, the user had access to a subset of modalities, and in the second phase they were given access to a full set of modalities. This meant that in the second phase all users, independent of their condition in the first phase, had access to both voice and keyboard, and one of either mouse or pen (though never mouse and pen together). Of those users who had access only to language modalities in the first phase, half of the users were given a mouse and the other half a pen as pointing modalities in the second phase, in order not to influence results based on pen/mouse differences.

On average, each user made 190 individual interactions during the full 40 minutes, of which 115 were made by mouse or pen, 67 by voice, and 8 by keyboard. This resulted in a data set (all 80 users) totaling 15239 individual interactions, of which 9226 were made by mouse or pen, 5355 by voice and 658 using the keyboard. Furthermore, we found that while users often used the different modalities available, there were no truly multimodal interactions in the sense that users used more than one modality to generate a single action. When using Archivus, each individual action was generated by a single modality, even though the modalities might have changed between actions. For example, if for a desired interaction search criteria needed to be specified, they were specified only by voice, without the assistance of either the keyboard or a pointing modality, while the next action, for example the selection of a book in the bookcase, would be done only by mouse. We had no case of interaction such as the users saying '*Show me this meeting*' while indicating a book in the bookcase using a pointing device. We hypothesize that this is because users are not accustomed to being able to interact with computers in this way, and since they were not explicitly shown that this was possible during the tutorial, it did not occur to them to try.

Assessing the quality of the data

Before beginning an analysis of the data, we wanted to assess whether all of the data that we had gathered was of sufficiently good quality to be included. In any experiment it is important to eliminate data that is considered to be outside the norms for a particular data set, as it can introduce unwanted variations in the resulting analysis. We were concerned for example, that there would be users in some of the conditions who did exceptionally well or particularly poorly compared to other users in that condition. If this was the case, then those users should be removed from the data set as they are not representative of the population. In order to find such users, we looked at three aspects - user performance based on success score (see section 6.8.2 for a detailed discussion of success scores), the number of questions answered, and the number of individual interactions made.

In order to determine eventual outliers in terms of success score we calculated the mean and standard deviation for success scores across all users within a condition and then noted which users did better or worse than one standard deviation from the mean. Next we performed the same analysis, but this time using the number of questions answered by each user rather than their success score. To determine which users were outliers in terms of performance we looked for those users who were more than one standard deviation from the mean, in *both* score and the number of questions answered. Based on this analysis we found 6 possible outliers.

However, we felt that the success score and the number of questions answered themselves were not sufficient to determine an outlier for our study since user performance is only one factor in assessing the usefulness of a modality. A large portion of the work presented in this thesis depends on how often modalities are used, alone or in relation to one another. Consequently, we decided that the number of interactions that an individual user made using a specific modality, when compared to other users in the same condition, was equally important in determining whether a user qualified as an outlier or not. To do this, we calculated the mean and the standard deviation for each modality in each phase of the experiment and overall for each of the 10 possible modality conditions. We then noted the users that were not within one standard deviation of the mean for each modality and each phase. For a user to be classified as an outlier, they had to be outside one standard deviation in both phases of the experiment and overall, *and* this had to be the case for at least two of the three modalities. There were only 3 outliers found using this method. Of the outliers found using both performance and the number of interactions, only one fell into both categories. However, we chose not eliminate this user from our data set as they were a weak outlier when looking at the number of interactions, which we felt was a more important factor than performance. Since there were no true outliers that fell two or more standard deviations from the mean and closer analysis (at the level of one standard deviation as discussed above) also revealed no true outliers, no users were removed from the data set.

Finally, in all of the analyses presented in this section with the exception of those in section 6.6 (modality switching), we did not include cases of *attempted* interactions using a pointing modality. These interactions are most common with pen use where the user attempts to make a selection but fails because they either did not apply enough pressure on the screen, or because the angle at which the pen was held to the screen was wrong. The reason for not including these interactions is that we were more interested in how modalities were used in relation to one another and to the system, and not in an evaluation of the user's ability to use a particular input device. We do however discuss the influence that this type of problem has on interaction in the section on modality switching and include it in the calculations in that section since this type of error was very common in the pen condition and therefore was potentially responsible for a large portion of switches in conditions that involved pen use.

Acronyms

Before moving on to a discussion of the analysis of the experimental data, we would like to remind the reader of the acronyms used to represent the different conditions and phases of the experiment as they will be frequently used in this chapter.

Acronym	Full text	Acronym	Full text
M	Mouse only	P	Pen only
V	Voice only	VK	Voice and keyboard
MK	Mouse and keyboard	PK	Pen and keyboard
MV	Mouse and voice	PV	Pen and voice
MVK	Mouse, voice and keyboard	PVK	Pen, voice and keyboard
P1	Phase 1	P2	Phase 2

Table 4: Acronyms for modality conditions and phases

The acronyms for the modality conditions in this analysis are used to represent a set of users who had that modality condition in the first phase (P1) of the experiment. For example, MV refers to the group of users who had access to only mouse and voice in P1. In phase two (P2), all users had access to a full set of modalities (voice, keyboard and either mouse or pen). However, there were cases in the analysis in which we wanted to examine how the set of modalities in P1 influenced modality use in P2, which meant that we would need to compare user groups within P2. To solve this problem, we decided to keep the nomenclature used for the P1 condition for P2 as well. Thus, when we refer to the MV condition in P2, we are referring to the group of users who had access to only mouse and voice in P1, and the fact that the keyboard was also available in P2 is implied.

6.2 Subjective user opinion of interaction with Archivus

We will begin our analysis with a discussion of the users' subjective opinions of the system and their experiences with it, as well as their perceptions of modality use as these subjective factors might influence the interpretation of the results. This data was gathered using a post-experiment questionnaire, which can be found in Appendix E2. Only the questions directly relevant to this research will be discussed here.

6.2.1 General impressions

To get a better idea of how users perceived their interaction with the system we asked them to specify how strongly they agreed with a series of statements. Their choices for a response were that they *strongly agree*, *agree,* have *no opinion*, *disagree*, or *strongly disagree* with the statement. The questions, along with the percentage of users that gave each type of answer are shown in Table 5.

In general, we can see that users had a positive reaction to both the system and their interaction with it. However, a few points regarding user satisfaction are worth commenting on. The first is that of those users who agreed or strongly agreed with the statement *I felt in control of the system* most of them were from conditions that did not include the mouse. In fact, on average, less than half of the users from mouse-based

Questions	Strongly agree or agree	Disagree or strongly disagree
The system was easy to use	85%	3%
I was comfortable working with the system	81%	4%
I felt in control of the system	54%	9%
I could use the system how I wanted to.	41%	10%
It was hard to learn to use the system.	9%	75%

Table 5: User responses to post-experiment questionnaire

conditions agreed or strongly agreed with the statement. This result is surprising since we assumed that because the mouse is such a familiar modality, it would leave users with a greater sense of control over the system. The fact that the opposite is the case can be attributed to the fact that the ways in which Archivus can be used are quite different from the ways other software is used, and that this was in fact confusing mouse users more than helping them. For users in non-mouse based conditions everything was new, which might have made interaction more difficult, but seemed to result in a higher overall acceptance of the system.

The next point is related to the statement *It was hard to learn to use the system*, where 75% of users said that it was not hard, while 9% said that it was. We found it interesting that the assessment of individual users as to how hard they found the system to learn does not correlate with their performance while using the system. Some of the users who found the system hard to learn did well using the system, while some who found it easy to learn did poorly (a discussion of what doing well and doing poorly means in this context can be found in section 6.8).

Finally, we asked users *Did you find the voice control of the system useful?* requiring a simple yes/no reply. 81% of the users said that they did find it useful, and only 4% thought that it was not useful. Of those who found it useful, the M and MK conditions had the lowest number of users per condition who gave that response, but the figures still correspond to more than half of the users in the condition.

6.2.2 Perceived usefulness of modalities

In order to investigate how users perceive the usefulness of various modalities, we asked them to rank each modality for accessing a particular functionality in the interface. They were told to give the most useful modality a ranking of 1, the next most useful a ranking of 2, the least useful a ranking of 3, and a ranking of 0 if they thought that the modality was not at all useful. The tables below, collectively referred to as Table 6, show the number of users that gave a particular rank for a particular modality (in the case of mouse and pen, they are put in a single category since they never co-occur and are functionally

equivalent). The numbers in bold indicate the modality that received the highest number of votes for each ranking.

Browsing in a book

	V	M/P	K
1	34	**42**	31
2	**25**	24	6
3	12	11	**18**

Finding specific information in a book

	V	M/P	K
1	**44**	26	14
2	14	25	**30**
3	14	**17**	12

Accessing predefined criteria buttons

	V	M/P	K
1	29	**43**	5
2	**23**	20	12
3	14	9	**15**

Accessing the search criteria list

	V	M/P	K
1	33	**36**	7
2	15	**19**	16
3	**10**	7	4

Accessing the bookcase

	V	M/P	K
1	**45**	31	13
2	17	**33**	17
3	15	11	**19**

Table 6: Usefulness rankings of modalities for different Archivus functionalities

We can see from these tables that for most interactions that involve manipulating the interface (and more specifically manipulating elements that are visible on the screen) such as browsing in the book, accessing the search criteria list, or using the predefined search criteria buttons, a pointing device is preferable. However, voice is preferred to pointing for accessing the bookcase. This is likely because the book icons are quite small on the screen, and users find it more convenient to say the name of the meeting rather than to select it by pointing, which in the case of smaller items is more difficult due to the small surface area. When looking for specific information in a book, which involves specifying search criteria, the language modalities are preferred to pointing, with voice considered to be more useful than keyboard.

6.2.3 Conclusions

Users' subjective opinion of the Archivus system, as shown through the data gathered in the post-experiment questionnaire, shows that overall users found the system sufficiently easy to learn and use. Moreover, the perception of the usefulness of the various modalities by the users in our study corresponds to the general perceptions about pointing and language input in multimodal interfaces from other studies [40]. This leads us to believe that overall the users' perception of the system did not have a negative influence on their performance during the experiments and consequently did not negatively influence the results of the experiments.

6.3 Learning effects

The reason that the experiment was performed in two phases, in the first of which users had access to only a subset of the modalities (P1) and access to all modalities in the second (P2), was to help determine whether there is a modality-learning effect. For example, we expected that participants who only had language input (and in particular voice) available in P1 would be more inclined to again use language input in P2. We based this assumption on the fact that the language-only conditions would give those users more experience with the novel modality than those participants who had limited (via keyboard) or no language input in P1. Table 7 shows, for each condition, the proportions of pointing and language (the total column) used in each of the phases, and where applicable, the percentages for each of the language modalities.

	Phase 1				**Phase 2**			
	Pointing	**Language**			**Pointing**	**Language**		
		Total	**Voice**	**Keyboard**		**Total**	**Voice**	**Keyboard**
M	-	-	-	-	71	29	*26*	*3*
V	-	-	-	-	32	68	*65*	*3*
P	-	-	-	-	55	45	*42*	*3*
MV	43	57	*57*	-	54	46	*41*	*5*
PV	36	64	*64*	-	44	56	*51*	*5*
MK	82	18	-	*18*	52	48	*44*	*4*
PK	79	21	-	*21*	41	59	*53*	*6*
VK	-	-	*92*	*8*	38	62	*58*	*4*
MVK	61	39	*32*	*7*	70	30	*22*	*8*
PVK	28	72	*68*	*4*	40	60	*56*	*4*

Table 7: Proportions of pointing and language used in each condition

We found that our assumption about the presence of a learning effect was valid. In fact, there is a clear learning effect for the M, V and VK conditions. Looking at the figures for P2 we see that in the M condition the mouse is used 71% of the time, which is much higher than in the other conditions involving mouse use in P1. Similarly, language is used 68% and 62% of the time respectively for the V and VK conditions, which is much higher than the general average of 48% for other conditions involving voice or keyboard use. In these calculations, we do not include the MVK and PVK conditions since no additional modality is added in the second phase, and thus the type of learning effect that we are referring to here is not possible.

It is also interesting to note that there is less learning effect with pen use. When examining pen use, we can see that in the P condition in P2, pen is used slightly more than language, whereas in all other pen-based conditions, which had an additional language modality, language is used more than pen. This suggests that there is a learning

effect with pen as well, but that it is much smaller than those of the mouse and language-only modalities.

Finally, we find no learning effect with keyboard use. When looking at the conditions that had only keyboard as the original language input modalities (MK and PK) we see that keyboard use drops significantly (~14%) in P2 when voice use is introduced. We believe that this is due to the inherent novelty of using voice as an input device. Voice is used more often when it is first introduced because it is novel and users are interested in trying it out, even though it is functionally equivalent to keyboard input. Moreover, keyboard use is minimal – on average 4.5% – in all conditions in P2, ranging from 3-6% in all cases except MVK, where it is 8%. The anomaly in the MVK condition could be attributed to the fact that there is a residual trace of the traditional MK paradigm that is influencing interaction. The novelty of using voice wears off by the second phase, at which point the user reverts more strongly to the traditional MK interaction paradigm, resulting in slightly higher keyboard use. The effect is not evident in the MK condition since the novelty of using voice is stronger than the traditional MK paradigm, resulting in more voice use.

6.3.1 Conclusions

The MVK and PVK conditions allow us to determine what types of proportions of use for each modality users might settle into overall if they are allowed all modalities from the beginning of their interaction with the system. We found that in general, users in the MVK condition always used pointing much more than language, and this difference became more marked in the second phase, where users in the MVK condition reached proportions similar to those of users in the M condition. A similar shift occurs with users in the PVK condition, where pointing use increases by 12% in P2, although in this case, the difference between the pointing and language is not as marked (40% pointing, 60% language) as it is for the MVK condition (70% pointing, 30% language).

From all of this data we can conclude that modalities (M, P and V) and modality types (VK – both language modalities) when used in isolation during early experience with the system (P1) result in a learning effect that propagates to later system use where these initial modalities are preferred over newer modalities that are introduced. However, there does not appear to be any learning effect for cases where modalities of different types (i.e. pointing + language modality) are introduced to the user at the beginning of their experience with the system. For example, rates of use of the keyboard are similar in all of MV, PV, MK and PK, despite only two of those conditions having had keyboard access

in the first phase. This is likely due to the strong preference for voice use over keyboard for language-based interaction.

These findings are important for cases where designers of systems want to encourage use of particular modalities, and also show that independent of what a single modality is, if a user learned to use the system with that modality, they will prefer to use that modality later on, which in turn suggests that users could learn to use only that modality if they had to.

6.4 Proportions of modality use

In this section we discuss the relationships between modalities in terms of how much they are used when combined with additional modalities. Table 8 shows the proportions in which each modality is used for each phase of the experiment, and overall, for each of the modality conditions. The first observation to be made here concerns mouse and keyboard use in the traditional MK paradigm. We can see from the data in P1 in the MK condition that the mouse is the dominant modality since it is used significantly more than the keyboard. We assume this to reflect the general trend for mouse-keyboard interaction in computer use, and in particular for mouse-keyboard interaction with the Archivus system.

	Phase 1			**Phase 2**			**Overall**		
	M/P	**V**	**K**	**M/P**	**V**	**K**	**M/P**	**V**	**K**
M	100	-	-	71	26	3	89	10	1
P	100	-	-	55	42	3	84	15	1
V	0	100	0	32	65	3	19	**79**	2
VK	-	92	8	38	58	4	22	**72**	6
MK	82	-	18	52	44	4	68	21	11
PK	79	-	21	41	53	6	61	26	14
MV	43	57	-	54	41	5	50	47	3
PV	36	64	-	44	51	5	40	**58**	2
MVK	61	32	7	70	22	8	66	27	7
PVK	28	68	4	40	56	4	36	**61**	3

Table 8: Proportions of use for each modality per phase and overall

Comparing the proportions of the use of mouse and pen in all of the conditions that involved those modalities (in either of the phases except for M and P interaction in P1) we see that mouse is always used more than pen. This suggests that among pointing modalities a novel input modality (the pen) is less likely to be used than a functionally equivalent familiar modality (the mouse). However, we notice that the opposite is true for language related modalities. Voice, a novel language-based input modality, is used much more frequently than the familiar keyboard. This is true not only when voice and keyboard are used together without pointing modalities (in P1), but also more generally

across all modalities in phase two. Bilici et al. [6] suggest that such behaviour could be attributed to the impact of the Rule of Matched Modality, which claims that '*users are likely to give their input in the same modality as the system gave its output*'. However, given the fact that in Archivus the same system output was provided in both voice and text, no definitive conclusions can be drawn.

When comparing the distribution of voice and keyboard use within only the language modalities (Table 9), on average voice is used 90% of the time, while the keyboard is used 10% of the time in all conditions that involve both modalities. These figures are quite similar to those found by Oviatt and Olsen in [126], despite the difference in the types of tasks that were being solved.

	Voice	**Keyboard**		**Voice**	**Keyboard**
M	90	10	**MK**	92	8
V	96	4	**PK**	90	10
P	93	7	**VK**	94	6
MV	89	11	**MVK**	**73**	**27**
PV	91	9	**PVK**	93	7

Table 9: Voice and keyboard use proportions during P2

The only exception is the MVK condition, where although the keyboard is still used less than voice, the distribution within the language modality is very different from those of the other conditions - 73% voice use and 27% keyboard use. Moreover, looking at the evolution of voice and keyboard use over the conditions that gave access to both in both phases of the experiment (Table 10), we see that in the VK and PVK conditions the proportions remain relatively stable across both phases while the proportions for the MVK condition change. We think that this can be attributed to the fact that users had access to the traditional mouse-keyboard paradigm as well as voice throughout the interaction and that access to the MK paradigm helped to reinforce keyboard use.

	Phase 1		**Phase 2**	
	Voice	**Keyboard**	**Voice**	**Keyboard**
VK	92	8	94	6
MVK	82	18	**73**	**27**
PVK	94	6	93	7

Table 10: Voice and keyboard use, in proportions, in P1 and P2

Interestingly, when the P2 figures for the VK condition are broken down into those users who had access to the mouse in P2 as opposed to pen, we find the opposite effect. When users were given pen input, the proportions changed to 85% voice use and 15% keyboard use while for those who were given the mouse, the figures were 92% voice use and 8%

keyboard use. However, these calculations were done on groups of users (3 for additional pen and 5 for additional mouse) that are too small to be statistically significant.

Finally, there is a difference in the proportions with which voice is used with two different pointing modalities. Specifically, voice is always used more frequently when it is combined with pen as a pointing modality rather than with the mouse. Table 11 shows the proportions of pointing and voice use for each of the modalities in both phases[5].

	Phase 1		Phase 2	
	M/P	**Voice**	**M/P**	**Voice**
M	-	-	71	26
P	-	-	**55**	42
MV	43	57	54	41
PV	36	64	44	**51**
MK	82	18	52	44
PK	79	21	41	**53**
MVK	61	32	70	22
PVK	28	68	40	**56**

Table 11: Proportions of pointing and voice in both phases

In the PV, PK and PVK conditions in P2, there is a higher proportion of voice use than pen use. However, in the parallel mouse-based conditions (MV, MV and MVK) there is a higher proportion of mouse use than voice use. Interestingly, Dybkjaer, Bernsen and Minke [41] cite a study in which the opposite effect was found, with pen being used more than voice, and which attributed this behaviour to the fact that users were more accustomed to using graphical user interfaces with a pointing device. Further study would be needed to determine whether the nature of the data used might also be playing a role in the different results.

In the P condition in P2 we see that there is more pen use than voice use which seems to show that the trend is not completely consistent. However, the difference between the pen and voice proportions is quite small (13%) when compared to that in the parallel M condition, where the difference between mouse and voice use is quite large (45%). These results are likely due to the fact that the combination of two novel input modalities encourages the use of the novel modalities, as we saw with the pen-based conditions, while the addition of a novel input modality to the stronger half of the MK paradigm (i.e. using the mouse) does not encourage the use of the novel modality to the same degree. We saw in the P condition that users also followed this trend (as shown when it was compared to the M condition). Moreover, the fact that pen use remained stronger than

[5] Keyboard use is not included in this table, which accounts for why the figures do not add up to 100%.

voice use in this particular condition while not in the other pen-based conditions can be explained by the modality learning effect that was found in the P condition (section 6.3).

6.4.1 Conclusions

In this section we were interested in looking at the relationships between modalities in terms of how much they are used when combined with other modalities. We began by looking at the traditional interaction paradigm of mouse-keyboard, and found that the mouse was the dominant modality as it was used much more frequently than the keyboard. Between the novel input modalities (pen and voice) we found the voice to be stronger as evidence by 1) its higher proportion of use throughout both phases of the PV condition and 2) by the degree of difference between use of mouse/pen and voice/keyboard (functionally equivalent input modes). Here we see that voice was used much more than keyboard (which was also found to be true in [16]), while mouse was used only slightly more than pen. We also found that mouse is always used more than the pen, independent of the modality condition it is in, and that voice is always used more when combined with the pen than with the mouse.

6.5 Evolution of modality use

In the previous section we established some general trends in interaction between the modalities. Next we were interested in looking at how modality use evolved over time. Two phases of 20 minutes each gave a total span of 40 minutes during which to observe interaction for each condition. We chose to break up those 40 minutes into 5 minute intervals in order to get a more fine-grained view of how interaction evolved. It is also important to remember that P2, where additional modalities are made available in eight out of the ten conditions, begins at the 20 minute mark.

We looked specifically at two aspects – the evolution of a single modality across the various conditions, and the evolution of pairs of modalities within a condition, comparing the results from parallel conditions (for example MV and PV).

6.5.1 Evolution of a single modality

We first wanted to see whether the use of a single modality changed over time, and whether it did so in a similar manner across all of the various conditions. If this was the case, it would suggest that the use of that modality was not dependent on other modalities.

Each graph in this section (Figures 5-8) shows the change in interaction over time for a single modality. Each curve in the graph represents one experiment condition (labeled according to the P1 condition in the legend) over the full 40 minutes of interaction with the Archivus system. Only conditions in which the modality in question was used in both phases of the experiment are considered in the graph. The x-axis shows the 5 minute intervals into which the 40 minutes were divided, with a vertical line marking the point at which P2 begins. The y-axis indicates how much (as a percentage of all interactions) the modality being examined was used.

Figure 5, below, shows the change over time of voice use over the 40 minutes of the experiment, broken down into 5 minute intervals. In the graph, we see the general trend

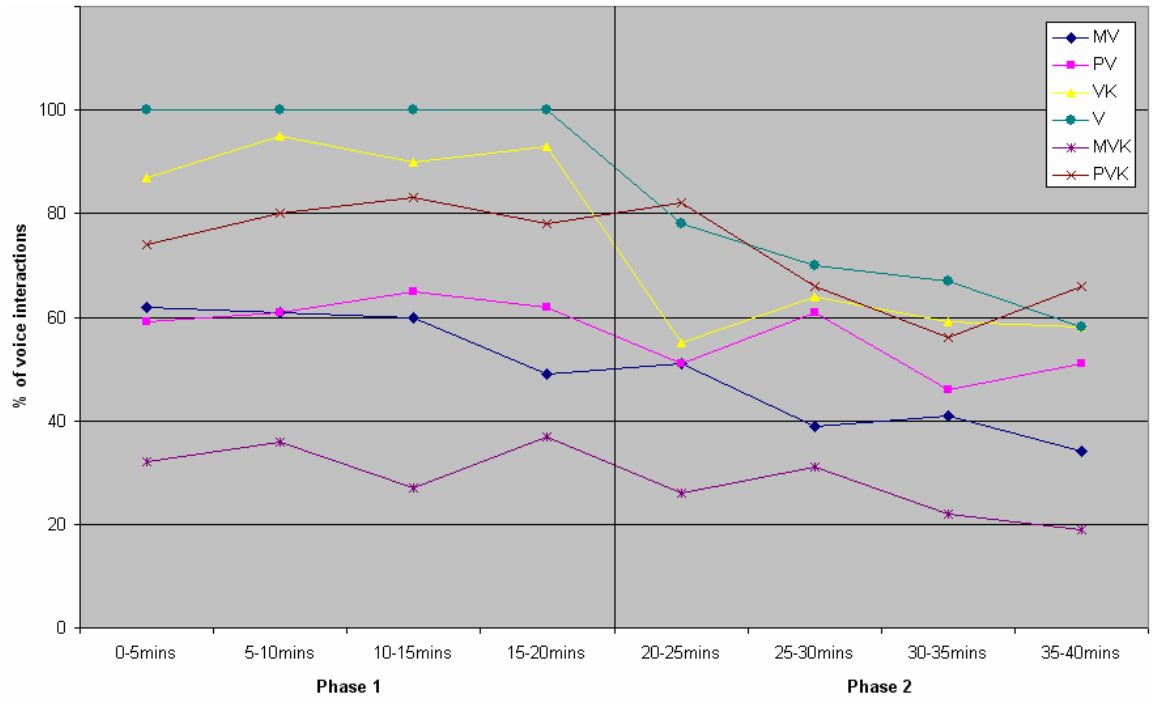


Figure 5: Change in voice use over time

for voice use to decrease over time, which contradicts results found by Rudnicky [27] which showed that voice was a preferred input modality even during extended use with a system. This difference is results might be attributed to the nature of the tasks being undertaken in the two studies.

We can also see that the V and VK conditions experience a rather sharp decrease in voice use while in the other conditions the decreases are much more gradual. The steep

decreases in the V and VK conditions are not surprising however, since they occur at the shift between phases where the user was also given access to a pointing device. What is interesting to note is the difference in behaviour between the MV and PV conditions after the first five minutes of interaction in P2. In the MV condition we see that there is a slight rise in voice use in the first five minutes which later decreases quite sharply, while in the PV condition there is a decrease in voice use, which then rises quite sharply, only to decrease again. In fact, we had expected the opposite to be the case due to the preference for keyboard interaction when the traditional MK paradigm is present (as shown in section 6.4) and the overall preference for voice interaction, but in particular in combination with the pen.

Keyboard use also clearly decreases over the 40 minutes in the conditions where it was present in both phases - MK, PK, VK, MVK and PVK. Figure 6, shows the change over time of keyboard use across the 40 minutes, broken down into 5 minute intervals.

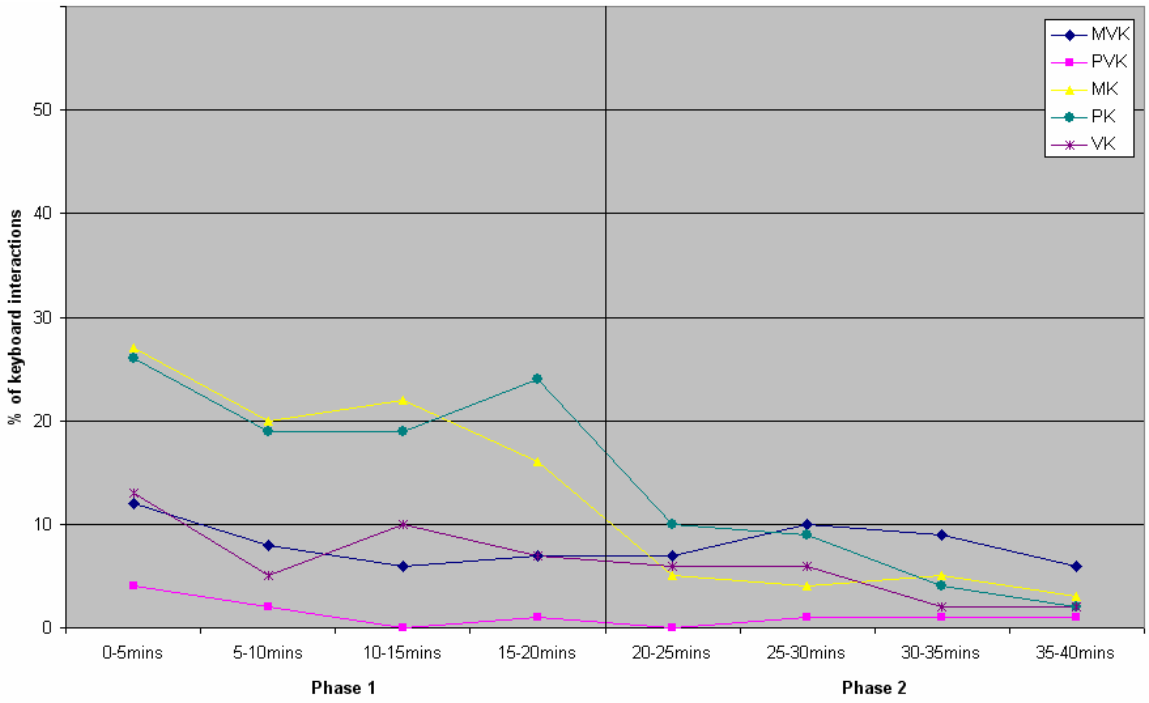


Figure 6: Change in keyboard use over time

However keyboard use tends to fluctuate quite a lot in the MVK condition, so it is not clear whether it would continue to drop off if the experiment had been carried on for a longer period of time. The other interesting point to note here is that keyboard use drops off much more steeply in P2 in those conditions where voice is introduced, which is not

surprising since voice has already been shown to be the stronger/preferred of the two language modalities.

Contrary to the general trend of language use decreasing over time, Figures 7 and 8 indicate that pointing use increases over time. Looking at Figure 7, which shows the change over time of mouse use across the 40 minutes, broken down into 5 minute intervals, we can see that the nature of the rise is independent of the conditions. In the MV and MVK conditions the rise is gradual and quite steady, but in the two conditions where voice is introduced in the second phase (M and MK), we see a similar amount of drop in mouse use, and then a fairly similar pattern in the subsequent rise in mouse use during P2, which differs from the rise in mouse use found in the conditions that had voice use throughout the 40 minutes (MV and MVK). This is likely due to the limited effect that the introduction of the keyboard as an additional modality has on interaction.

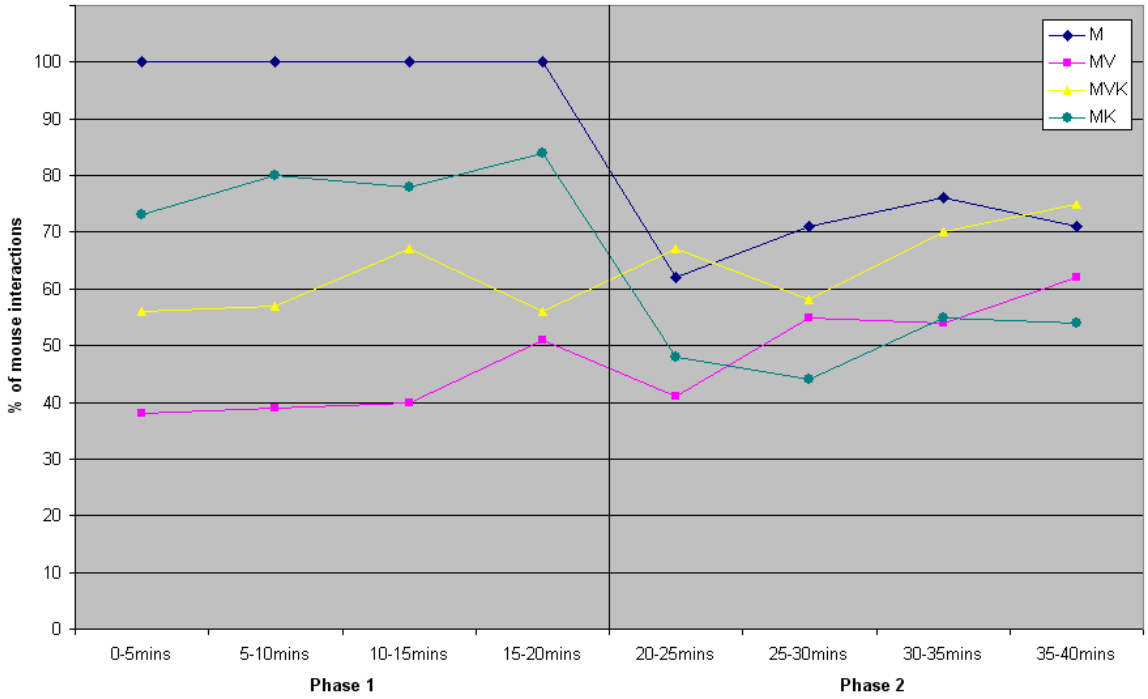


Figure 7: Change in mouse use over time

Figure 8 shows the change over time of pen use across the 40 minutes, broken down into 5 minute intervals. Here we see a very similar pattern in change in pen use for the conditions in which voice was added during the second phase (P and PK) as we saw for change in voice use for M and MK. However, we can also see that overall there is much less pen use in the conditions that have voice access throughout (PV and PVK) than there was mouse use in the parallel mouse-based conditions. Furthermore, in those conditions

the rise in pen use only becomes marked in the second phase of the experiment, while its use was quite steady in the first phase. This result is surprising because there are no additional modalities added in the PVK condition, and in the PV condition, only the keyboard is added, which has been shown to have little impact on proportions of modality use in general, and when combined with pen in particular (section 6.4).

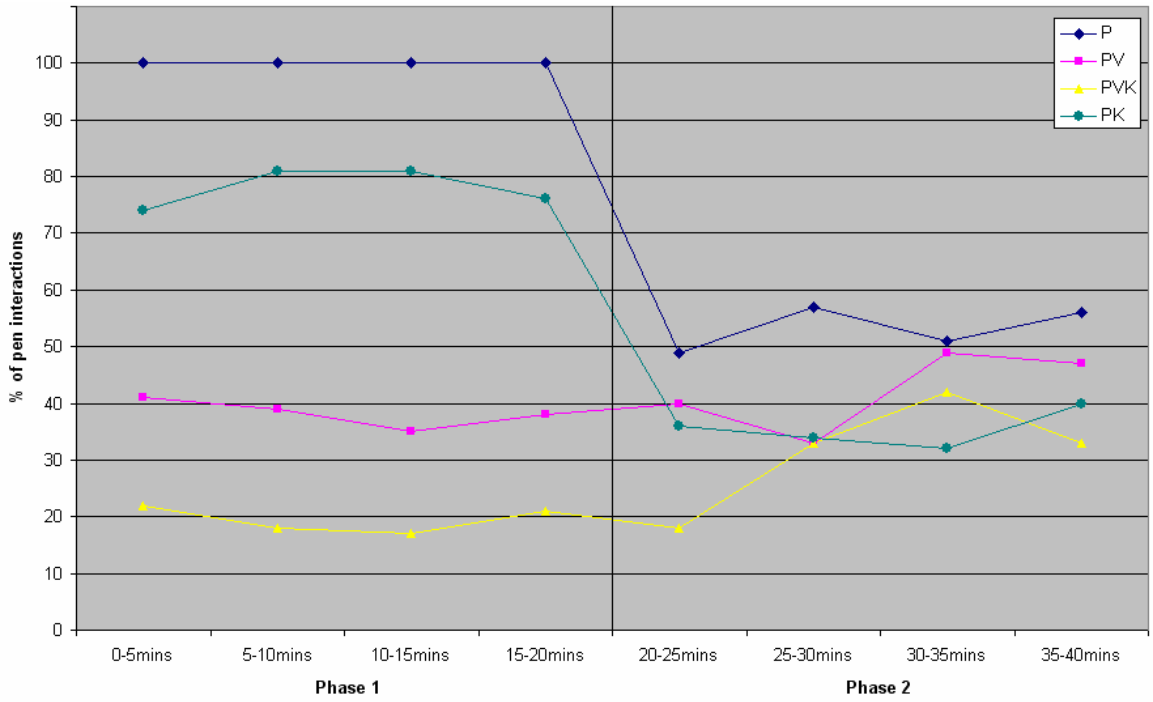


Figure 8: Change in pen use over time

6.5.2 Evolution within a condition

Finally, we wanted to compare the changes in relationships between modalities within a single condition and compare those changes to changes that occur in parallel conditions. By parallel conditions we mean those conditions where exactly the same modality is introduced in P2. For example, MV and PV are parallel conditions, since the keyboard (and only the keyboard) is introduced in both in P2. The V and VK conditions are a special case since 1) functionally equivalent rather than identical modalities are introduced in P2, and 2) in the V condition, the keyboard is also being introduced in addition to the pointing modality.

Each graph in this section (Figures 9-18) shows the change in the relationships between modalities over time within a single experiment condition. Each curve in the graph represents one modality over the full 40 minutes of interaction with the Archivus system.

The x-axis shows the 5 minute intervals into which the 40 minutes were divided, with a vertical line marking the point at which P2 begins. The y-axis indicates how much (as a percentage of all interactions) a modality was being used.

Voice and voice-keyboard

The first pair of conditions that we compared were V and VK (Figures 9 and 10), where the addition of a pointing modality plays the most significant role since the addition of keyboard has been shown to have little impact on the interaction. In both cases there was more voice use than pointing in the second phase, and by the end of the experiment voice and pointing use had converged to almost the same levels in both conditions. But, we can see a much faster convergence of the voice and pointing modalities in the VK condition (Figure 10), although the convergence in the V condition (Figure 9) is much smoother.

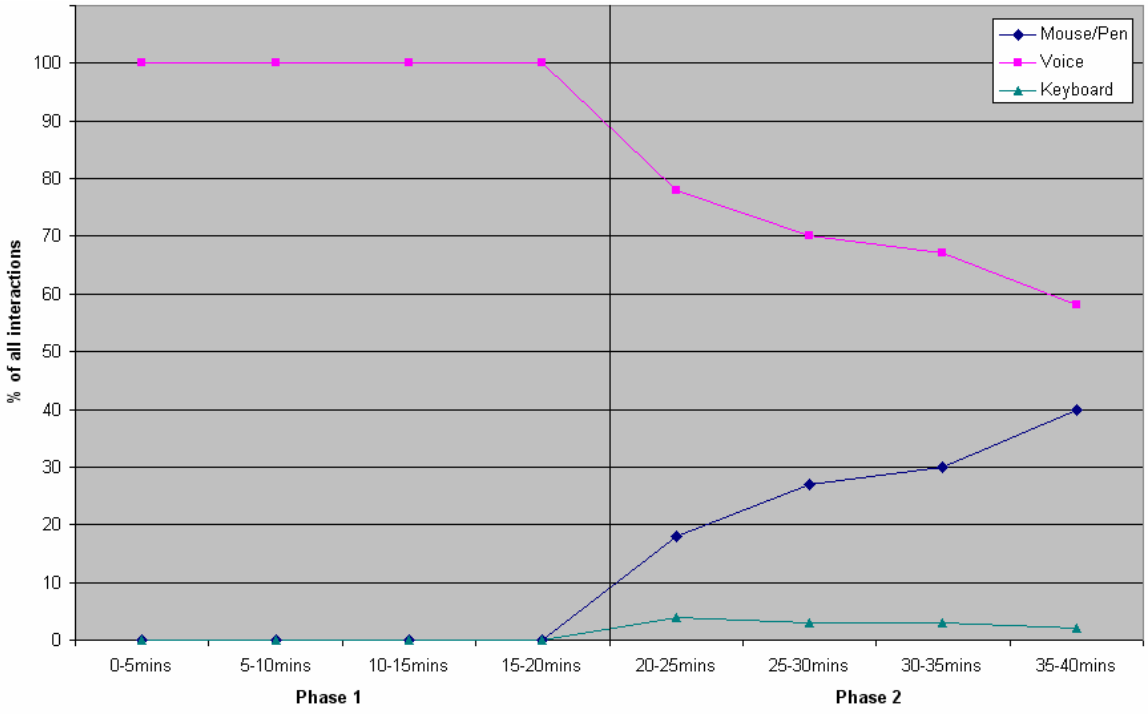


Figure 9: Modality change over time for the V condition

Despite the fact that the introduction of the keyboard does not significantly impact interaction, we believe that it is the factor that is responsible for the slower convergence rate in the V condition. Since to users in that condition the keyboard is never the less a new modality (in terms of access rather than novelty), they try to use it, perhaps in order to decide which of voice and keyboard is more convenient for specifying search criteria. Users in the VK condition will not have had this problem since by P2 they will have

already drawn their conclusions about the relative usefulness of both of the modalities, and it is only the usefulness of the pointing modality that needs to be considered in P2, resulting in a faster convergence.

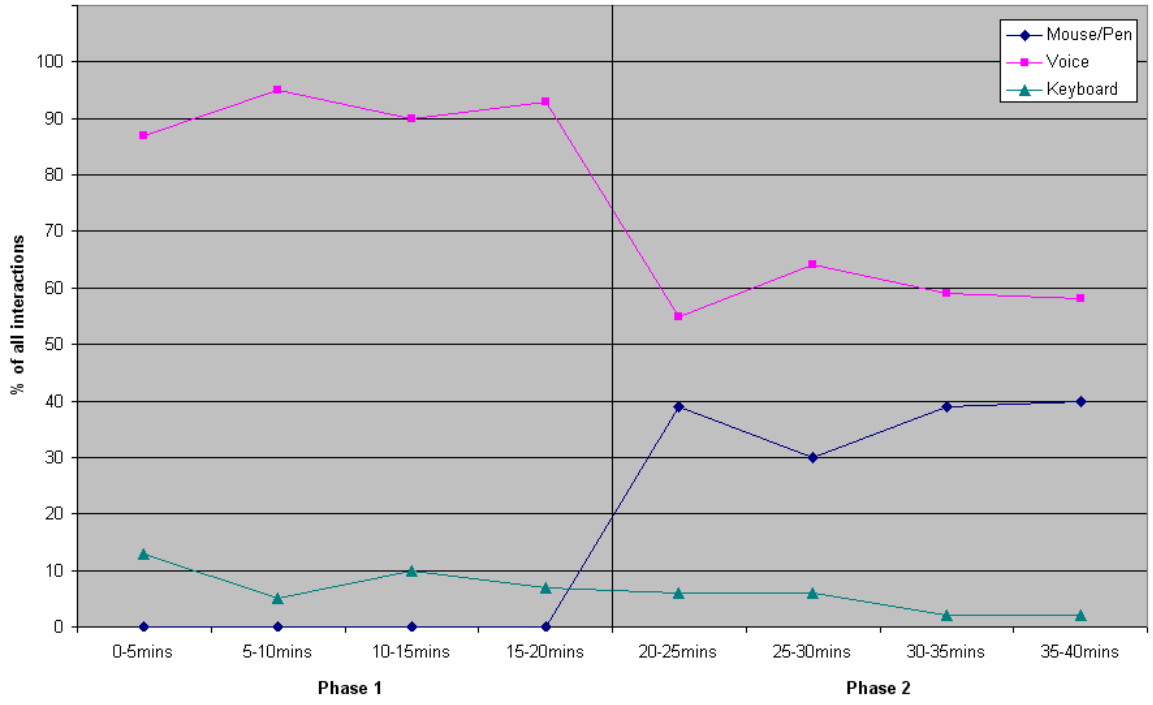


Figure 10: Modality change over time for the VK condition

Mouse and pen

The next pair of conditions were M and P, where both voice and keyboard were added (Figures 11 and 12). First, we can see that there is little impact made by the introduction of the keyboard. Next, we note that while in both cases there is more use of pointing than voice, the degrees of convergence are dramatically different in the two conditions. Pen and voice converge quite tightly almost immediately and then fluctuate throughout P2 but never diverge more than 10% from the point of initial convergence, but the same is not true for the mouse and voice combination.

In the M condition, the mouse and voice converge quite rapidly, but not to the same degree as in the P condition. While they also almost immediately diverge, they continue to do so until almost the very end of the experiment. It is only at that point that they begin to converge again. However, it is not clear whether the convergence would continue if the experiment lasted longer. The lines in the M condition however are fairly smooth, unlike in the P condition, which suggests that overall, users were a little bit more decided as to

which modalities they wanted to use to interact with the system at a particular point in time.

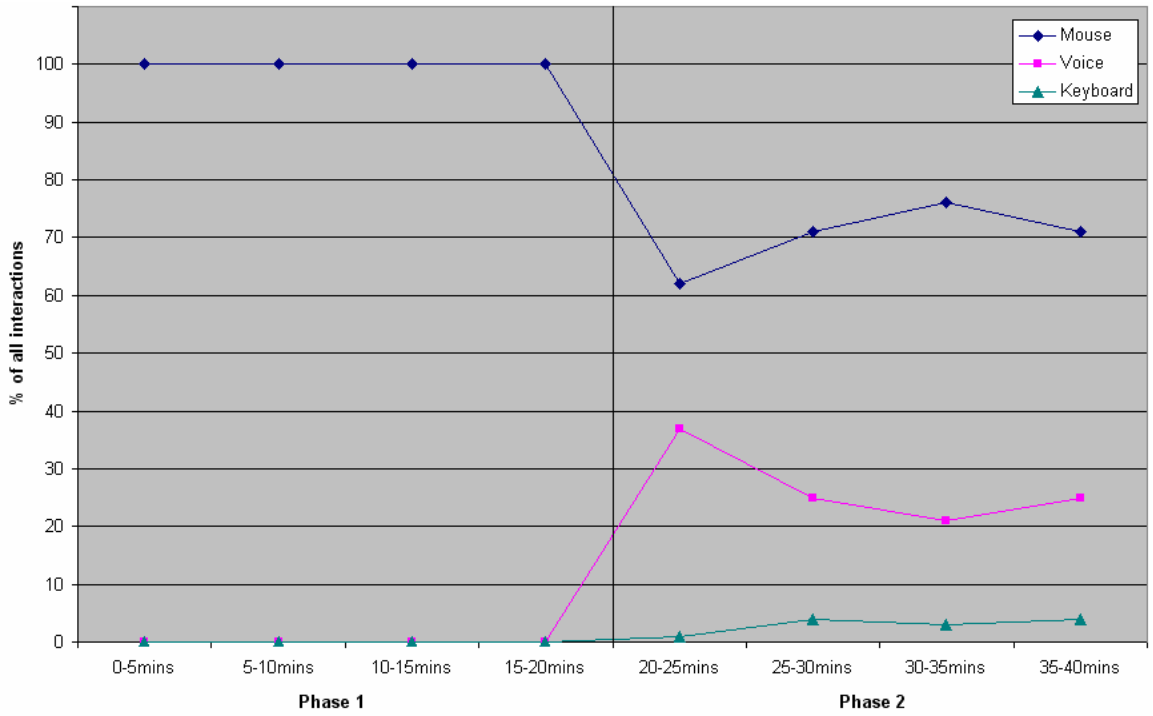


Figure 11: Modality change over time for the M condition

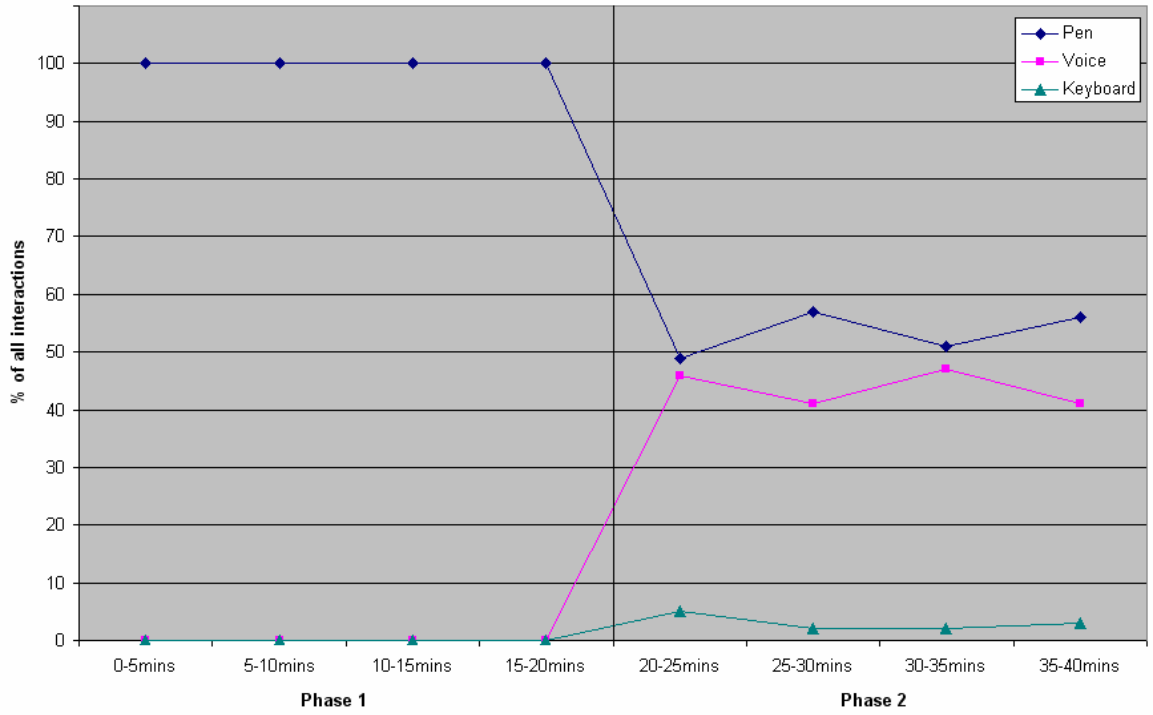


Figure 12: Modality change over time for the P condition

Mouse-voice and pen-voice

We next compared the MV and PV conditions (Figures 13 and 14), where only the keyboard was added in the second phase. In general, the distances between the pointing and voice curves in the two conditions are not very disparate. Moreover, at the beginning of the first phase we see similar patterns of use between pen and mouse in relation to voice, but as time progresses we see that pointing use rises much faster in the MV condition. We can also remark that in general, pen use is more stable than mouse use over time as evidenced by the relative smoothness of the curves in the MV condition, and the fact that the introduction of the additional modality in P2 has a much less disruptive effect (a lot of sudden fluctuations in the curves). Moreover, as can be seen in Figure 13, towards the end of the second phase the use of mouse and voice are quite far apart and appear to be diverging. In the pen condition (Figure 14) they are quite close together and it is not clear whether this pattern would be maintained or whether it would diverge or converge if the experiment had been longer.

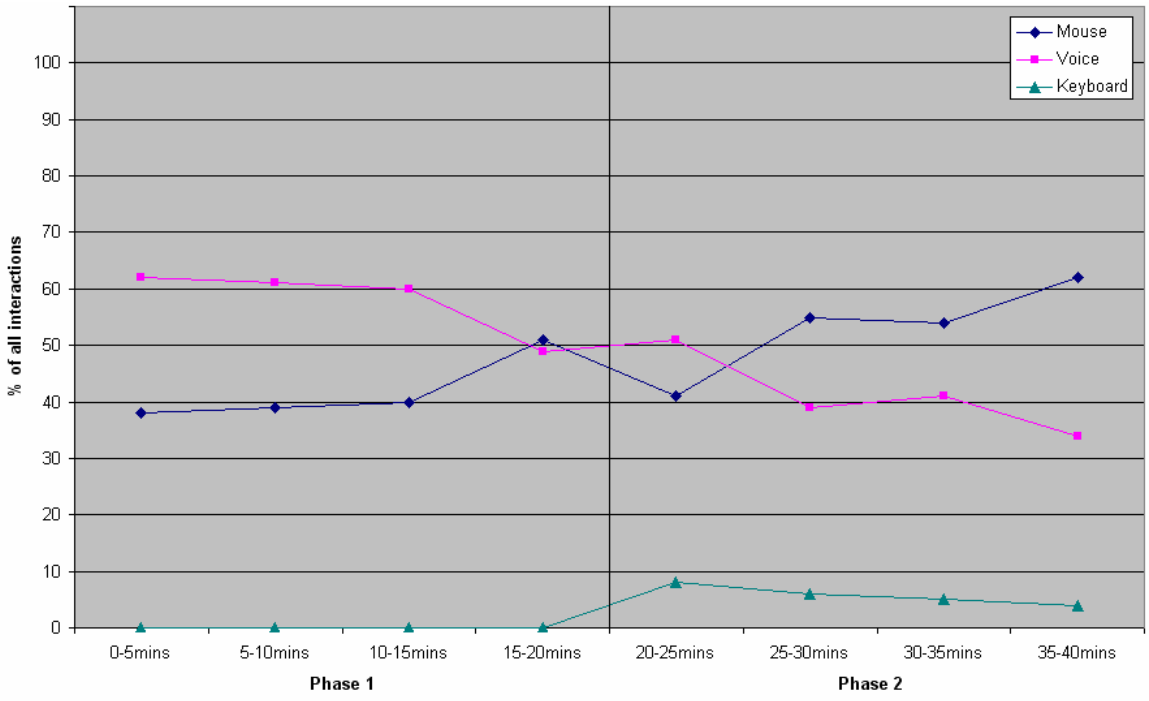


Figure 13: Modality change over time for the MV condition

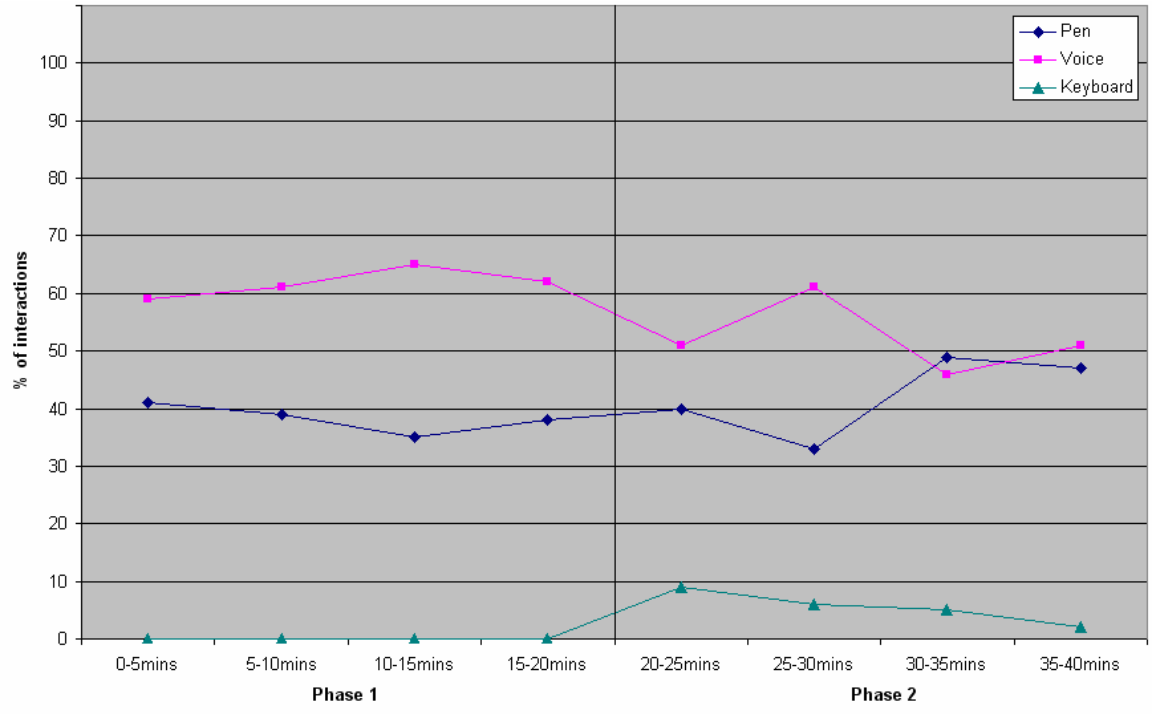


Figure 14: Modality change over time for the PV condition

Mouse-keyboard and pen-keyboard

Next we compared the MK and PK conditions (Figures 15 and 16) where only voice is introduced in the second phase. Here, we see that in the PK condition (Figure 16), there is immediately more voice than pen use once voice is introduced, and that this trend is maintained throughout P2. In the MK condition (Figure 15) on the other hand, there is equal voice and mouse use immediately after voice is introduced, followed by a slight rise in voice use, but as the interaction progresses, mouse use becomes dominant. Moreover, the curves are smoother in the PK condition, which suggests that there is less hesitation in how the two modalities are used.

Finally, we notice a trend that was also present in the MV and PV conditions. That the introduction of a secondary language modality causes more disruption in modality use when the pointing modality is the mouse than it does when the pointing modality is the pen. Here, disruption is defined as sudden and significant fluctuations in modality use that are likely the result of the user experimenting with new interaction patterns. We feel that the reason for this is again the fact that with mouse-based conditions, the traditional interaction paradigm of MK plays a factor. In the MK case, the novel modality of voice is being added, and evidence in section 6.4 has shown that users are interested in trying out the novel modality. The addition of this modality however interrupts the interaction patterns that users have established using the MK paradigm. In the case of MV, users

establish interaction patterns with the stronger half of the traditional paradigm, in combination with a novel modality. When the keyboard is added, completing the traditional interaction paradigm, users appear to be tempted to use that paradigm, since it is one that they are familiar with, which in turn interrupts their previously established interaction patterns. In the pen condition however, either a novel modality is being added to another novel modality, or a less dominant familiar modality to two novel modalities.

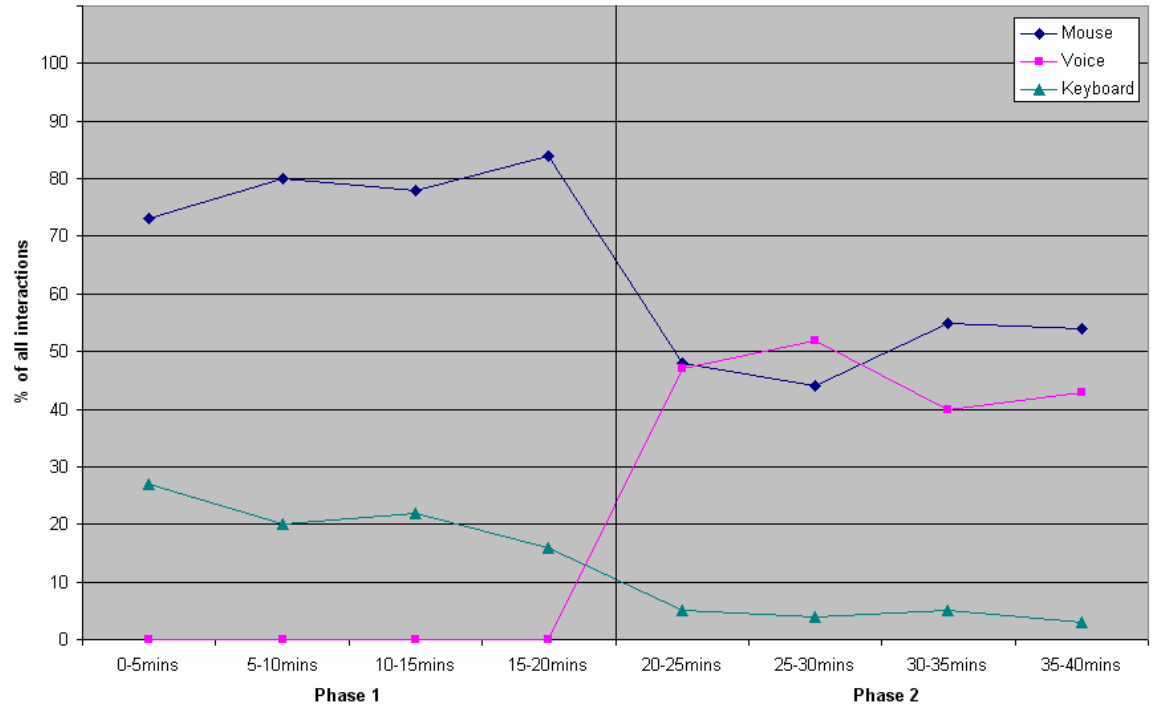


Figure 15: Modality change over time for the MK condition

One might assume that this trend for more disruption in mouse-based conditions would also hold for the M and P conditions. However, this is not the case. The likely explanation is the influence of the learning effect, which is shown to be strong for the M and P conditions, and in particular for the P condition. Moreover, adding two language modalities at once could also be giving an advantage to the pointing modality. If users are not sure which of the language modalities to use, they might simply default to the available pointing modality instead, whose use had already been entrenched due to the learning effect.

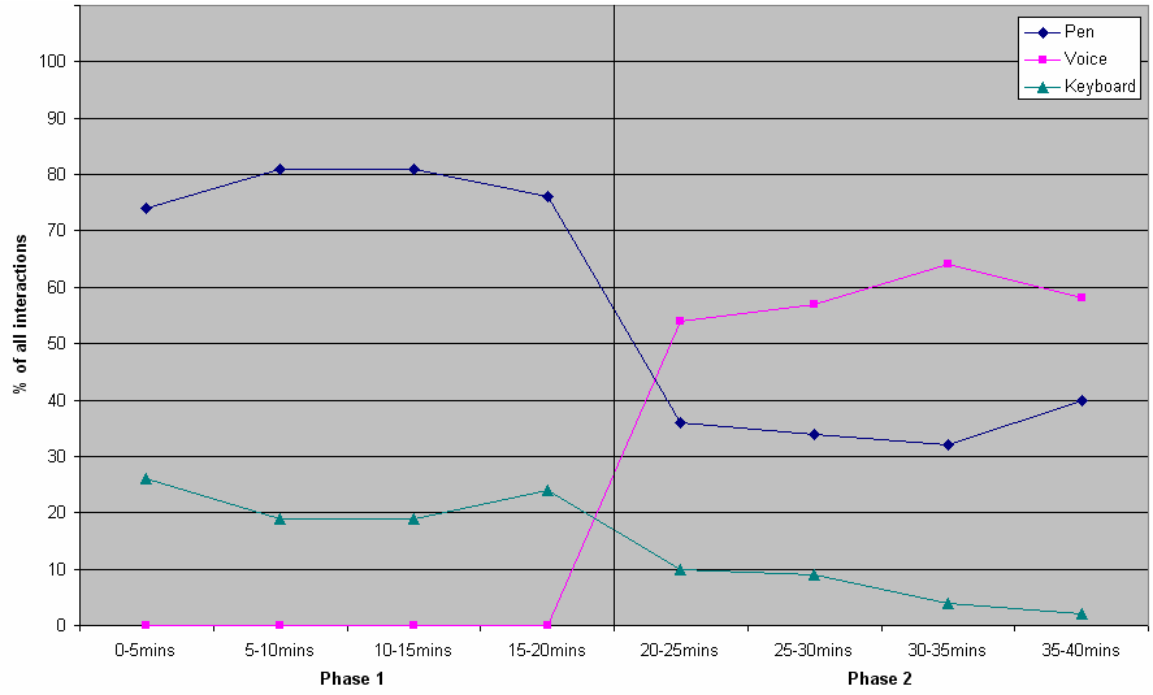


Figure 16: Modality change over time for the PK condition

Mouse-voice-keyboard and pen-voice-keyboard

Finally, we compare the MVK and PVK conditions (Figures 17 and 18), where no additional modalities are added in the second phase. We expected the relationships between the different modalities to either be quite steady or to change gradually but consistently over time. In fact, we found neither of these to be true. First, we noticed that the MVK condition (Figure 17) showed a lot of fluctuation in the proportions of interactions over time, while the PVK condition (Figure 18) proved to be much steadier. We believe that again the continual presence of the traditional MK paradigm in the MVK condition is responsible. Users who have access to the MK paradigm are torn between using it, and wanting to try the novel input modality of voice, resulting in much more erratic interaction patterns until later stages of the experiment where they have probably either established their preferred interaction patterns, or the novelty of the voice has worn off. In the PVK condition on the other hand, the fact that there are two novel input devices present leads users to establish new interaction patterns without any influence from previous experience, which results in a much steadier interaction flow. However, the steadiness is not consistent throughout the interaction as evidenced by the behaviour after the first 20 minutes of interaction.

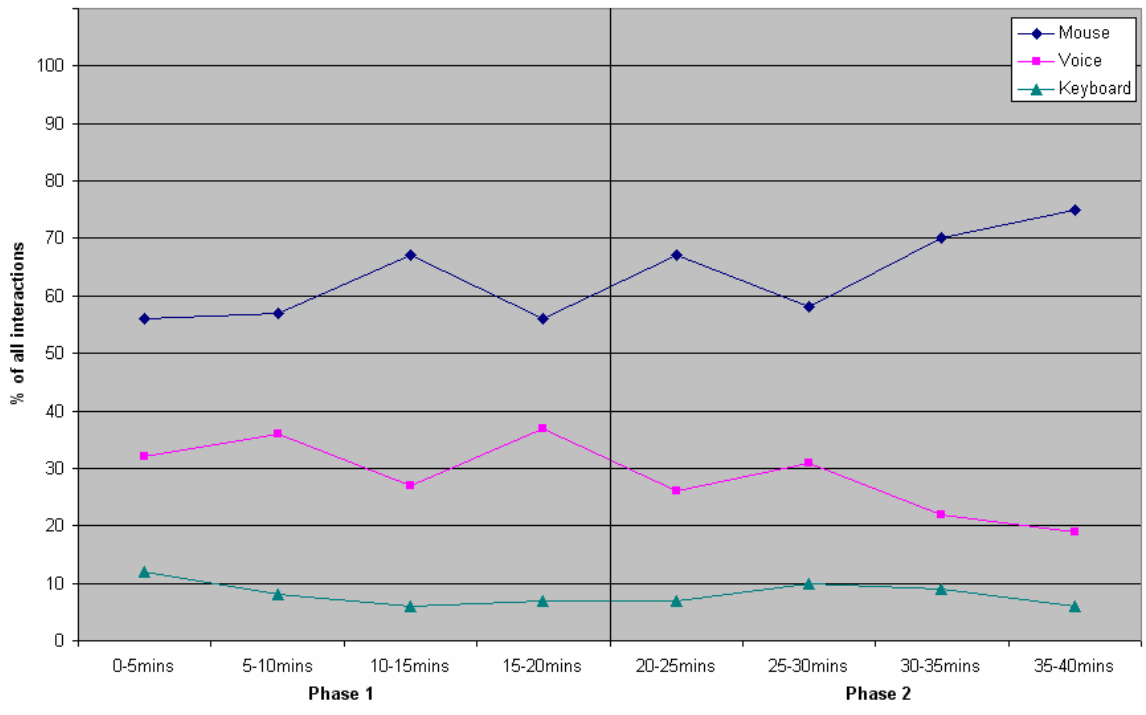


Figure 17: Modality change over time for the MVK condition

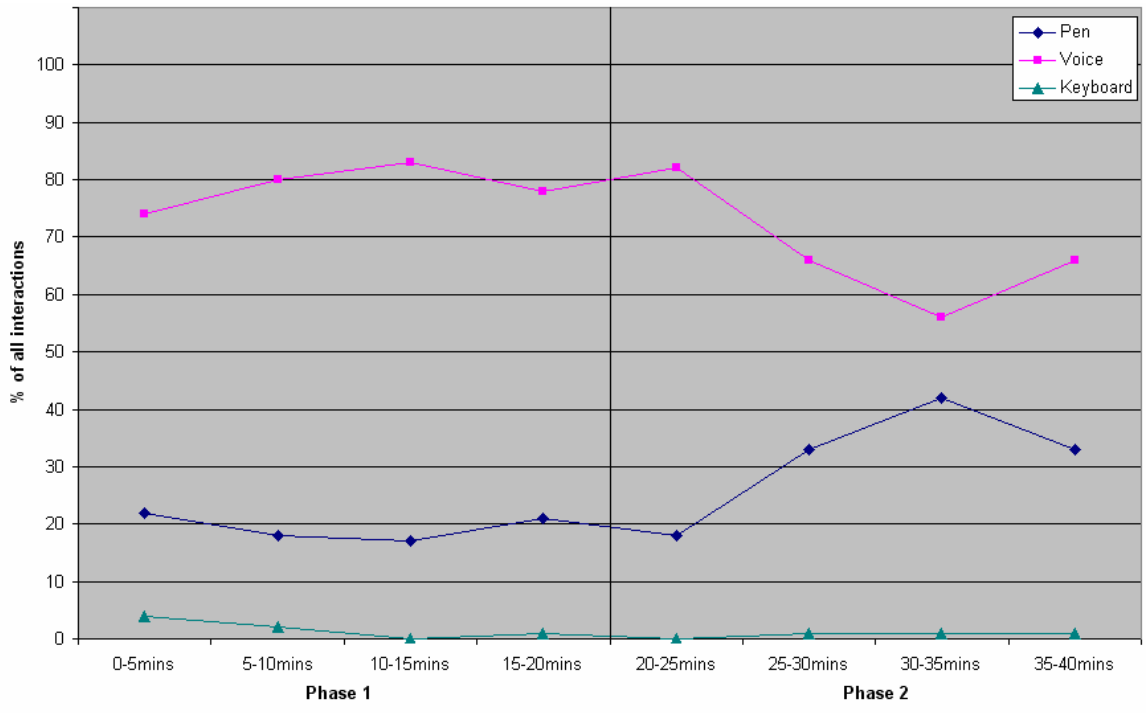


Figure 18: Modality change over time for the PVK condition

Secondly, neither condition showed a slow and gradual change. Rather, in both conditions the change occurred towards the second half of the experiment, with the MVK condition showing a slightly later onset. We hypothesize that the change occurring during later stages of interaction is due to the fact that that is how long it takes users to determine their interaction preferences. Another explanation, though less plausible, is that the short break between the two phases in which the experimenter went into the room to give the user a new set of questions may have played a role. The MVK condition (Figure 17) exhibits a degree of divergence of the modalities that suggests that it would continue until reaching a natural equilibrium at some point in time beyond the length of these experiments. The PVK condition on the other hand shows an initial convergence of the modalities, followed by a slight divergence, which makes it much more difficult to hypothesize in which direction the trend would continue. Overall, it appears that MVK behaviour settles over time, while PVK behaviour, which is quite consistent in early use, becomes less stable in later use.

6.5.3 Conclusions

When investigating the evolution of modalities over time, we were interested in two aspects in particular. The first was whether the use of a single modality evolved over time in the same way across all modality conditions, which would demonstrate that modality use was largely uninfluenced by its co-occurrence with other modalities. The results from section 6.5.1 show that the use of language modalities (voice and keyboard) decreases over time, while the use of pointing modalities (pen and mouse) increases over time, with mouse rising faster than pen. However, these results only hold at the general level. A more detailed investigation of the data showed that the addition of voice in P2 has a significant impact on how other modalities are used, much more so than the addition of any of keyboard, mouse or pen does. But, we also noticed that the addition of *any* secondary language modality had more of an impact on conditions that involve mouse than on those that involve pen.

The second aspect that we wanted to investigate was the impact that different modalities had on each other within a condition, and whether the pattern holds for parallel conditions. Results showed that in most cases, the patterns did not hold due to the difference in impact that voice use had on the mouse and the pen. First of all, the tendency to use voice more with the pen modalities was further corroborated. Moreover, the addition of voice to the mouse modality resulted in a much greater disturbance in choices of interactions than it did when added to pen. Furthermore, we saw that the use of pen was generally steadier in the PV, PK and PVK conditions than it was in the parallel mouse conditions. Mouse use however was steadier in the M condition than pen use was

in the P condition. Surprisingly, we found that the evolution of relationships between modalities was neither gradual nor steady in the conditions that allowed access to all modalities during interaction (MVK and PVK). There was significantly more fluctuation at the beginning of interaction in the MVK condition, which settled during later stages of interaction, while the opposite effect was found in the PVK condition.

Finally, we found that in all conditions, the addition of the keyboard had little impact on the interaction relationships between modalities.

6.6 Modality switching

When different modalities are available simultaneously, user behaviour when switching between modalities becomes important. In particular we are interested in how often users switched between modalities, and whether their behaviour altered when they encountered a problem in their interaction.

All modality switches were annotated by hand by a single annotator. They were noted as either a *smooth switch*, where there was a switch that did not appear to be motivated by a problem in the interaction, or a *problem switch.* Problem switches were cases where the annotator could see that the user had encountered a problem, or what the user perceived as a problem. There were several different situations that we considered as problems. In voice interaction these were cases where the speech recognizer (the wizard) failed or misrecognized a command, or the user used out-of-scope vocabulary. Specifically, they were cases where the system response was '*Sorry, I didn't hear you*' (misrecognition) or '*Sorry, I can't find that*' (out of scope) or a variation of these. For keyboard interaction, only out-of-scope vocabulary posed problems for interaction, to which the system replied '*Sorry, I can't find that*'. For mouse interaction, problems involved cases of users trying to click on an element on the screen that was not clickable. Most often this occurred when users tried to click on the hits tabs in the book (section 4.5.2). Pen interaction faced the same problem as mouse interaction, but also that of the users not using enough pressure or the correct angle with the pen in order for a selection to be acknowledged by the system. It is important to note that problems could also be encountered by users who had access to only one modality. For example, a user can encounter a problem when having access to only the mouse, in which case they cannot switch to a different modality, but must use the same modality to try to resolve the problem. In such cases, a problem was acknowledged to have happened (it was annotated as such) but there was no switch.

6.6.1 Error production

We first wanted to see whether some modalities produced a larger number of errors than others. We expected voice input to produce the most errors, since users were not familiar with this modality and had had little training as to the lexical capabilities of the system. Table 12 shows the total numbers of problems (for both phases) for each of the modalities in each of the conditions, along with the total number of problems in each modality (across all conditions) and, in the last row, the normalized proportion of errors per modality.

	Mouse/Pen	Voice	Keyboard
M	91	6	0
P	306	12	0
V	63	62	3
MK	88	9	12
PK	306	13	11
VK	78	49	8
MV	61	30	7
PV	163	26	2
MVK	65	18	10
PVK	329	27	3
Total	1550	252	56
Percent	17	5	9

Table 12: Total number of problems induced by each modality in P1 and P2

To our surprise, a large proportion of problems (17%) were produced by the pointing modalities. Moreover, we can see that in fact among the pointing modalities, there were substantially more problems encountered when using pen than when using the mouse. Anecdotal evidence suggests that this is due to the problem of applying the wrong pressure or pen angle. Between the language modalities, we can see that a larger number of errors were made when using keyboard than when using voice. This result is somewhat surprising. We had assumed that more errors would be made using voice since this type of interaction was less familiar to users. However, the high number of errors made using the keyboard can be accounted for by typing mistakes made by users, and the relatively small overall number of interactions made using the keyboard.

6.6.2 Nature of the errors

Next we wanted to examine the nature of the errors that appeared. We developed two error categories that were applied to each of the modalities. The first type of error, which we call a *miss*, occurred when a user tried to use a modality in a way that the system did not understand. For the pointing modalities, this meant failed selection on parts of the interface that were in fact selectable. For voice and keyboard input, this meant input that

the system could not understand/process, or did not hear (in the case of voice). The second type of error were *scope* errors, which are similar (at least in the case of pointing interaction) to what Donald Norman describes as description errors [100]. For pointing interactions, this meant trying to select items in the interface that were not selectable. For the language modalities, this meant all interactions that involved out-of-scope vocabulary. Table 13 shows the proportions of smooth interactions, misses and scope problems within each modality for each of the conditions for the total of the two phases.

	Mouse/Pen			Voice			Keyboard		
	Smooth	**Miss**	**Scope**	**Smooth**	**Miss**	**Scope**	**Smooth**	**Miss**	**Scope**
M	95	1	**4**	97	3	0	91	**9**	0
V	79	12	9	94	4	2	83	0	**17**
P	82	**17**	1	95	4	1	100	0	0
MV	93	2	**5**	96	3	1	84	**11**	5
PV	76	**18**	6	97	2	1	93	0	**7**
MK	91	2	**7**	96	3	1	92	**5**	3
PK	67	**27**	6	96	3	1	93	**5**	2
VK	75	12	13	94	4	2	86	6	**8**
PVK	45	**34**	21	95	3	2	86	2	**12**
MVK	93	1	**6**	96	3	1	91	**6**	3

Table 13: Proportions of smooth interactions, misses and scope problems

We can see from the table that in the pointing modalities, conditions with pen interaction have a higher number of misses than scope problems, while the opposite is true for mouse interactions. We did not consider the V and VK conditions in the analysis of pointing errors since both of these conditions had some users who had pen input and others who had mouse input in P2. As mentioned before, the high number of misses with pen interaction is likely due to problems with pressure and pen angle.

For voice interaction meanwhile, the difference in proportions between misses and scope problems is less marked, but in all cases there are more misses (situations where the system did not hear or understand the user) than there are out-of-scope errors. For keyboard errors, the results are much more erratic, and there appears to be no discernable pattern.

Looking at the proportions of each type of error across all modalities and as a proportion of all interactions we note that there are more misses than scope errors (8% vs. 5%).

6.6.3 Proportion of errors

Thus far, we have discussed which modalities are the most error prone, and the nature of those errors. Next we wanted to know what proportion of all interactions were smooth, and what proportion were due to problems. Table 14 shows the proportions of smooth

and problem interactions between modalities for each phase, and overall, for each of the conditions.

	Phase 1		Phase 2		Overall	
	Smooth	**Problem**	**Smooth**	**Problem**	**Smooth**	**Problem**
M	95	5	95	5	96	4
P	83	17	82	18	83	17
V	94	6	84	16	88	23
MK	89	11	93	7	92	9
PK	71	29	80	20	75	25
VK	93	7	84	16	88	12
MV	95	5	92	8	94	6
PV	85	15	84	16	84	16
MVK	90	10	95	5	93	7
PVK	77	23	71	29	73	27

Table 14: Proportion of smooth and problem switches in P1, P2 and overall

The first thing we notice is that the M, V, VK, MV, MK and MVK conditions had relatively low proportions of problem interactions (less than 11%), while the pen based conditions, P, PK, PV and PVK had between 15% and 29% of problem interactions. As has already been mentioned, pen use proved to be particularly problematic due to the fact that users had trouble adapting to the angle with which the pen needed to make contact with the screen and the amount of pressure needed for a selection to register with the system, in addition to the problem of knowing which parts of the screen were selectable and which were not. While from those cases the PK and PVK conditions seem to have proportions of problems that are quite a bit higher than the other two, there is no evidence to support the notion that it is the presence of the keyboard that is the factor that is posing additional problems. There are no problems with keyboard use in the PVK condition, and in the PK condition only 3 % of all problems are caused by the keyboard.

We can also notice that in the M, P, MV and PV conditions, the proportions of smooth and problematic interactions stay virtually the same in both phases of the experiment. In the V, VK and PVK conditions the number of problem interactions increases in the second phase, while in the MK, PK, and MVK condition, the number of problem interactions decreases in the second phase. The increase in the number of errors for the language-only modality combinations (V and VK) can be explained by the addition of the pointing modality in the second phase, which has been shown to have a higher proportion of errors than the language modalities. It is unclear however, why there should be an increase in problems with the PVK condition.

A decrease in problems in the MK and PK conditions can be explained by the introduction of the voice modality. It has already been shown that voice is used much

more than keyboard when users are given a choice between the two, and generally produces fewer errors than pointing input. Consequently, an increase in the use of language as input, and a simultaneous decrease in the use of pointing, results in an overall reduction of problematic interactions. Similarly, little difference was seen in the proportions of interactions for MV and PV – the addition of keyboard interaction induces little keyboard use, and in general keyboard use has been shown to have few errors. It is much harder however to explain the constancy in the M and P conditions, even when voice is added, since logically, as in the MK and PK conditions, the addition of voice should reduce the number of problems, in particular in the P condition since voice accounts for 47% of all interactions in the second phase for that condition.

6.6.4 Proportion of modality switches

Next we wanted to examine the proportions of cases where users switch modalities when they encounter a problem, as opposed to the proportion of cases where they continue to use the same modality to resolve the problem. Table 15 shows the proportions of cases where the modality was changed (switch) to cases where it was not (same), for both phases of the experiment. We also show the figures for smooth interaction to better illustrate the changes that occur.

	Problem switches					**Smooth switches**			
	Phase 1		**Phase 2**			**Phase 1**		**Phase 2**	
	Switch	**Same**	**Switch**	**Same**		**Switch**	**Same**	**Switch**	**Same**
M	-	-	9	91	**M**	-	-	18	22
P	-	-	**24**	76	**P**	-	-	**28**	72
V	-	-	**29**	71	**V**	-	-	23	77
MK	17	83	4	96	**MK**	**24**	76	**28**	72
PK	3	97	5	95	**PK**	**31**	69	25	75
VK	15	85	**33**	67	**VK**	12	88	24	76
MV	**32**	68	**25**	75	**MV**	17	83	**31**	69
PV	9	91	9	91	**PV**	18	82	25	75
MVK	**35**	65	**35**	65	**MVK**	**31**	69	**27**	73
PVK	18	82	14	86	**PVK**	**25**	75	24	76
average	18%	82%	18%	82%	**average**	23%	77%	25%	75%

Table 15: Proportion of switches in problem and smooth interactions

We can see from the data that on average, users switch modalities about a quarter of the time during smooth interaction. However, the number of modality switches decreases to 18% in cases of problem interactions. This suggests that when users encounter a problem they prefer to try to resolve that problem using the same modality rather than immediately switching modalities. This contradicts findings cited by Bilici et al. in [6], which claim that users tend to switch modalities when they experience difficulties.

We can also see that the highest number of modality switches during problem interactions in P1 occurs in the MV and MVK conditions, while in P2 the P, V, VK, MV and MVK conditions have high amounts of switching. We consider a high number of switches to be anything above the average number of switches across all conditions. During smooth interaction, the highest number of switches occurred in the MK, PK, MVK and PVK conditions, while in the second phase, the P, MK, MV and MVK conditions had the highest number of switches. This shows that only the MVK, P, and MV modalities tend to exhibit a higher likelihood to switch modalities in general, but this conclusion is not very solid.

6.6.5 Nature of modality switches

Finally, we wanted to look at the nature of the switches when problems arise. We were interested in two aspects in particular. The first was if a language interaction fails, whether users tend to switch to the functionally equivalent modality, or whether they prefer to switch to a pointing modality. The second was which language modality users will switch to if a pointing modality is causing a problem. We hypothesize that given the influence of the traditional MK interaction paradigm, mouse users will tend to shift to the keyboard, and pen users to voice, since we have already established that there is a marked coupling between pen and voice use. In order to examine this, we considered only interactions in phase two of the experiment, so that only cases where users had equal access to all of the modalities were taken into account.

Table 16 shows the raw number of cases where each modality was used to resolve a problem in P2 for each condition and for each of the modalities available. For example, P→K means that a problem was encountered using the pen, and the keyboard was used to resolve it, while V→V means that a problem was encountered using voice, and voice was never the less used to resolve the problem.

	P→V	**P→K**	**P→P**		**V→P**	**V→K**	**V→V**		**K→P**	**K→V**	**K→K**
M	1	0	34	**M**	3	2	1	**M**	0	0	0
P	16	0	91	**P**	5	3	4	**P**	0	0	0
V	14	0	49	**V**	10	0	27	**V**	0	2	1
MK	1	0	33	**MK**	1	0	8	**MK**	0	0	0
PK	3	**6**	108	**PK**	1	1	11	**PK**	2	0	2
VK	13	2	63	**VK**	9	1	12	**VK**	2	3	1
MV	2	0	39	**MV**	5	3	6	**MV**	1	2	3
PV	6	0	74	**PV**	3	2	6	**PV**	0	0	2
MVK	4	1	23	**MVK**	2	1	2	**MVK**	3	0	0
PVK	7	1	222	**PVK**	6	2	3	**PVK**	1	1	1
total	67	10	736	**total**	45	15	80	**total**	9	8	10
percent	8	1	91	**percent**	32	11	57	**percent**	33	30	37

Table 16: Modalities used to resolve problems in P2 for each condition

We can see from the table that in general, when voice interaction fails, users prefer to switch to the pointing modality that is available rather than to the functionally equivalent keyboard. This trend is particularly marked in the V and VK conditions, and in no cases are there more switches to keyboard than to a pointing device. In keyboard interaction however, there is an almost equal split between changes to the functionally equivalent modality of voice, and to pointing. Moreover, 4 out of the 6 conditions in which any type of modality switch occurs exhibit switches from keyboard to voice use. These figures provide further evidence that voice is preferred over keyboard when users are given a choice, as is pointing, but there is not a clear preference between voice and pointing when a mistake is made using the keyboard. However, when a pointing modality is causing a problem, users overwhelmingly choose to use the same modality to resolve the problem, but if they do switch, overall it is most often to voice. This is particularly true for the P, V, and VK conditions. It is only the pen-keyboard combination that shows the opposite effect, with a switch to keyboard being preferred.

In an extension of the previous discussion, we wanted to look in more detail at whether users preferred to switch to a functionally equivalent modality or not. Table 17 shows, for each phase and condition, the proportion of switches that occur to a functionally equivalent modality (same) and those that do not (different).

	Phase 1		**Phase 2**	
	Same	**Different**	**Same**	**Different**
M	-	-	95	5
P	-	-	77	23
V	-	-	72	28
MK	83	17	96	4
PK	68	32	94	6
VK	-	-	71	29
MV	68	32	86	14
PV	90	10	92	8
MVK	75	25	78	22
PVK	84	16	91	9
average	67	33	85	15

Table 17: Proportion of switches within and between categories

We see that in P1, there is a general tendency not to change categories, whether by using the same modality or switching to one that is functionally equivalent (in the case of language use), and this trend is strengthened in P2. Thus, if a user is using language, they will keep using language if they encounter a problem (whether it is the same modality or a different one), and if they are using a pointing device, they will keep using that pointing device. Further investigation, which was not possible with the data at hand, would be

needed to determine whether this results from the user's personal interaction preferences, the interaction strategies that they are adopting, or some other factor.

6.6.6 Conclusions

When investigating cases of problems in interaction and the impact that they have on modality use, we made some interesting discoveries. The first was that between all of the input modalities tested, pointing modalities produced the highest number of errors, with pen use producing more errors than mouse use. Between the language modalities more errors were produced by voice than by keyboard. We also looked at the distribution of the types of problem cases (misses or scope problems) and found that in general, there are more misses than scope problems, which is an encouraging trend for the inclusion of language interaction in interfaces such as Archivus.

We also found that even when all input modalities were available users tended not to switch modalities when faced with a problem – on average, only 18% of all interaction in problem situations resulted in switches in modalities. This was particularly true when a pointing modality was responsible for the problem. However, if they did switch, users tended to switch to voice rather than keyboard. If voice use failed, the preference was to switch to a pointing modality, but if keyboard use failed, there appeared to be no particular preference between switching to voice or to pointing. Thus, we can see that the keyboard was never preferred as a back-up input device. Moreover, we found that there is a trend not to switch categories (for example from pointing to language, or vice versa), but this result is likely overly influenced by the amount of users who choose not to switch modalities.

6.7 Functional equivalence

In our work we introduce *functional equivalence* which is the notion that the exact same actions can be performed by two different modalities, and that these actions are then processed in exactly the same way by the system. In Archivus the pen and mouse are functionally equivalent, as are voice and the keyboard. Although the designers of the system knew that mouse and pen, as well as voice and keyboard, are functionally equivalent, we were interested in knowing whether users perceived them as such. We assumed that if users who used the mouse used it in a similar way to those who used pen, then the modalities were indeed perceived as equivalent. The same would hold true in the case of voice and keyboard.

6.7.1 Mouse vs. pen

Given that mouse and pen input are functionally equivalent, and that pen users were explicitly told by the experimenter that they could use the pen the same way as they would use a mouse, we expected the number of interactions made with each of those modalities to be quite similar. If this was not the case, that would imply that there is a novelty factor introduced by the pen modality that alters how users use it. When comparing pen and mouse interactions, we noticed that users make more interactions with the mouse than they do with the pen. Table 18 shows the average numbers of interactions (in raw numbers) for each modality in each phase per condition.

	Phase 1			Phase 2		
	M	**V**	**K**	**M**	**V**	**K**
M	**148**	-	-	79	22	3
P	132	-	-	45	30	3
MV	**50**	51	-	**56**	42	6
PV	34	50	-	36	40	4
MK	**77**	-	16	**44**	37	3
PK	56	-	14	32	37	5
MVK	18	43	2	33	42	2
PVK	51	26	6	66	20	7

Table 18: Average number of interactions per modality

In order to eliminate as many confounding factors as possible, we focus our analysis on parallel pairs of conditions where only the pen and mouse vary - M/P, MV/PV, MK/PK and MVK/PVK. We will focus our analysis on those values in the table which are marked in black. In these conditions the use of voice and keyboard are almost equivalent between the pairs, which allows us to do an accurate comparison between mouse and pen interactions for each pair. However, in some conditions the voice and keyboard interactions vary too much between the pairs to be able to draw any significant conclusions about the contrast between pen and mouse use. We include them for completeness, but mark them in grey.

The data shows clearly that mouse users make more interactions than pen users. In the case of the M/P pair in P1 we can see that there are more mouse than pen interactions, even though users in the P condition answered more questions than those in the M condition. Data from the MV/PV and MK/PK conditions further supports this conclusion. However, in both the MV/PV and MK/PK pairs, the mouse users had never the less answered a higher number of questions, which was not the case for the M/P pair. In order to see whether there were never the less more mouse interactions occurring we calculated how many more questions on average these users answered (1.5 for the MK condition, and 0.75 for the MV condition) and then calculated on average how many interactions are

made per question in each of those conditions (7.8 for MK and 10 for MV). There were 21 more interactions using mouse than pen in the MK condition, but only around 12 of those interactions can be explained by the additional questions answered. This means that regardless of the difference in questions, there were still more interactions being made with the mouse than with the pen. The same is true to the MV/PV condition, where only 7.5 of the 16 additional interactions can be explained by the extra questions done.

We hypothesize that the lower number of interactions with pen is due to the fact that because it is a novel input modality and perceived differently from the mouse, users are more careful with how and when it is used, and possibly pay more attention during the tutorial, learning the system more carefully as a consequence. The combination of these two factors then results in fewer interactions needed to achieve the same results. Another possible explanation is that users are accustomed to clicking, and the interaction happens on a horizontal plane, which makes it relatively easy and cost-efficient. With the pen, the physical aspect of interaction is new (users have to figure out the most convenient way to switch between the pen and the keyboard for example) and the interaction requires changing planes, which might make users less likely to over-use the modality.

6.7.2 Voice vs. keyboard

We also wanted to see if voice and keyboard are viewed as functionally equivalent by users. To do this, we looked at how these modalities are used when they are in isolation with a common pointing modality. Table 19 shows the proportions of pointing and language interactions between MV/PV and MK/PK in P1. We chose to use proportions in this case rather than raw numbers as we did in the pen/mouse analysis since in these conditions the differences in the raw numbers of pointing interactions were too great to allow us to draw conclusions about the associated language interactions. Furthermore, in this case we only look at P1 interactions in order to avoid cases where both voice and keyboard were present in the same condition since their co-occurrence might alter user attitude towards them.

	Pointing	**Language**
MK	82	18
MV	43	**57**
PK	80	20
PV	40	**60**

Table 19: Proportions of pointing and language (MV/PV & MK/PK) in P1

From the results in the table we can see that even though there is a slight tendency to use mouse more than pen in each of the cases, the proportions are similar enough to be able

to draw some conclusions about how users view voice and keyboard, and that in fact, voice and keyboard are not seen as equivalent modalities. The data shows that voice is much more likely to be used than the keyboard, and that in fact, there is a slightly higher tendency to use language with pen as a pointing device rather than the mouse. A possible explanation for this is the novelty of using voice as an input modality, which encourages users to use language. Another is that while users are used to speaking full natural language sentences, they are much less used to typing them in a searching and browsing context, since most internet browsers for example are based on keyword or simple phrase searches rather than full semantically driven natural language searching.

6.7.3 Conclusions

From this data, we can conclude that despite being functionally equivalent input modalities, neither pen/mouse nor voice/keyboard are perceived as such. In the case of voice and keyboard, this result is not particularly surprising. First of all, users are used to interacting with the keyboard in a certain way in search and browsing applications. This type of interaction usually involves only sporadic use of the keyboard in comparison to pointing since command and control actions for the interface are not available with the keyboard. Moreover, these interactions primarily involve the use of keywords and very short phrases using a specific syntax rather than full free-form phrases. Finally, keyboard use forces a physical effort that is neither natural nor quick for novice users. Voice however is both natural and quick as an input modality, since users are accustomed to using it to communicate in their daily lives. Moreover the use of free-form natural language is natural with the use of voice, whereas short phrases and keywords are less natural, since voice is rarely used to communicate this way in regular use. These differences easily account for a large gulf between how voice and keyboard interaction are perceived by a user, even though they are functionally equivalent at the system level.

Although the difference seems to be less marked than that of the voice/pen pair, the fact that pen and mouse use are not perceived as functionally equivalent is much more surprising. The way that the physical artefacts themselves are used is quite similar in the sense that both are tangible and in most cases have to be reached for in order to be used. Moreover, even though mouse use might be more familiar, pen users had been explicitly told that the pen could be used in the exact same way as a mouse. We hypothesize that the primary reasons for this difference are the novelty of using a new type of input device, and the problems users had in establishing the correct angle and pressure needed to make selections with the pen, factors that are not present when using the mouse.

6.8 Task completion

Task completion is one of the most common ways to evaluate user interfaces since it allows evaluators to determine how successfully an interface allows users to complete their tasks [7, 115]. In our case however, it was not the interface *per se* that we were evaluating, but rather which modality combinations (experiment conditions) allowed users to complete their tasks most effectively[6]. We use 3 different factors to measure user performance. The first is what we call the *success score*. This is a normalized overall measure of how well the user did in answering questions. The second factor is how many questions they answered in the allotted time. This gives an indication of whether some modalities are 'faster' than others. The final factor is the *correctness score*, which is a measure of how correct, on average, users were when answering questions. The results from these three measures, when taken together give an indication of which modality combinations users found to be the most effective for using the Archivus system.

Given the widespread use of the mouse and keyboard and the fairly short amount of time with which users had to familiarize themselves with the system, we expected that the MK condition would be the most effective in finding the answers to the questions.

6.8.1 Number of questions completed

We first wanted to look at whether there were any marked differences in the average numbers of questions completed in each of the phases, and overall, for each of the conditions (Table 20).

	Phase 1	Phase 2	Total		Phase 1	Phase 2	Total
M	9.75	9.845	19.625	**PK**	9.125	9.625	18.75
P	10.125	8.625	18.75	**MV**	10.125	10.5	20.625
V	10.625	8.875	19.5	**PV**	9.375	9.375	18.75
VK	11.25	10	21.25	**MVK**	9.125	9.125	18.25
MK	10.625	9.625	20.25	**PVK**	9.375	8.625	18

Table 20: Number of questions completed per condition and phase

On average across all modality combinations, there were 9.95 questions answered in P1, 9.422 questions answered in P2, and 19.4 questions answered overall. The table shows that in the first phase, the highest number of questions were answered in the V (10.625) and VK (11.25) conditions. There are two plausible explanations for this. The first is that

[6] It is important to note that the experiments were designed to examine how users used the modalities, and NOT to directly test their performance with them.

since voice is a novel modality, users took more care in learning how to use it to manipulate the system during the tutorial. This resulted in them having a better understanding of both how to use the modality, and of how the system itself functions. The second reason is that the nature of the system is such that it takes fewer steps to reach a specific answer using language-only than it does using only a pointing device or a combination of language and a pointing device, which in turn implies a faster completion time for language modalities (similar results were found by Karl [70] in a study comparing voice and mouse input in a word processing task). With only a very few exceptions involving individual cases of interaction, this is true even when the time taken by the wizard to process the language input is taken into account. Taking both of these factors into account together, users in the V and VK conditions could answer more questions in the same amount of time as those using other modality combinations.

Meanwhile, the lowest numbers of questions answered were in the PK (9.125) and MVK (9.125) conditions. We think that in the case of PK, this is due to both the fact that it simply takes longer to switch between the pen and keyboard than between any of the other modalities, and the fact that users had a novel input device, the pen, coupled with the weaker half of the traditional MK input paradigm. It is possible that the fact that they had access to a pen, which behaves functionally in the same way as the mouse but is not as familiar to use was blocking them on a subconscious level. A similar explanation accounts for the low number of questions answered in the MVK condition. Users had access to the traditional paradigm of MK, but were also given voice as an input modality, which as has already been discussed, proved to be a tempting modality to 'try out'. Consequently, users were torn between the familiar and their desire to try a new modality, which impeded their overall performance.

Between the two phases there is a general decrease in the number of questions answered. We think that this is due to the fact that in most cases, the introduction of new modalities forces the user to slightly change the interaction strategies they have adopted in the first phase, particularly when a novel input modality is introduced. This results in slower interaction, since the processing time of the user is now also being dedicated to selecting from a larger number of modalities and to adapting a new strategy. Thus fewer questions are answered in the same amount of time. However, looking more closely at the data reveals that the decrease is not uniform. In fact, the number of questions answered in the PV and MVK conditions stays the same, and there is a very slight increase in the number of questions answered in the M, PK and MV conditions. The fact that there is no change in the MVK condition is likely due to the fact that since no modalities are added, users continue to work with the interaction patterns established in the first phase, resulting in

no change in the number of questions answered. A similar explanation applies to the PV condition. In section 6.4 we have shown that when given the choice, users prefer voice interaction to keyboard interaction when pen is the pointing modality involved. Consequently, the addition of the keyboard modality brings little added value to interaction with the Archivus system in this condition, resulting in users maintaining their interaction habits. The problematic case here is the PVK condition which in principle should give similar results to MVK since no modalities are added. However, the number of questions answered in this condition does drop, and we see no valid explanation for why this happens.

The rises under the M, PK and MV conditions can be explained though. In the case of MV, interaction rises because the MK traditional interaction paradigm is being introduced into the modality set. Consequently, users who already have some success with the MV combination experience an increase in confidence (even if only at the subconscious level) when they are also given access to a keyboard, resulting in a higher score. The same phenomenon is responsible for the rise in the M condition, only in this case, it is further boosted by the fact that introducing language as an input modality decreases the number of interactions that a user has to make with the system to achieve the same results, which gains them time that they can use to answer more questions. This hypothesis is supported by the work of Grasso, Ebert, and Finin who in [21] show that multimodal interfaces involving speech demonstrate shorter task completion times.

Moreover, it appears that this gain is sufficiently large to account for any learning factors that arise by the introduction of voice as an input modality. Given these facts, one might assume that there should also be a rise in the P condition. We see however that this is not the case, despite the fact that mouse and pen can be used in exactly the same way in the system. The reason for this is likely the fact that the pen is a novel input modality, to which another novel input modality (voice) is added along with the familiar but infrequently used input modality of the keyboard. Thus, users have to learn how to use two novel modalities together, which slows down the interaction, without having the advantage of the traditional interaction paradigm that is present in P2 of the M condition. The same explanation also applies to the PK condition.

In the second phase the MV (10.5) and VK (10) conditions answered the highest number of questions while MVK (9.125), PVK (8.625) and P (8.625) answered the fewest. The reasons for the low numbers in the MVK, PVK and P conditions have already been explained in previous paragraphs, as has the reason for a high number of questions answered in the MV condition. The high number of questions answered in the VK

condition is more difficult to account for. The only explanation that we can see is that the users have already established good interaction methods using language. Consequently, although the addition of a pointing modality normally decreases the number of questions answered due to an increase in the amount of interactions necessary to achieve the same results, in this case, the success of the VK interaction is greater than the decrease introduced by pointing, resulting in an overall continued increase. The difficulty with this explanation is that it does not account for why the number of questions answered in the V condition drops much more.

When looking at the overall (for both phases) scores however, which we feel are more indicative of the long-term behaviour that will be established for each condition, we see that the highest numbers of questions answered were in the VK (21.25), MV (20.625) and MK (20.25) conditions. As an interesting aside, the P, PK and PV conditions all had the same overall number of questions answered (18.75).

6.8.2 Success scores

While the number of questions that users can answer in a given amount of time with each modality is important, how well they answer those questions is also important, and can be examined in two ways. The first is to look at the total scores for each user in each modality condition over a fixed period of time. The other is to look at how well they scored on average at the level of an individual question.

During the experiments, users had to write their answers on question cards. Their answers were then marked as correct, not answered (if the card was left blank), or incorrect immediately after the experiment was finished. However, given the large proportion of non-native English speakers (85%) with varying competencies in English, we were worried that problems independent of the system, such as users misunderstanding or misinterpreting a question or the content of the database might lead to results that are not necessarily representative of how these users would use the system if it were in their native language. Results discussed in section 6.8.5 show that native English speakers seem to have a slight advantage while using the system. Additionally, we noticed during the experiments that some users would find the right information, but simply not realize that they had found it, for example by not taking the time to read carefully enough. Finding the correct answer is directly relevant to the scenario of use for a system such as Archivus. The user should be able to find the information that they are looking for and have some certainty that this is in fact the correct information. However, in the strict sense of examining which modalities are used when interacting with the system, the distinction between correct and incorrect answers is not important, since it is the overall

interaction patterns that we are interested in. Consequently, we decided to focus on answers independent of correctness in the analysis presented here and tried to normalize for the two problems mentioned above by developing a 4-point scale on which each user answer was ranked. The scale is described below.

0 points – wrong place and wrong answer, wrong place and right answer (guesses)
1 point – no answer was given
2 points – user was in the right place but gave the wrong answer
3 points – user was in the right place and gave the right answer

We felt that a distinction between 0 and 1 point was necessary to separate those who made guesses, as opposed to those who realized that they could not find the answer and gave up. Since users were told that they could give up, and that they would not be penalized for it, guessing should not have taken place. A single annotator was responsible for generating all of the normalized scores using the raw scores on the question cards and video data from the experiments. A user's success score was calculated by adding the scores for all of the questions that they completed. Table 21 shows the success scores, as an average of the total scores for each user in a condition, for each of the two phases, and overall for each of the conditions.

	Phase 1	**Phase 2**	**Total**		**Phase 1**	**Phase 2**	**Total**
M	23.75	22.5	46.25	**PK**	24.625	24	48.625
P	28	22.25	50.25	**MV**	26	26.25	52.25
V	27.75	24.25	52	**PV**	25.875	22	47.875
VK	27.875	23	50.875	**MVK**	24.125	19.5	43.025
MK	28.375	23.5	51.875	**PVK**	24.625	19.125	43.75

Table 21: Success scores (averaged over a condition) in each phase and overall

We can see from the table that during the first phase, the MK (28.375), P (28), VK (27.875) and V (27.75) conditions had the highest success scores, while the M (23.75) and MVK (24.125) conditions had the lowest. It is not surprising to see that the MK condition was the most successful in the first phase. This is the condition that all users are already familiar and relatively comfortable with. There was no shift from their standard interface interaction paradigm, so they could focus on understanding the system and its content and the introduction of an unfamiliar modality did not interfere with this. In the V, VK and P conditions, we believe that the reason for the high success rate is two-fold. First, the users were using new modalities with a new system. This might have led them to pay more attention during the tutorial to what the modalities could do, and how they impacted the system. The fact that they paid more attention would then lead them to a more careful and planned use of the system, resulting in more overall success.

The second reason is that in fact all 4 of the conditions that had the highest scores also answered the greatest number of questions in the first phase. One might assume that this link in and of itself is the explanation for the higher scores, but examination of the conditions with the lowest success scores shows that this is not necessarily the case, and that how users approached each of the modalities is also a factor. Specifically, users in the M condition answered the 5th largest number of questions (out of the possible 10 conditions), even though they had the lowest success score. This suggests that it is not the quantity of questions alone that is responsible for the higher scores. Similarly, the P condition got the second highest score, but only the 4th largest number of questions. An explanation for the surprisingly low success score in the M condition can likely be attributed to a false sense of security with mouse use, and over-zealous clicking. Users are so familiar with mouse use, that they subconsciously establish certain interaction paradigms that they follow when using it, and their interaction paradigms are likely to appear even in cases where the interface might be of a different nature than what they are used to. This results in users making interactions with the interface that are inappropriate or finding results that are not entirely what they were expecting, which in turn leads them to answer questions incorrectly, resulting in the success score/question quantity discrepancy that can be seen in the data. In the case of the MVK condition on the other hand, we believe that the problem lies in the addition of another input modality, in this case voice, to the traditional MK interaction paradigm. We believe that users are tempted to try voice (they use it 32% of the time in this condition), but because of the presence of the mouse and keyboard, they are less careful in learning how it can be used with the interface, since they know that they will not have to rely on it. Not learning to use voice as properly as in other cases, coupled with the (apparently false) sense of security that the MK modality combination brings, results in both low success scores and a low number of questions answered.

In the second phase, we no longer see the correspondence between the number of questions answered and a high success score. During the second phase the highest success scores were obtained by the MV (26.25), V (24.25) and PK (24) conditions. For example, V condition users, while getting the second highest success score, only answered the 8th largest number of questions, while the VK users, who had the second highest number of questions answered, had only the 5th highest success score.

There are several other points that are interesting to note in the second phase. In all cases except those of MV and PK (two of the top 3 success scorers), the score drops between the first and second phases of the experiment. In the MV and PK conditions, the score stays virtually the same. Thus, it appears that in fact these two conditions don't really

earn a high success score status in P2 due to an improvement in the strict sense of the word, but rather due to the fact that all of the other conditions decrease dramatically in success score by comparison. This notion seems to be supported by the fact that there is no definitive correlation between success score and quantity of questions answered. Moreover, due to a methodological oversight with about a quarter of the users, not all of them answered the same questions in P2. Since the effect was the same independent of the questions being answered, we know that it is not due to question difficulty in P2. Rather, we think that it is due to the fact that by the time users have reached P2, they are already somewhat familiar and comfortable with the system and, as a result, are less careful in the steps they take to search for an answer and in discovering that answer in the content. They have more confidence in their use and knowledge of the system, are used to search engines such as Google giving the most pertinent answers first, and don't always verify that this is the case with the results which Archivus provides, which are more complex and of a different nature than those of internet searches.

The V condition is the only one that does not drop significantly enough to alter its high success score. Given its high success score and the fact that relatively few questions were answered, we are led to believe that the explanation is in how users approach the use of the modality. The addition of the new modalities induces hesitation as to which modalities to use in which situations, which means that user interaction with the system becomes slower, resulting in fewer questions being answered. Moreover, the use of a pointing modality implies more steps to be taken in an interaction, which in turn implies that more time is needed, again resulting in fewer questions being answered in the same amount of time. However, experience with the system, and possibly a more thorough knowledge of it due to the voice-only interaction in the first phase (as explained above), could be resulting in a higher number of correctly answered questions, thus accounting for a higher success score.

The lowest success scores in P2 on the other hand were obtained by the MVK (19.5) and PVK (19.125) conditions. We believe that the explanation for the MVK condition is the same as it was in the first phase, and that the drop in score is a result of users being less careful in their analysis of resulting answers, since the number of questions answered for this condition is the same in both phases. The explanation for the poor results for PVK lie along similar lines. Here, users have two modalities that they are unfamiliar with, voice and pen, and one with which they are familiar, but that is not used as often in browsing and searching interactions. As in the case of MVK, the presence of the keyboard provides a backup that users can resort to if the pen or the voice fail, so they are less careful in how they use the modalities and in how they can be used in the interface, resulting in

both lower success scores and fewer number of questions answered. It is important to note that in P1 PVK was tied for third worst success score.

Looking at each phase of the experiment separately allows us to observe the evolution of a user's ability to answer questions using Archivus, but it is with the overall figures that we gain a more concrete picture of which conditions might be most successful in the long term. The highest success scores overall were gained by the MV (52.25), V (52), and MK (51.875) conditions. These results support the theory that although MK remains a strong interaction paradigm for this domain, the use of voice, particularly when introduced on its own or with the mouse, the stronger half of the MK traditional paradigm, is used and useful for interaction with the system.

6.8.3 Correctness scores

Success scores give an idea of how well users did in general when answering questions over a fixed time period, but it is also necessary to look at how well, on average, they answered individual questions. We call this value the *correctness score*, which is calculated as the average success score for a given condition divided by the number of questions answered in that condition. Recall that since the normalized scale for determining user answers had a maximum of 3 points, the maximum correctness score is also 3. This type of score is important in order to distinguish between cases where the success score is the same or similar, but user behaviour is not. For example, the same success score can be achieved by users who answer a larger number of questions poorly, and users who answer fewer questions, but more correctly. Table 22 below shows the correctness scores for each condition per phase and overall.

	Phase 1	Phase 2	Overall
P	2.8	2.6	2.68
V	2.6	2.7	2.67
PK	2.7	2.5	2.59
MK	2.7	2.4	2.56
PV	2.8	2.3	2.55
MV	2.6	2.5	2.53
PVK	2.6	2.3	2.43
MVK	2.6	2.1	2.39
VK	2.5	2.3	2.39
M	2.4	2.3	2.35

Table 22: Average per question correctness score ranked by overall score

The lowest correctness score was found in the M condition. This is likely due to the fact that people are overconfident or overzealous using the mouse, and take less care in how they go about finding the correct answer, resulting in fewer correct answers. The VK

condition, which had the second lowest correctness score at 2.39, might have suffered from a similar problem in that although the condition allowed users to answer the highest number of questions and get a relatively high score, they also made more mistakes when answering questions. However, it must be kept in mind that an average score of above 2 is still quite good. It means that at the very least, users were finding the right information, but may have had problems in interpreting it in order to answer the question correctly.

From the table we can also see that there is a decrease in the correctness score between P1 and P2 for every condition except V, where there is a very slight increase (0.1). This suggests that as users become more comfortable with the system, they gain a false sense of confidence in how to use it and the type of results that it can provide for them in different situations. As a result, they begin to make more mistakes. Since this phenomenon happens across virtually all of the modalities, it seems to be dependant on the relationship between the user and their knowledge of the system rather than the user and the modalities in question.

When both the success score and the number of questions answered are taken into account, MK and MV seem to be the most effective conditions for interacting with the system. However, if we look strictly at correctness scores P and V score the highest with scores of 2.68 and 2.67 per question respectively. Meanwhile, MVK and PVK seem to be the least effective both in terms of success scores and the number of questions answered, and are on the lower end of the scale for correctness scores (2.43 and 2.39 respectively).

6.8.4 Distribution of success among modalities

We also ranked the 80 individual users in the experiment by their success score and the number of questions that they were able to answer, and looked in more detail at the top and bottom 20 users for each case. We were interested in seeing what the distribution of the 10 modality combinations was in the top and bottom 20 scores, and in particular, whether there were more individual success or failure cases in the same conditions as the average figures from sections 6.8.1 and 6.8.2 indicated. Figure 19, shows the number of individual success (light bars) and failure (dark bars) cases for each condition.

We can see that in fact the number of top 20 individuals does not indicate which condition will be the most successful for overall scores. The M condition has only the 8th best success score despite having as many of the top 20 most successful individuals as the MV and MK conditions which are among the top 3 highest success scorers, and having more individual high scorers than the V condition, which had higher average success

scores. Similarly, having many low success scorers does not indicate an overall low success score as seen in the MK condition.

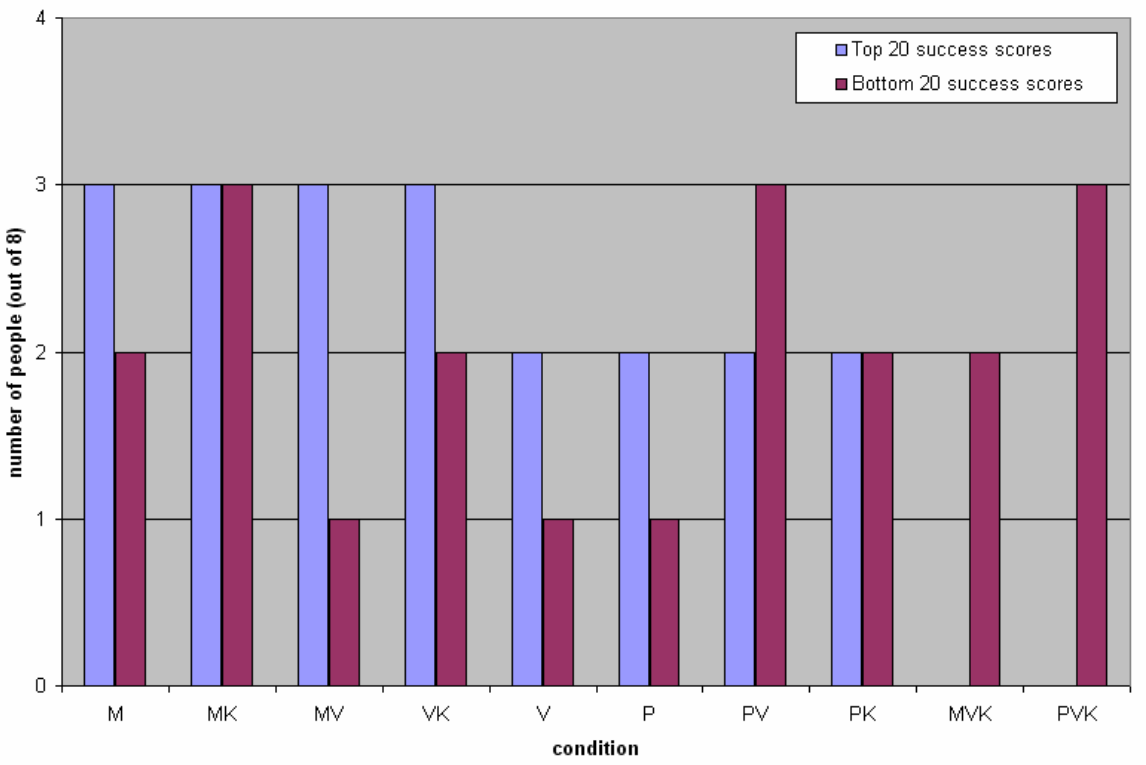


Figure 19: Number of instances of top and bottom 20 success scores per condition

We performed a similar analysis on the top and bottom numbers for questions answered. Figure 20, shows the number of instances of high question answerers (light bars) and low question answerers (dark bars) for each condition. From this figure we can see that no condition did particularly badly since there are no cases of 3 users in a condition for the bottom 20 number of questions answered. However, the PVK condition had no high question answers whatsoever.

Again, we see that the individual results are not indicative of overall results, since users in the M condition, despite having the same number of individual high and low question answers as users in the MK and MV conditions, placed much worse when looking at averages. Similarly, for low scores, PVK and MVK did not have more individual instances of low question answers than other conditions in which users generated more answers.

From these results we can conclude that the number of individual users who do well or poorly in a task using a certain modality combination is not indicative of whether that modality combination in general is useful or not for doing that task.

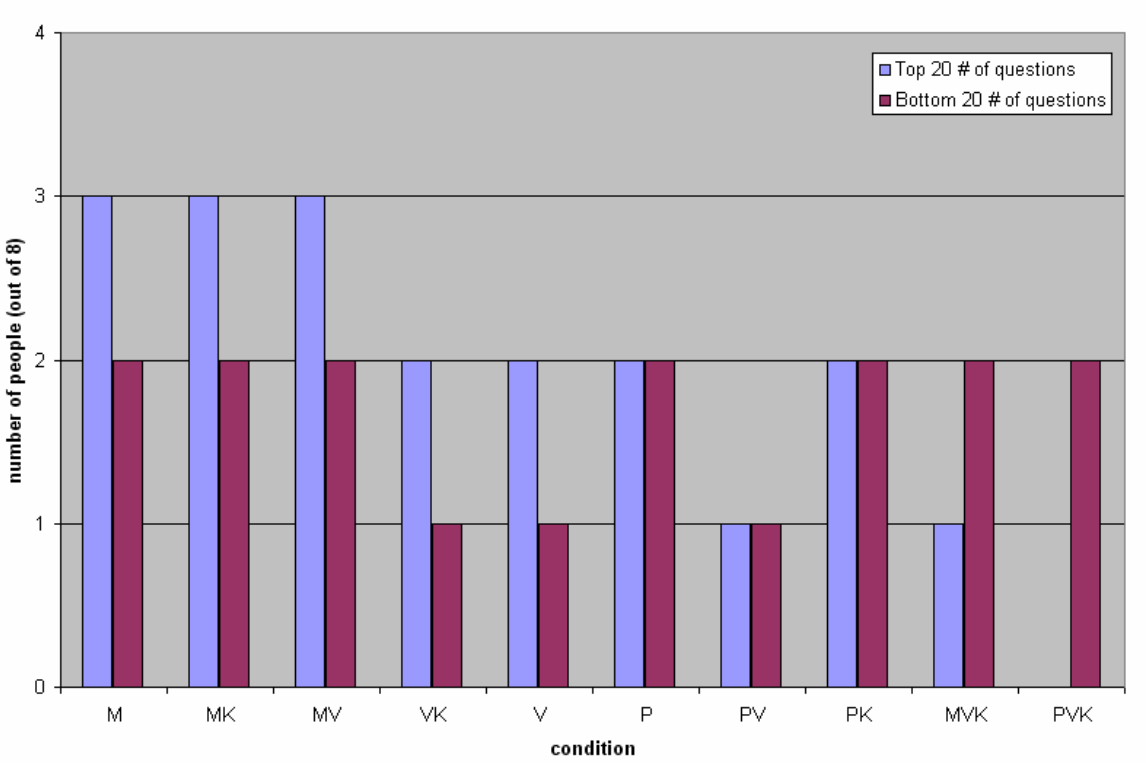


Figure 20: Frequency of top and bottom 20 for # of questions answered per condition

6.8.5 Influence of user background

Finally, one of the factors that we felt was important to look at was whether the users' personal background played a role in their overall performance with the system. In particular, we wanted to look at whether there were differences between native and non-native speakers of English, and between male and female participants, and whether experience with speech recognition systems and the amount of time that users spent with a computer on a daily basis had an influence on their success in finding answers. To do this we looked at the proportions that each of these factors held in both the top 20 and bottom 20 individual users for both the success score and the number of questions answered.

Native vs. non-native English speakers

Out of the 80 users in the experiment, only 12 (15%) were native English speakers. Out of those 12 users, 9 (75%) had top 20 success scores, and 8

(67%) had a top 20 score for the number of answered questions. Moreover, only 1 of the native English speakers had a bottom 20 success score, and the same person was also in the bottom 20 for the number of questions answered. This suggests that native English speakers have an advantage over non-native speakers when using the Archivus system. Unfortunately, we could not control for the distribution of native English speakers in the data set, so while some conditions have 2 native speakers (M, P, V, VK, and PVK), others have only 1 (MV and PV) and some have none at all (MK, MVK and PK).

Male vs. female performance

Out of the 80 users in the experiment, 33 (41%) were female. Only 6 of them (18%) had top 20 success scores, as opposed to 14 (30%) out of the 47 males, whereas 10 (30%) had success scores in the bottom 20, as opposed to 10 (21%) out of the 47 males. Of the top 20 question answerers, only 5 were female (15%) as opposed to the 15 males (32% of the male population), and from the bottom 20, 9 (27%) were female as opposed to the 11 males (23% of the male population). These figures suggest that women do a bit worse than men when using the system. However, since the female users were distributed fairly evenly among the different conditions – most conditions (M, P, MK, PV, MV and PVK) had 3 females each, one condition (MVK) had only 2 females, and three conditions (V, VK and PK) had 4 females – we don't think that this had a significant impact on the results[7].

Previous experience with speech recognition systems

32 out of our 80 users (40%) had controlled a computer with their voice before. Out of those, 9 (28%) had a top 20 success score, and 8 (25%) had a top 20 questions answered score, while 7 (22%) had a bottom 20 success score and 7 (22%) had a bottom 20 questions answered score. The differences between these numbers are not large enough to conclude that previous experience with speech recognition systems played a role in user performance.

[7] Note: we did not have data as to the educational background and professional experience of the users, so a more thorough analysis was not possible.

Amount of time spent daily with a computer

Out of our 80 users, 28 (35%) spent more than 7 hours per day using a computer. From those, 9 (32%) had success scores in the top 20, and 8 (29%) had scores in the top 20 for the number of questions answered. But, 7 (25%) also had success scores in the bottom 20, and 8 (29%) were in the bottom 20 for the number of questions answered. Conversely, 33 of the users (41%) spent less than 4 hours per day with a computer. From those, 7 (21%) had success scores in the top 20, and 6 (18%) were in the top 20 for the number of questions answered, while 9 (27%) scored in the bottom 20, and 6 (18%) in the bottom 20 for the number of questions answered. While there is some variation in the number of users for each case, the differences are not large enough to say with certainty that experience with a computer was a significant factor in how successful users were with Archivus.

Overall, we can conclude that while native English speakers did have an advantage during the experiments none of the other factors played a role in the performance of the individuals involved in using the system. This is shown by the relatively low percentages, which indicate that the users who did place in the top and bottom 20 did not constitute a significant proportion of the population in question.

6.8.6 Conclusions

In order to consider efficiency in a holistic sense, we need to take into account all 3 factors for task completion – the number of questions answered, the success score and the correctness score. Given that in all of the conditions users had access to all of the modalities in the second phase of the experiment it is difficult to determine which particular modality combinations would have been most successful if additional modalities had not been added. Based on performance in phase one, we can hypothesize that MK, P, V and VK would have been the most effective, since each appears in the top 4 for at least two of the three factors.

We can also use this data to hypothesize about how the modalities and the system should be introduced to the user by looking at the overall results for the period of time including both phases. These results show that both the MK and MV conditions have high scores and high numbers of questions answered, and their correctness scores are 2.56 and 2.54 respectively, while V has a high correctness score and a high success score. This makes these conditions the most effective to learn to interact with Archivus in order to achieve the best results in the long term. In this conclusion we can also discount the influence that

the presence of native English speakers has since each of the conditions had 0 (MK), 1 (MV) or 2 (V) native English speakers.

Finally, we found MVK and PVK to be the least effective modality combinations. They consistently placed towards the lower figures in all three factors in both phases, and since there were no modalities added over time, we can conclude that providing all three of the modalities at once is not very helpful to the user. An interesting question that poses itself here is whether the problem lies in learning with all three modalities or the general availability of all three modalities. If it is a question of the availability of all three modalities, then we can draw the further conclusion that in fact, the MK and V conditions are the most useful overall for interacting with Archivus, since they are the only modalities that appear to be strong both in P1 and overall. We can also go further and claim that in this case, the MV condition which was also strong overall would be effective since in fact the performance in that condition appears to get better over time while data shows that the addition of keyboard has little impact on the interaction (keyboard is used only 5% of the time in phase two). However, in their work Le Bigot, Jamet, and Rouet [53] did not find that the use of voice lead to an increase in performance, which leads us to conclude that a more detailed analysis is necessary.

6.9 Modality use and task type

The final aspect that we were interested in looking at was whether there were any correlations between a particular modality, or modality combination, and the type of task being solved. By the type of task we specifically refer to the type of information that the user is being asked to find in order to answer one of the questions from the question set. In our work we used a classification of the Archivus question set which was developed by Marita Ailomaa, who was one of the people involved in the design and development of the Archivus system and participated as a wizard in the experiments described in this thesis. The classification can be found in Appendix H. However, it must be noted that the question set was developed to be as varied as possible in order to test the different capabilities of the system with the different data types that were available in the database. Moreover, while we tried to distribute the question types evenly among the phases, we discovered that the amount of questions answered per phase during pilot experiments to finalize the experimental protocol was much higher than the number of questions answered by users during the actual experiment. As a result, in some cases very few questions of a particular type were answered, and sometimes none at all, or a type of

question was answered in one phase, but not in the other. Despite these drawbacks, we wanted to see whether any trends were apparent.

We first looked at whether there were any similarities in the modality combinations that scored better or worse than average for each individual question. To do this, we calculated the average and standard deviation across each of the 10 conditions for each question that was done by at least one user in each of the conditions. This resulted in 11 possible questions in P1 and 10 possible questions in P2. For each question, we then noted which modality combinations scored higher and lower than one standard deviation. We were looking for example for cases where within a single question, all or most conditions that involved voice scored consistently better than those that did not. We did not find evidence for this in any of the questions in either phase. The only noteworthy result was that of question 29 in phase 2 (*Denis showed 4 possible versions of the movie club advertising poster at the meeting*) where the M, V and P conditions did better than all of the others. However, we don't believe that this result is significant since questions of a similar type did not produce the same effect. We also examined whether there were similarities in the behaviour of certain conditions at a more abstract level, namely across similar question types. Again, we did not find this to be the case.

Additionally, we performed the same analyses as those described in the previous paragraph, but looking at the average (across all participants in a modality combination) total number of interactions that were used by each condition to complete a question. Again, we noted instances where the total number of interactions was greater or smaller than one standard deviation. In this case, we found that the M and P conditions consistently used a higher number of interactions (in 64% of the questions for M and 55% for P) to solve questions in phase one, while users in the PVK condition often used a lower number (36% of the questions). In the second phase, the M condition continued to exhibit the trend of using more interactions (in 40% of the questions), but while P normalized, MVK also showed an increase in the number of interactions used (also in 40% of the questions). The MK and MV conditions also showed fewer interactions used (in 30% of the questions). All other conditions involved 20% or fewer of the questions. Finally, we looked at whether any of the question type sets evoked similar high/low trends in the number of interactions, and again found that they did not.

The preliminary data discussed in this section suggests that performance on different task types is not influenced by modality choice, although further experiments targeting specifically this problem would be needed to verify these results.

6.10 Conclusions

In this chapter, we have analyzed the data from the Wizard of Oz experiments with the Archivus system and discussed the results. In particular, we looked at the users' subjective opinion of the system, learning effects with different modalities, proportions of modality use and their evolution over time, how users perceive functionally equivalent modalities, task completion, and the relationship between task types and the modalities used during interaction with them. In the following chapter, we begin by reminding the reader of the most important findings, and then discuss whether or not multimodal interaction brings added value to the multimedia meeting browsing and retrieval domain.

7. Conclusions

In this chapter we will discuss how the results from the experiments, presented in Chapter 6, help answer the research questions laid out in Chapter 3 and draw some conclusions about whether multimodal input brings added value to interaction in the multimedia meeting browsing and retrieval domain, and if it does, what the nature of that interaction is. In particular, we were interested in the benefits and drawbacks that novel input modalities such as voice and pen bring to interaction, especially in the presence of more familiar modalities such as the mouse and keyboard. We had chosen to make the assessment based on the answers to 6 central questions:

1. How often are different modalities used, alone and in combination, for meeting browsing and retrieval tasks?
2. Are some modalities more suited to finding certain types of information than others?
3. Do certain modalities or modality combinations make the system easier to learn, leading to an increase in performance in the long term?
4. Does modality use change when a user encounters a problem during interaction?
5. How do users perceive different modalities?
6. Does learning to use a system with a particular set of modalities influence how those modalities are used when other modalities also become available?

We remind the reader that we had defined *added value* in terms of increased performance when compared to standard mouse and keyboard input, the usefulness of multiple modalities, and overall subjective user satisfaction when interacting with an interface. We had also stressed the importance of three concepts that are central to our analysis – 1) familiar vs. novel input modalities, 2) the traditional interaction paradigm (TIP) of mouse and keyboard use, and 3) functional equivalence.

In the following sections, we first summarize the most pertinent findings from the experiments as they apply to each of the research questions before discussing, in section 7.6, whether they support the claim that multimodality brings added value to interaction in the multimedia meeting browsing and retrieval domain. We will not, however, address question 2 from the list above since, as explained in section 6.9, the experimental

protocol did not allow us to gather a sufficient amount of data for statistically significant analysis of the correlation between modality use and task types.

7.1 Use of modalities

Understanding how different modalities are used in various multimodal input situations plays an integral role in assessing their usefulness. In our work we were particularly interested in how often different modalities were used during interaction, whether their use was dependent on co-occurring modalities, and how their use evolved over time. We found that in general:

- the mouse was the dominant modality in the traditional interaction paradigm
- the combination of two novel input modalities encourages the use of novel modalities
- among pointing modalities mouse (familiar) was always used more frequently than pen (novel)
- among language modalities voice (novel) was used more frequently than keyboard (familiar), particularly when combined with the pen as a pointing modality
- the traditional interaction paradigm helped reinforce keyboard use

When examining the evolution of modality use we found that:

- the introduction of a secondary language modality caused more disruption in modality use when the pointing modality was the mouse than it did when the pointing modality was the keyboard
- there was a general trend for voice use to decrease over time, which contradicts other studies [27]
- there was a general trend for keyboard use to decrease over time, which became much more marked when voice was introduced
- when comparing the V and VK conditions, voice and pointing use converged to the same levels, but the convergence was faster in the VK condition
- in the MVK and PVK conditions the rise and fall of different modalities was not gradual over time and the relationship between modalities was not steady over time

- pen use was much more stable over time than mouse use

7.2 Task completion

How well tasks are completed using a modality or set of modalities is an important factor in determining how easily a user can learn to interact with a system given those modalities, and how useful they are. Consequently, we were interested in examining whether certain modality combinations lead to an increase in performance when using a multimedia meeting browsing and retrieval system. We found that:

- in terms of overall performance, the MK and MV conditions seemed to be the most effective while the MVK and PVK conditions seemed to be the least effective
- for success scores over both phases the MV, V and MK conditions had the highest scores while MVK and PVK had the lowest scores
- for correctness scores over both phases the P and V conditions had the highest scores while the M and VK conditions had the lowest scores
- for the number of questions answered over both phases the VK, MV and MK conditions had the highest number of questions while MVK and PVK had the lowest
- there was a general decrease in both correctness and success scores between P1 and P2
- the distribution of individual high and low scores was not indicative of which condition would do the best or the worst overall
- native English speakers had an advantage using the system

7.3 Problems and modality choice

It is reasonable to assume that during interaction with any system, users will eventually encounter problems that are caused by the modality that they are using. We were interested in investigating how users handled this type of problem in a multimodal context, the nature of the problems that they encountered, and which modalities were the most likely to cause problems. We found that:

- a large proportion of problems was produced by pointing modalities, and there were more problems with pen than with mouse use (despite

the fact that performance using pen was quite good in terms of questions answered and success scores)

- in problems produced by language modalities, there were more problems with voice than with keyboard use
- in general when users encountered a problem they preferred to resolve it using the same modality (very large proportion of cases), or one that was functionally equivalent
- when voice use failed users tended to switch to pointing rather than the functionally equivalent keyboard
- when a pointing modality failed, users were most likely not to switch modalities at all, but if a switch did occur, it was more likely to be to voice
- only the MVK, MV and P conditions exhibited a slightly higher likelihood to switch modalities when a problem was encountered
- pen interaction had more misses than scope problems, but the opposite was true for mouse interaction
- there were more misses than scope problems overall across all modalities

7.4 User perception of modalities

How users perceive and contrast individual modalities are important factors to consider when choosing input modalities for an application since they are likely to influence how those modalities will be used with it. For example, how users perceive the usefulness of a modality, determined from subjective user opinion via questionnaire responses, can help to highlight the types of tasks for which a user is, or is not, willing to use that modality. Also, although some input modalities may be functionally equivalent at the technical/system level, that fact does not imply that they are perceived as such by users. In terms of how users perceive the usefulness of modalities we found that:

- for interactions involving manipulation of the interface a pointing device was preferable, except when the items to be selected were small
- for specifying search criteria, language modalities were preferred, with voice being preferred over keyboard

In terms of functional equivalence, we found that:

- more individual interactions were made with the mouse than with the pen
- voice is much more likely to be used than keyboard

7.5 Learning effect

In our work we were also interested in investigating whether, for a new application, early experience with a particular set of input modalities had an impact on modality use during later stages of interaction, when other modalities were also made available. We found that:

- there was a clear learning effect in the M, V and VK conditions
- there was a learning effect in the P condition, but it was less marked
- there was no learning effect in the K conditions but the keyboard was never used in isolation as the mouse, pen and voice modalities had been

7.6 Conclusions

There are 12 general conclusions about the use and usefulness of different modalities for input in a multimedia meeting and retrieval interface that can be drawn from the results presented in Chapter 6 and summarized in the preceding sections.

1. The presence of the traditional input paradigm (mouse-keyboard interaction) has an impact on the relationship between modalities. For example, mouse use is less affected by the addition of other modalities than pen use is, and the presence of the mouse-keyboard combination during later stages of interaction for conditions where that combination was not available in early stages has been shown to increase keyboard use. This effect was not found with the pen-keyboard combination, which suggests that it is in fact the TIP that is responsible for the increase in keyboard use.

2. When the influence of familiar and novel modalities in functionally equivalent modality pairs is examined, the use of mouse (a familiar modality) is much more entrenched as a default input modality (when compared to the novel pen) than keyboard (familiar) use is when compared to the functionally equivalent (novel) voice modality. This implies that not all novel input devices are considered

equally novel. The novelty of using voice is greater than the novelty of using the pen.

3. Despite being functionally equivalent input modalities, neither pen/mouse nor voice/keyboard are perceived as such by users, as evidenced by the different ways in which those modalities are used when they co-occur with other modalities in parallel situations. Moreover, the difference is stronger in the case of voice/keyboard than it is in the case of mouse/pen for the reasons explained in section 6.7.3.

4. There is a high proportion of voice use at any point at which voice is included in a set of modalities. In the case where voice use is included from the early stages of interaction, this means that voice is also used in a high proportion during early stages of interaction. If voice is included at a later stage of interaction, then voice is used in a higher proportion immediately after its inclusion. This suggests that the novelty of being able to interact using voice is very tempting to users, independent of which other modalities might be simultaneously available. However, the trend for voice use to decrease over time also suggests that one of two things happens. Either the user's acceptability of voice as an input device changes and they become less tolerant of errors, prompting them to decrease their use of voice, or the novelty wears off as they spend more time with the system. Further study is needed to determine which of these factors is responsible, and at which point in time voice use stabilizes, reaching a plateau.

5. The addition of voice to a set of modalities has an impact on the use of modalities. Although this is true for all modality combinations, it is particularly the case when voice is added to a modality set in which the pointing modality is the mouse. We believe that this effect is due to the entrenchment of the use of mouse in western computer culture. Users have certain pre-conceived perceptions of mouse use and have established patterns for the use of the mouse through past experience. When users learn to use a new system with a modality set that includes the mouse, they apply interaction patterns that they have already establish through interaction with other systems to the new application. We have already discussed the fact that all users are tempted to try voice use due to the novelty introduced by the modality. When the temptation to use voice conflicts with the user's desire to use established interaction patterns, the result is a higher degree of disturbance in the flow of use, as the user wavers between the different modalities available to them. This is less the case with pen use because the interaction patterns that are

established during early stages of interaction with the system are not as entrenched as those with mouse use, resulting in a lower level of disturbance. Users will not have had the same degree of previous experience with the pen as they had with the mouse, and there will be no transfer effect from previous mouse use for pen users since we have already established that the pen and mouse are not perceived as functionally equivalent by users.

6. A novel modality, when combined with a modality that is also novel, is used more than if it had been combined with a familiar modality. For example, there is more use of voice when it is combined with pen than there is when it is combined with mouse. This also holds if a third (familiar) modality is present.

7. Observations of our data show that use of the pen modality is generally more stable throughout interaction and in particular when additional modalities are included at later stages of interaction. The wider implication here is that if designers are choosing modalities for an application for which they know that other modalities will be added in the future, the pen might be a better choice than the mouse to include as a base modality since users will be less disturbed by the addition of new modalities.

8. Language based modalities exhibit a particularly strong learning effect, meaning that users who learn to use a system with only language as an input modality are much more likely to continue to use language as an input modality even when other input modalities become available. The mouse-only condition experiences an equally strong learning effect. However, the learning effect in the pen-only condition, while still present, is much less significant. This leads us to believe that the learning effect in the mouse condition is only an *apparent* learning effect, and that in fact, the *actual* learning effect is similar to that of the pen. The external difference is due to the fact that users have previous experience with the mouse, which is applied to the new application and boosts the real learning effect. Le Bigot et al. [51] suggest that the effort taken to learn interaction using a certain modality will have an impact on the learning effect, which corroborates our findings. Language-only interaction is harder to learn when using a new system since there is less transfer from familiar modalities than there is in the pen case (see discussion in section 6.7.3).

9. Contrary to findings by Oviatt and Olsen [126] that users tend to switch modalities when they encounter a problem, we found that in fact users

overwhelmingly preferred to resolve a problem using the same modality. However, if a switch did occur, users tended to switch to a modality that was not functionally equivalent. For example, problems with a pointing device resulted in a switch to voice, while problems with voice resulted in a switch to a pointing device. Only problems encountered when using the keyboard seemed to exhibit no preference for the type of modality being switched to.

10. Keyboard use was very low in general across all modality combinations, and appeared to have little impact on co-occurring modalities.

11. In terms of overall performance (determined based on success scores, correctness scores, and the number of questions answered – section 6.8), the MK and MV conditions performed the best with the V condition following closely behind. This suggests that while the use of a pointing device is quite important when interacting with a system such as Archivus, it is not strictly necessary. Moreover, it shows that even though they are functionally equivalent, mouse as a pointing device is more successful overall than pen, despite the fact that it was shown that more individual interactions are made when using the mouse than when using the pen. Better performance with the mouse is most likely due to the fact that users have more previous experience with it. But, as Dix et al. [9] point out, it could also be attributed to the fact that during pen use the screen is partially obscured by the user's hand and arm, which can affect their performance. Meanwhile, the MVK and PVK conditions were found to be the least successful, which suggests that allowing users to have too many choices between modalities is not recommended.

12. User performance seems to decrease when additional modalities are added to the set of modalities with which the user already has experience for an application. This was demonstrated by a decrease in both success and correctness scores between P1 and P2 in 7 out of the 8 experiment conditions where one or more modalities were added in P2. However, the same effect can be seen for the remaining two conditions (MVK and PVK), where the same three modalities were available during both phases of the experiment. Further study is needed to determine whether the addition of modalities is truly detrimental to interaction (in which case the effect is due to a methodological problem in the current experiments since in the real world modalities are rarely added after an application has been launched) or whether another factor such as the user becoming tired or bored with the task is the cause.

When taken together, our findings lead us to the conclusion that while multimodality in the strict sense of the word (having more than two modalities available) does not give added value to interaction in the multimedia meeting browsing and retrieval domain, the use or inclusion of voice as an input modality does. However, as Cohen and Oviatt [16] observe based on a number of studies cited in the literature, '*it is not obvious why people should want to speak to their computers in performing many tasks – in particular, their daily office work*'. Since the multimedia meeting browsing and retrieval domain is quite similar in many respects to daily office work, further study is needed to determine the exact reasons for the willingness to use voice suggested in the results presented in this thesis. Some ideas as to how to go about doing this are discussed in the following chapter.

8. Future Work and Possible Extensions

In this chapter we discuss possible extensions to the work presented in this thesis and which are meant to complement the results found. The extensions are grouped into three general categories - 1) further analysis of existing data, 2) analysis requiring a change in the experiment protocol, and 3) analysis requiring a change in the Archivus system.

8.1 Further analysis on existing data

There are several analyses on the data set that resulted from the experiments described in this thesis which were not performed due to either time constraints or the fact that they were not tightly coupled with the research themes addressed in this work. However, we feel that they are never the less interesting themes that merit further investigation in future work, so we discuss them briefly here.

Influence of the tutorial

Given the problems in designing a tutorial that would not influence users in their interaction or modality choices (section 5.4), it would be important to determine the degree of influence that the steps and actions shown in the tutorial have on later interaction by users, and how much that would affect the results presented in this study. This type of analysis would have to be done on a per modality basis since tutorials for each modality combination vary slightly depending on the nature of the modalities involved.

Error recovery strategies with different modalities

In this thesis we discuss the proportions and types of errors made by users, and how they differ across different modalities and modality combinations. An extension of this work, inspired by the work of Halverson, et al. [28] would be to examine the error recovery strategies adopted by users, and whether there are any correlations between modalities and particular error recovery strategies.

Functionalities accessed and modality choice

Studies such as those by Whittaker and Walker [72] and Oviatt, de Angeli and Kuhn [122] suggest that certain modalities are favored for performing certain tasks or actions. The results from our post-experiment questionnaire, as described in section 6.2.2, indicate that the types of actions for which users *think* specific modalities are the most useful correspond with the results from these studies.

However, it would be helpful to perform an analysis on the users' actual actions rather than relying on their subjective opinion. Unfortunately, we could not perform this type of analysis on our experimental data due to time and technical constraints. The Archivus interface, being flexibly multimodal, provides the ideal platform for this type of analysis since any interface element can be accessed with any modality, and search criteria can be specified with any modality as well, which gives experimenters an accurate perspective on the user's choice. Furthermore, such an analysis would allow for an evaluation of users' interactions based on the CARE properties developed by Coutaz, Nigay and Salber [74], which form '*a simple framework for reasoning about multimodal interaction from both the user and the system perspective*'.

8.2 Analysis requiring a change in experiment design

In this section, we discuss additional studies that would require a change to the design of the experiment protocol.

Task completion times

Task completion time is often used as a measure to evaluate usability, and can also be applied when comparing modality combinations [42]. We did not analyze task completion times in our work because we had observed that users did not consistently signal the end of a task and the start of a new task, despite being explicitly asked to do so both in the experiment instructions document and by the experimenter. While task *completion* could be analyzed using post-experiment analysis of the video data that was recorded, the results were not fine-grained enough to draw scientific conclusions about the differences in task completion *times* across the different modality combinations. However, Cohen and Oviatt [16] suggest that user preferences for modalities may be influenced by time-to-input rather than by overall performance measures. Therefore, it would be worthwhile to develop a new evaluation protocol that would more strictly enforce the indication of task completion and thus allow for the investigation of whether such influences play a role in the results presented here.

Increased experiment length

In section 6.5, where we discussed the evolution of modality use, we saw that user behaviour with different modalities changed over time, and in particular that language use decreased over time while pointing use increased. Due to time and

resource constraints, we were not able to run experiments that were long enough to reveal the point at which interaction between modalities stabilizes. While we believe that the results presented in this thesis are a valid indicator of the direction that modality use will follow, longer studies to confirm these results would be welcomed. Moreover, studies over a longer period of time would give users in-context practice with the various modalities, and in particular with voice, which as Bell et al. [36], Le Bigot, Jamet, and Rouet [53] and Strum et al. [123] suggest, can influence user behaviour.

Interaction patterns

In our work, we have not looked at the nature of the interaction patterns (the order in which modalities are chosen, and for which concrete tasks they are chosen) that users adopt for problem solving during their interaction. This was done for three reasons. The first is that it speaks more to the evaluation of the Archivus system itself rather than to how modalities are used to access information using the system. The second is that a flexibly multimodal interface allows for too much variability in how modalities can be used together. The third is that there were many different ways in which users could answer a single question - for example the date on which a meeting happened could be found using the predefined criteria buttons, rearranging books in the bookcase, or opening the meeting book. This, when combined with all of the modalities that it was possible to use at each step in the interaction, led to a combinatorial explosion. However, Sturm et al. [123] found that there were clear differences in interaction patterns adopted by different users, so it would be interesting to see whether the same holds for interaction with the Archivus system, and how those patterns might influence modality choice. However, experiments of this kind would have to be done on only a limited set of questions and with single modalities or modality pairs, in order to make analysis manageable.

Task type and modality choice correlations

The way in which the experiment presented in this thesis was conceived did not allow for a sufficiently detailed evaluation of whether there were correlations between the choice of a modality or modality combination to perform a task, and the type of task. It would be worthwhile to delve deeper into this issue in future studies, as we suspect that correlations are present. In addition to the types of analysis described in section 6.9, we would also follow work by Tricot [107], which suggests looking at whether there are differences associated with finding information that is explicit in a text or information that has to be inferred. The

question set used in the Archivus experiments was not designed to distinguish between these two cases, so a vast majority of the questions involved finding information that was explicit either in the data or in the interface elements themselves.

Browsing vs. searching tasks

The Archivus system allows for either searching on the multimedia data in the database or simply browsing it, and it would be interesting to determine whether interaction using different modalities changes depending on which of these activities the user is doing. However, the questions that were used in the experiments described in this thesis were not targeted to specifically elicit either searching or browsing behaviour. Although finding the answers to most questions required both searching and browsing, some questions could be answered using only one of these. Moreover, there was a lot of variability in how users mixed searching and browsing to answer specific questions. Consequently, it was impossible to determine whether there were differences in modality use that were correlated to either of these activities in particular. In order to test for such a correlation, we would need to define an experiment in which users were forced to either exclusively search or browse using the different modalities.

Cognitive factors

Neither the competencies of the author nor the availability of resources allowed for an investigation into the cognitive factors that might be affecting modality choice in this work. Le Bigot et al. [51], for example, show that there is a relationship between a modality chosen for an interaction and the cognitive costs implied in planning for interaction and for solving particular tasks. Similarly, Grasso [40] shows that interfaces that require the use of speech impose higher memory requirements than those that do not. It would be interesting to determine the extent to which such factors affect the use of multimedia and multimodal systems such as Archivus, and whether the simultaneous availability of different input modalities reduces them.

Further inspired by the work of Le Bigot et al. [51], who found that '*voice recognition errors resulted in an increase in stress and mental load which, in turn, led to an increase in the number of voice recognition errors*', it would be interesting to investigate whether the same holds true in the context of Archivus, whether it also applies to errors made with other input modalities, and if so, whether it is to the same degree.

Experimental vs. real-use studies

Karlgren [18] notes that there are differences in user behaviour in experimental studies and real-use studies, and that subjects in experimental studies tend to try harder '*both because of curiosity and the novelty of the situation and to perform well in a situation where they are observed*'. Unfortunately, at this point in time doing real-use experiments in the multimedia meeting browsing and retrieval domain is impossible since there are no institutions that record and annotate meetings in the ways necessary on a regular basis. Nevertheless, it would be useful to know if a real-use situation would significantly alter the use of modalities in the meeting browsing and retrieval context.

8.3 Analysis requiring a change in the Archivus system

In this section, we discuss studies that require changes to be made to the current implementation of the Archivus system, such as the addition of a speech recognizer (which was simulated in the experiments described in this thesis), and the addition of dialogue strategies.

Quality of the speech recognizer

Although the wizard simulation of the speech recognition modules planned for the Archivus system was intentionally not perfect, the simulated recognition rates were much higher than what can reasonably be expected from current speech recognition technology. For the purpose of this work, where the system is a research prototype and not a commercial product, we were more interested in the academic question of whether the addition of speech, assuming fairly good speech recognition quality and language processing capabilities, was useful. If even good speech recognition was not found to be useful, then the addition of speech recognition with current speech recognizer capabilities would be even less useful. However, given the fact that speech *was* found to be a useful input modality for this application, it now becomes important to investigate how users would react when interacting with a real (as opposed to a simulated) speech recognizer. There are two factors that are likely to influence user behaviour and/or acceptability. The first is how long the system takes to process voice input. If processing time is too long, users are less likely to continue using voice unless they feel that it would give them a significant advantage over other modalities. The second is the recognition rates that would be required to maintain

an interest in using voice if a real speech recognition module was added to the system. For example, Cohen et al. [55] cite work that claims that a recognition rate of 94% is needed if speech use is to remain as useful as pointing devices.

Impact of dialogue strategies on modality choice

The Archivus system was designed as a multimodal *dialogue* system. As was discussed in section 4.4, the current implementation of the system does not include an active dialogue strategy, but rather an architecture that allows for easy integration of a dialogue strategy when one has been developed. Once dialogue strategies are implemented in the system, it would be interesting to examine whether and how modality use changes in the presence of these strategies.

8.4 Concluding remarks

In this thesis we explored a number of research themes addressing the question of whether multimodal interaction brings added value to multimedia meeting browsing and retrieval. This was done through a large-scale user study with the Archivus multimodal system, in whose design and development we participated. The study, for which we had designed and put into practice the experimental protocol, used a multimodality-enabled version of the Wizard of Oz experiment methodology. The data resulting from the study was then analyzed, and a number of conclusions were drawn. We hope that our findings help to further the field of multimodal interface design and open new research directions by identifying a number of points that require further investigation.

Bibliography

1. Moore, D. *The IDIAP Smart Meeting Room*. 2002, IDIAP: Martigny (Switzerland). p. 13.

2. Tucker, S. and S. Whittaker. *Accessing Multimodal Meeting Data: Systems, problems and possibilities*. In *Machine Learning for Multimodal Interaction - First International Workshop, MLMI 2004*. S. Bengio and H. Bourlard, Editors. 2005, Springer Verlag: Martigny, Switzerland. p. 1-11.

3. Sinha, A.K. and J. Landay. *Embarking on Multimodal Interface Design*. In *IEEE International Conference on Multimodal Interfaces*. 2002. Pittsburgh, PA, USA.

4. Coutaz, J., D. Salber, and S. Balbo. *Towards Automatic Evaluation of Multimodal User Interfaces.* Knowledge-Based Systems, 1993. **6**(4): p. 267-274.

5. Bernsen, N.O. *Multimodality in Language and Speech Systems - From theory to design support tool.* In *Multimodality in Language and Speech Systems*. B. Granström, D. House, and I. Karlsson, Editors. 2002, Kluwer Academic Publishers: Dordrecht. p. 93-148.

6. Bilici, V., et al. *Preferred Modalities in Dialogue Systems*. In *International Conference on Spoken Language Processing - ICSLP2000*. 2000. Beijing, China.

7. Walker, M. *Natural Language in a Desktop Environment*. In *3rd International Conference on Human Computer Interaction (HCI'89)*. 1989. Boston, MA, USA.

8. Lisowska, A., M. Rajman, and T.H. Bui. *ARCHIVUS : A System for Accesssing the Content of Recorded Multimodal Meetings*. In *MLMI - Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms*. 2004. Martigny, Switzerland.

9. Dix, A., et al. *Human Computer Interaction*. Second ed. 1998, England: Prentice Hall.

10. Norman, D.A. *Emotional Design: Why we love (or hate) everyday things*. 2004, New York: Basic Books. 257.

11. Lisowska, A. *Multimodal Interface Design for the Multimodal Meeting Domain: Preliminary Indications from a Query Analysis Study*. 2003, University of Geneva: Geneva. p. 30.

12. Jacko, J.A. and A. Sears. *The Human Computer Interaction Handbook*. Human Factors and Ergonomics, ed. G. Salvendy. 2003, Mahwah, New Jersey: Lawerence Erlbaum Associates. 1277.

13. Mayhew, D.J. *Requirements Specification within the Usability Engineering Life Cyle*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 913-921.

14. Jokinen, K. and A. Raike. *Multimodality - Technology, visions and demands for the future*. In *1st Nordic Symposium on Multimodal Interfaces*. 2003. Copenhagen, Denmark.

15. Hinckley, K. *Input Technologies and Techniques*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: 2003.

16. Cohen, P.R. and S. Oviatt. *The Role of Voice Input for Human-machine Communication.* National Academy of Sciences, 1995. **92**(22): p. 9921-9927.

17. Buxton, W. *HCI and the Inadequacies of Direct Manipulation Systems*. In *SIGCHI Bulletin*. 1993. p. 21-11.

18. Karlgren, J. *The Interaction of Discourse Modality and User Expectations in Human-computer Dialog*. In *Computer and Systems Sciences*. 1992, University of Stockholm: Stockholm.

19. Dybkjaer, L. and N.O. Bernsen. *Usability Issues in Spoken Language Dialogue Systems.* Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue System Engineering, 2000. **6**(3/4): p. 243-272.

20. Bradford, J.A. *The Human Factors of Speech-based Interfaces: A research agenda.* SIGCHI Bulletin, 1995. **27**(2): p. 61-67.

21. Grasso, M.A., D.S. Ebert, and T.W. Finin. *The Integrality of Speech in Multimodal Interfaces.* ACM Transactions on Computer-Human Interaction, 1998. **5**(4): p. 303-325.

22. Dahlbäck, N. and A. Jönsson. *Empirical Studies of Discourse Representations for Natural Language Interfaces*. In *4th Conference of the European Chapter of the ACL (EACL'89)*. 1989. Manchester, U.K.

23. McTear, M. *Spoken Language Technology: Enabling the conversational interface.* ACM Computing Surveys, 2002. **34**(1): p. 90-169.

24. Anastopoulou, S., C. Baber, and M. Sharples. *Multimedia and multimodal systems: Commonalities and differences*. 2001, Educational Technology Group, School of Engineering, University of Birmingham,: Birmingham.

25. Bretan, I. and J. Karlgren. *Transparent Natural Language Interaction through Multimodality*. In *ERCIM Workshop on Multimodal HCI*. 1993. Nancy, France.

26. Lefebvre, P., G. Duncan, and F. Poirier. *Speaking with Computers: A multimodal approach*. In *EUROSPEECH'93*. 1993. Berlin, Germany.

27. Rudnicky, A. *Mode Preference in a Simple Data-retrieval Task*. In *APRP Workshop in Human Language Technology*. 1993. San Mateo: Morgan Kaufmann.

28. Halverson, C.A., et al. *The Beauty of Errors: Patterns of error correction in desktop speech systems*. In *INTERACT '99*. 1999. Edinburgh, Scotland: IOS Press.

29. Lai, J. and N. Yankelovich. *Conversational Speech Interfaces*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 698-713.

30. Oviatt, S. *Breaking the Robustness Barrier: Recent progress on the design of robust multimodal systems.* Advances in Computers, 2002. **56**: p. 305-341.

31. Karat, C.-M., J. Vergo, and D. Nahamoo. *Conversational Interface Technologies.* In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 169-186.

32. Ogden, W.C. and P. Bernick. *Using Natural Language Interfaces*. In *Handbook of Human Computer Interaction*, M. Helander, Editor. 1996, Elsevier Science Publishers B.V.

33. Walker, M. and S. Whittaker. *When Natural Language is Better than Menus: a field study*. 1989, Hewlett Packard Laboratories: Bristol, England. p. 1-9.

34. Dahlbäck, N., A. Jönsson, and L. Ahrenberg. *Wizard of Oz Studies - Why and how*. In *International Workshop on Intelligent User Interfaces*. 1993. Orlando, FL, USA: ACM Press, New York.

35. Pirker, H., G. Loderer, and H. Trost. *Thus Spoke the User to the Wizard*. In *6th European Conference on Speech Communication and Technology (Eurospeech'99)*. 1999. Budapest, Hungary.

36. Bell, L., et al. *Modality Convergence in a Multimodal Dialogue System*. In *Götalog Fourth Workshop on the Semantics and Pragmatics of Dialogue*. 2000.

37. Oviatt, S., P. Cohen, and M. Wang. *Toward Interface Design for Human-language Technology: Modality and structure as determinants of linguistic complexity.* Speech Communication, 1994. **15**(3-4): p. 283-300.

38. Oviatt, S. *User-centered Modeling for Spoken Language and Multimodal Interfaces.* IEEE Multimedia, 1996. **3**(4): p. 26-35.

39. Dybkjaer, H., N.O. Bernsen, and L. Dybkjaer. *Wizard-of-Oz and the Trade-off Between Naturalness and Recogniser Constraints*. In *Eurospeech 2003*. 1993. Berlin, Germany.

40. Grasso, M.A. *Speech Input in Multimodal Environments: A proposal to study the effect of reference visibility, reference number, and task integration*. 1996, University of Maryland, Baltimore Campus: Baltimore, Maryland. p. 51.

41. Dybkjaer, L., N.O. Bernsen, and W. Minker. *Usability Evaluation of Multimodal and Domain-Oriented Spoken Language Dialogue Systems*. In *4th International Conference on Language Resources and Evaluation*. 2004. Lisbon, Portugal.

42. Oviatt, S., et al. *Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-art systems and future research directions.* Human Computer Interaction, 2000. **15**(4): p. 263-322.

43. Allen, J., et al. *Towards Conversational Human-Computer Interaction.* AI Magazine, 2001. **22**(4): p. 27-37.

44. Bernsen, N.O. and L. Dybkjaer. *Evaluation of Spoken Multimodal Conversation*. In *ICMI-04*. 2004. State College, Pennsylvania, USA: ACM.

45. Sturm, J., F. Wang, and B. Cranen. *Adding Extra Input/output Modalities to a Spoken Dialogue System*. In *2nd ACL SIGdial Workshop on Discourse and Dialogue*. 2001. Aalborg, Denmark.

46. Jönsson, A. *A Dialogue Manager for Natural Language Interfaces*. In *2nd Conference of the Pacific Association for Computational Linguistics*. 1995. University of Queensland, Brisbane, Australia.

47. Jönsson, A. *Dialogue Actions for Natural Language Interfaces*. in *Fourth International Joint Conference on Artificial Intelligence (IJCAI-95)*. 1995. Montreal, Canada.

48. Androutsopoulos, I., G.D. Ritchie, and P. Thanisch. *Natural Language Interfaces to Databases - An Introduction.* Journal of Natural Language Engineering, 1994. **1**(1): p. 29-85.

49. Nerbonne, J. *Natural Language Interfaces and the Turing Test*. In *Natural Language Interfaces: From Laboratory to Commercial and User Environments*, M. Franciska, G. de Jong, and A. Nijholt, Editors. 1993: Twente. p. 15-21.

50. Moore, R. and A. Morris. *Experiences Collecting Genuine Spoken Enquiries using WOZ Techniques*. In *Workshop on Speech and Natural Language (HLT'91)*. 1992. Harriman, New York: Association for Computational Linguistics.

51. Le Bigot, L., et al. *Mode and Modal Transfer Effects on Performance and Discourse Organization with an Information Retrieval Dialogue System in Natural Language.* Computers in Human Behaviour, 2006. **22**(2): p. 467-500.

52. Cohen, P., D. McGee, and J. Clow. *The Efficiency of Multimodal Interaction for a Map-based Task*. In *Applied Natural Language Processing Conference (ANLP'00)*. 2000. Seattle, Washington: Morgan Kaufmann.

53. Le Bigot, L., E. Jamet, and J.-F. Rouet. *Searching Information with a Natural Language Dialogue System: A comparison of spoken vs. written modalities.* Applied Ergonomics, 2004. **35**(6): p. 557-564.

54. Christian, K., et al. *A Comparison of Voice Controlled and Mouse Controlled Web Browsing*. In *ACM ASSETS 2000*. 2000. Arlington, Virginia: ACM Press.

55. Cohen, P., et al. *The Efficiency of Multimodal Interaction: A case study*. In *International Conference on Spoken Language*. 1998.

56. Oviatt, S. *Ten Myths of Multimodal Interaction.* Communications of the ACM, 1999. **42**(11): p. 74-81.

57. Maybury, M. and J.R. Lee. *Multimedia & Multimodal Interaction Structure*. In *The Structure of Multimodal Dialogue II*, M.M. Taylor, F. Neel, and D.G. Bouwhuis, Editors. 2000, John Benjamins: Amsterdam. p. 295-308.

58. Sutcliffe, A. *Multimedia User Interface Design*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 245-262.

59. Coutaz, J. and J. Caelen. *A Taxonomy for Multimedia and Multimodal User Interfaces*. In *1st ERCIM Workshop on multimodal human-computer interaction*. 1991. Lisbon, Portugal.

60. Salber, D. and J. Coutaz. *Requirements for Multimodal Wizard of Oz Platforms*.

61. Salber, D. and J. Coutaz. *Applying the Wizard of Oz Technique to the Study of Multimodal Systems*. In *3rd International East/West Human Computer Interaction Conference*. 1993. Moscow, Russia: Springer Verlag Publications.

62. Oviatt, S. *Multimodal Interfaces*. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Assoc.: Mahwah, NJ, USA. p. 286-304.

63. Reeves, L.M., et al. *Guidelines for Multimodal Interface Design.* Communications of the ACM, 2004. **47**(1): p. 57-59.

64. Oviatt, S. *Taming Recognition Errors with a Multimodal Interface.* Communications of the ACM, 2000. **43**(9): p. 45-51.

65. Nass, C. and L. Gong. *Speech Interfaces from an Evolutionary Perspective.* Communications of the ACM, 2000. **43**(9): p. 36-79.

66. Oviatt, S. *Multimodal Interface Research: A science without borders*. In *International Conference on Spoken Language Processing (ICSLP'2000)*. 2000. Beijing, China: Chinese Friendship Publishers.

67. Oviatt, S. and P. Cohen, *Multimodal Interfaces that Process What Comes Naturally.* Communications of the ACM, 2000. **43**(3): p. 45-53.

68. Oviatt, S. *Designing Robust Multimodal Systems for Universal Access*. In *Workshop on Universal Accessibility of Ubiquitous Computing (WUAUC'01)*. 2001. Alcacer do Sal, Portugal: ACM Press.

69. Mignot, C., C. Valot, and N. Carbonell. *An Experimental Study of Future 'Natural' Multimodal Human-computer Interaction*. In *ACM INTERANCT'93 and CHI '93*. 1993. Amsterdam, Netherlands: ACM Press.

70. Karl, L., M. Pettey, and B. Schneiderman, *Speech Activated versus Mouse-Activated Commands for Word Processing Applications: An empirical evaluation.* International Journal of Man-Machine Studies, 1993. **39**(4): p. 667-687.

71. Proctor, R.W. and K.-P. Vu, L. *Human Information Processing: An overview for human-computer interaction*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 35-51.

72. Whittaker, S. and M. Walker. *Toward a Theory of Multimodal Interaction*. In *AAAI'91 Workshop Notes - InterCHI Adjunct Proceedings*. 1991. Amsterdam, Netherlands.

73. Oviatt, S., et al. *Toward a Theory of Organized Multimodal Integration Patterns during Human-Computer Interaction*. In *International Conference on Multimodal Interfaces (ICMI'03)*. 2003. Vancouver, British Columbia, Canada: ACM Press.

74. Coutaz, J., L. Nigay, and D. Salber. *Multimodality from the User and System Perspectives*. In *ERCIM'95 workshop on Multimedia Multimodal User Interfaces*. 1995. Crete, Greece.

75. Wahlster, W. *SmartKom: Fusion and Fission of Speech, Gestures and Facial Expressions*. In *1st International Workshop on Man-Machine Symbiotic Systems*. 2002. Kyoto, Japan.

76. Stein, A. and U. Thiel. *A Conversational Model of Multimodal Interaction*. In *AAAI93, Eleventh conference on Artificial Intelligence*. 1993. Washington, D.C., USA: AAAI Press/ The MIT Press.

77. Elting, C., et al. *The Use of Multimodality within the EMBASSI System*. In *Proceedings of M&C2002, Usability Engineering Multimodaler Interaktionsformen*. 2002. Hamburg, Germany.

78. Bohus, D. and A. Rudnicky. *LARRI: a language-based maintenance and repair assistant*. In *IDS-2002*. 2002. Kloster Irsee, Germany.

79. Gourdol, A., et al. *Two Case Studies of Software Architecture for Multimodal Interactive Systems: Voicepaint and a voice-enabled graphical notebook*. In *IFIP 92*. 1992: North-Holland Publishing Co.

80. Waibel, A., et al. *Multimodal Interfaces.* Artificial Intelligence Review, 1995.

81. Dybkjaer, L., N.O. Bernsen, and H. Dybkjaer. *Knowledge Acquisition for a Constained Speech System using WoZ*. In *6th Conference of the European Association for Computational Linguistics (EACL)*. 1993. Utrecht, Netherlands.

82. Bernsen, N.O. *What is Natural Interactivity?* In *Workshop: From Spoken Dialogue to Full Natural Interactive Dialogue: Theory, Empirical Analysis and Evaluation at LREC'2000*. 2000. Athens, Greece: European Language Resources Association.

83. Maybury, M. *Communicative Acts for Multimedia and Multimodal Dialogue*. In *The Structure of Multimodal Dialogues II*, M.M. Taylor, F. Neel, and D.G. Bouwhuis, Editors. 2000, John Benjamins Publishing: Amsterdam. p. 375-392.

84. Bernsen, N.O. *Modality Theory in Support of Multimodal Interface Design*. In *AAAI Spring Symposium on Intelligent Multi-media Multi-modal Systems*. 1994. Stanford, USA.

85. Bernsen, N.O. *Foundations of Multimodal Representations.* Interacting with Computers, 1994. **6**(4): p. 347-371.

86. Bernsen, N.O. and S. Luz. *SMALTO: Advising interface designers on the user of speech in multimodal systems*. In *Third IEEE workshop on Multimedia Signal Processing*. 1999. Elsinore, Denmark: IEEE.

87. Bernsen, N.O. and L. Dybkjaer. *A Theory of Speech in Multimodal Systems*. In *ESCA Workshop on Interactive Dialogue in Multimodal Systems*. 1999. Irsee, Germany: European Speech Communication Association.

88. Bernsen, N.O. and L. Dybkjaer. *Is Speech the Right Thing for your Application?* In *International Conference for Spoken Language Processing (ICSLP'98)*. 1998. Sydney, Australia: Australian Speech Science and Technology Association.

89. Lee, J.R. and M. Maybury. *Multimedia and Multimodal Interaction Structure*. In *The Structure of Multimodal Dialogues II*, D.G. Bouwhuis, Editor. 2000, John Benjamins Publishing: Amsterdam. p. 295-307.

90. van der Veer, G.C. and M. del Carmen Puerta Melguizo. *Mental Models*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 52-80.

91. Brave, S. and C. Nass. *Emotion in Human-Computer Interaction*. In *Handbook of Human Computer Interaction*, J. Jacko and A. Sears, Editors. 2002, Lawrence Erlbaum Associates: New York. p. 251-271.

92. Nass, C. and Y. Moon. *Machines and Mindlessness: Social responses to computers.* Journal of Social Issues, 2000. **56**(1): p. 81-103.

93. Lalanne, D., et al. *The IM2 Multimodal Meeting Browser Family*. 2005: Switzerland. p. 17.

94. Bett, M., et al. *Multimodal Meeting Tracker*. In *RIAO*. 2000. Paris, France.

95. Larsen, L.B. *Usability Evaluation of Spoken Dialogue Systems*. In *4th International Conference on Language Resources and Evaluation, LREC'2004*. 2004. Lisbon Portugal.

96. Raskin, J. *The Humane Interface: New directions for designing interactive systems*. 2000: Addison-Wesley, ACM Press. 256.

97. Redish, J. and D. Wixon. *Task Analysis*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 922-940.

98. Sire, S. and D. Lalanne. *Smart Meeting Minutes Application Specification*. 2002, University of Fribourg: Fribourg (Switzerland).

99. Holzblatt, K. *Contextual Design*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 941-963.

100. Norman, D.A. *The Design of Everyday Things*. 2002, New York: Basic Books. 257.

101. Beaudouin-Lafon, M. and W. Mackay. *Prototyping Tools and Techniques*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 1006-1031.

102. Tricot, A. *Recherche d'Information dans des Documents Non-linéaires et Récupération Volontaire en Mémoire*. In *1° Colloque « Jeunes Chercheurs en Sciences Cognitives »*. 1994. La Motte d'Aveillans.

103. Scanlon, J. and L. Percival. *UCD for Different Project Types, part 1*. 2002, IBM developerWorks.

104. Bui, T.H. and M. Rajman. *Rapid Dialogue Prototyping Methodology*. 2004, Swiss Federal Institute of Technology (EPFL): Lausanne (Switzerland).

105. Haddock, N.J. *Multimodal Database Query*. In *COLING '92*. 1992. Nantes, France.

106. Melichar, M., et al. *Rapid Mutlimodal Dialogue Design: Application in a Multimodal Meeting Retrieval and Browsing System*. In *MLMI-05*. 2005. Edinburgh, Scotland.

107. Tricot, A. and J.-F. Rouet. *Activites de Navigation dans les Systemes d'Information*. In *Psychologie Ergonomique: tendences actuelles*, J.-M. Hoc and F. Darses, Editors. 2004, PUF: Paris.

108. Constantine, L. *Use and Misuse of Metaphor*. 1998, Constantine & Lockwood, Ltd.

109. Watzman, S. *Visual Design Principle for Usable Interfaces*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 263-285.

110. Card, S.K., G.G. Robertson, and W. York. *The WebBook and the Web Forager: An information workspace for the world-wide web*. In *Conference on Human Factors and Computing Systems (CHI '96)*. 1996. Vancouver, British Columbia, Canada: ACM Press.

111. Ozsoyoglu, G., et al. *Electronic Books in Digital Libraries*. In *IEEE Advances in Digital Libraries Conference*. 2000. Washington, D.C.

112. Brewster, S. *Nonspeech Auditory Output*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 220-240.

113. Dahlbäck, N., et al. *Spoken Interaction with Computers in a Native or Non-Native Language - Same or different?* In *Interact 2001*. 2001. Tokyo, Japan.

114. Flanagan, J. and I. Marsic. *Issues in Measuring the Benefits of Multimodal Interfaces*. In *IEEE Conference on Acoustics, Speech and Signal Process (ICASSP '97)*. 1997. Munich, Germany: IEEE Computer Society.

115. Dumas, J.S. *User-based Evaluations*. In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 1093-1117.

116. Klemmer, S.R., et al. *Suede: A Wizard of Oz prototyping tool for speech user interfaces*. In *ACM Symposium: User Interface Software and Technology (UIST '00)*. 2000. San Diego, California, USA: ACM Press.

117. Whittaker, S. and P. Stenton. *User Studies and the Design of Natural Language Systems*. In *4th Annual meeting of the European Chapter of the Association for Computational Linguistics*. 1989. Manchester, England.

118. Qvarfordt, P., A. Jonsson, and N. Dahlbäck. *The Role of Spoken Feedback in Experiencing Multimodal Interfaces as Human-like*. In *ICMI'03*. 2003. Vancouver, British Columbia, Canada: ACM.

119. Price, D.E., et al. *Off to See the Wizard: using a "Wizard of Oz" study to learn how to design a spoken language interface for programming*. In *32nd ASEE/IEEE Frontiers in Education Conference*. 2002. Boston, MA, USA: IEEE.

120. Rajman, M., et al. *Extending the Wizard of Oz Technique for Multimodal Language-enabled Systems.* In *LREC 2006*. 2006. Genoa, Italy.

121. Cheng, H., et al. *A Wizard of Oz Framework for Collecting Spoken Human-computer Dialogs*. In *8th International Conference on Spoken Language Processing (Interspeech-ICSLP'04)*. 2004: Jeju island, Korea.

122. Oviatt, S., A. de Angeli, and K. Kuhn. *Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction*. In *Conference on Human Factors in Computing Systems (CHI'97)*. 1997. Atlanta, Georgia, USA: ACM Press.

123. Sturm, J., et al. *The Effect of Prolonged Use on Multimodal Interaction*. In *ISCA Workshop on Multimodal Spoken Dialogue in Mobile Environments*. 2002. Kloster Irsee, Germany.

124. Mehlenbacher, B. *Documentation - not yet implemented, but coming soon!* In *The Human Computer Interaction Handbook*, J. Jacko and A. Sears, Editors. 2003, Lawrence Erlbaum Associates: Mahwah, New Jersey. p. 527-543.

125. Erbach, G. *Evaluating Human Question Answering Performance under Time Constraints*. 2004.

126. Oviatt, S. and E. Olsen. *Integration Themes in Multimodal Human-Computer Interaction*. In *International Conference on Spoken Language Processing*. 1994. Yokahama, Japan.

Appendices

Appendix A: User requirements questionnaire

Questionnaire: Eliciting input for a multimodal database query system

Introduction

In the context of the IM2 (Interactive Multimodal Information Management, www.im2.ch) project we are developing a system that stores electronic recordings of meetings between human beings, in a variety of formats - video, audio, text transcripts etc. In addition to the content of the meeting itself, any files that are associated with the meeting, such as presentation slides and distributed papers are also stored in electronic form. The fact that the information is stored in a variety of annotated formats means that the user can ask questions about the actual content of the meetings, in addition to requesting to see or hear parts of the meeting.

A user is able to access the information in the system by simply posing a question to the system (much as they might if they were trying to get the same information from a colleague) about the information in those meetings, or requesting to see some or all of a particular meeting. The user can ask the question through any combination of typing on a keyboard, speaking and using a pointing device (such as a mouse or laser pen). What we are currently interested in is what aspects of a meeting people would want to know about and how they would pose their questions.

Scenarios

Below you will find four scenarios in which someone might want to use the system that has been described. Please pick one, and then list as many questions as you can think of that you would want to ask the system to get the answers you need, given the context described in the scenario. Note: If you are willing to do this for more than one scenario, that would be very helpful, but we ask that the questions be listed separately for each scenario.

1. Imagine that you are managing a project, but are too busy to attend all of the meetings related to it. This isn't a problem, because you know that all meetings in your institute have been recorded and stored in the IM2 system. You want to find out how particular members of the group are contributing overall to the project.

2. Imagine that you are managing a project, but are too busy to attend all of the meetings related to it. This isn't a problem, because you know that all meetings in your institute have been recorded and stored in the IM2 system. You want to find out how the flow of ideas for the project has been progressing, what directions the project is taking and what decisions are being made.

3. Imagine that you have missed a meeting about a project that you are working on. You want to catch up on what happened in that meeting (what was discussed, what was decided etc), and you know that all meetings in your company are recorded and stored in the IM2 system. Also remember that since all of the meetings are stored, you can also ask questions about previous meetings that you attended, if you feel that that can be helpful.

4. Imagine that you have just been hired at a company to work on a project. The project actually started six months ago, so you have some catching up to do. Fortunately, all of the meetings regarding this project have been recorded and are stored in the IM2 system which you will use to help you catch up.

About you

It would be helpful if you could answer the following questions, but you are not obligated to do so.

1. Are you involved in any way with the IM2 (Interactive Multimodal Information Management) project? If yes, please specify how.

2. What sort of computer experience do you have? Please erase all those options that do not apply to you, leaving only those that best describe your experience.

    Basic personal computer use (word processing, internet browsing etc.)
    Extensive personal computing use
    Computer programming experience
    A computer science (or related) degree
    Experience with building databases
    Experience with natural language processing

3. What is your professional position? Again, please erase those that do not apply to you.

    Researcher
    Manager/director
    Other (please specify)

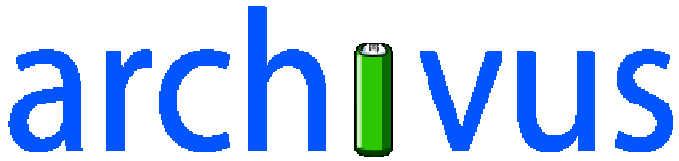
4. What is your area of expertise?

5. Please specify your native language:

6. Would you be willing to volunteer to participate in a hands-on experiment in this area in the future?

Thank you for taking the time to help us with our research. We greatly appreciate it.

Appendix B: Archivus experiment consent form



Experiment Consent Form

Within the Interactive Multimodal Information Management (IM2) project we have developed an interface, Archivus, which allows users to access a database of stored meetings. Through this interface, a user should be able to review meetings or relevant sections of meetings in a quick and efficient manner. The purpose of the experiment in which you will be participating is to determine whether the Archivus interface meets the needs of real users such as yourselves.

By signing this consent form, you agree to be recorded (both audio and video) while interacting with the Archivus system, and allow those recordings to be reviewed and analyzed by the experimenters for research purposes.

We, the experimenters, promise not to make the recorded data publicly available, reveal any personal information you provide (expect as part of general demographic notes), nor publish your image and/or voice without your prior consent.

Name: _____

Date: _____

Signature: _____

Participant number: _____

You will only be asked to sign this form once, even if you agree to participate in multiple experiments.

Appendix C: Archivus evaluation description

Evaluation Scenario

To help us evaluate the system we would like you to pretend that you have just been given a job at SomeCompany Inc. Your new manager is very busy though, so he hasn't had much time to fill you in on the details of your new job. However, he has asked you to find and check some information for him. Sometimes this involves checking whether he has remembered certain facts from a meeting correctly, and at other times it means finding bits of information that he has forgotten. To help you find the information, your manager has told you to use the Archivus system. SomeCompany Inc records all of its meetings in a SmartMeeting room. In this room, meetings are recorded by video cameras and microphones. All of the slides used in the meetings, all documents that anyone brought into the meeting on paper, and any handwritten notes taken by the meeting participants are also stored in electronic form. Finally, all of the meetings have been transcribed, so there is a text version of the meeting available as well. The Archivus system lets you access all of this data. Your manager has also given you the *Archivus Tutorial*, which takes you through some step-by-step examples of how to use the system, as well as the *Archivus Manual*, which explains the system in more detail. You can keep both of these documents with you while you are working with the system.

The questions that your manager wants you to answer are written on cards which the experimenter will give you when they come back to the room.

It is important to remember that we are testing how well the Archivus system helps you to find the information and NOT your ability to answer the questions correctly. We ask that you try to find the correct answer as quickly as you can, but you should not rush to finish, or try to guess the answers.

Now, you should go ahead and work through the Tutorial. If you have any other questions, please ask the experimenter when they come back.

Appendix D: Archivus tutorial examples – P, V and MVK conditions

Appendix D1: The P condition

Tutorial 1

You want to find out what questions Denis asked during the April 21st meeting.

1. Select the *Date of Meeting* predefined criteria button  near the bottom of the screen. You will see a list of all of the meetings that are available, and the dates on which they took place. Select the line that says 'April 21st'. You will notice that three things have happened. The first is that the book for the Furniture 2 meeting has **opened automatically**. This is because it is the **only meeting that occurred on April 21st**. The second is that the colours of the books in the bookcase have changed. The Furniture 2 meeting book is **light green**, which means that it **matches your criteria**, while all of the others which don't match are dark green. The third thing is that **'April 21st'** has been **added as a criterion to the Current Search Criteria list** in the bottom left part of the screen.

2. Now that you have the correct meeting, you still need to find the questions that Denis asked. Select the *Dialogue Elements* predefined criteria button  . You will see two options, 

and 

. Select *Smaller sections of the meeting*. You will see a list of choices in the interactive pane. Select 'Question'. The criteria 'question' will be added to the Current Search Criteria list and the book will open again. This time, you will see **yellow results tabs** 

which show all of the places in the meeting **where questions were asked.** You can see the **actual questions highlighted in yellow** in the book.

3. You may have noticed that Archivus is speaking to you and trying to give you advice. If you do not want to hear what Archivus says, select the loudspeaker icon  **next to the text version of the advice**. You will see a red 'X' appear over the icon  . To turn the sound back on, select that icon again.

4. To see only the questions that Denis asked, select the *Speaker*  predefined criteria button. You will see the names of the speakers from the **active** meeting appear. Select 'Denis'. If you look at the Current Search Criteria list, you will see that 'Denis' has been added to the list, and the book changes to show **only the places** in the meeting **in which Denis asked questions**.

5. On the right side of book, you can also see **several green/blue tabs which are the general tabs**. Select the tab labeled *Documents* 

. You will see a list of all of the documents that were used or referred to in that meeting, including one called *Denis's furniture choices.* Select it. You can now see the document itself, which contains the slides that Denis presented during that meeting. You can **browse through the document**

using the blue up and down arrows to the right of the document. To close the document select the Close button on the bottom right. Although an example is not shown in this tutorial, sometimes you can find references to documents in the text of the book as well. They will look like a citation (for example [3]). Selecting a citation will open the document directly.

6. **Before you go on to the next example** you should reset the system by selecting the ***Task Finished*** button Task finished.

Tutorial 2

Imagine that you want to find out why a coffee machine was mentioned during the meetings.

1. Select the *Content* predefined criteria button  . You will see three categories appear, Topic, Keyword, and Document. Select *Topic*. You will see a list of all of the topics in the meetings. Select the letter 'C' that appears in the alphabet at the top of the screen. You will be taken to the section of the list that starts with 'C'. Select the blue arrow that is on the bottom right side of the screen two time. You will see the list change. Now, towards the top of the list you will see 'coffee machine'. Select it. You can see that three things have happened. The first is that the books on the bookcase have changed. Only two books are now **light green**, which means that **only those books contain information about coffee machines**. Also, you will notice that 'coffee machine' has been added to the Current Search Criteria list. Finally, you will notice that the list has disappeared and been replaced with the options for the *Content* criteria button. You might also notice that these options have now changed to reflect your new criteria. Since you won't need that anymore, you can close it by selecting the Close button on the bottom right.

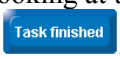
2. To see the context in which the coffee machine was mentioned in each of these meetings you have to look inside the meeting book. Select the 'Furniture 3' book in the bookcase. Archivus will open the meeting book for you, and you will see six **yellow results tabs** 22 that show you the **pages in the book on which the coffee machine is mentioned**. You can **scroll** through the different parts of the meeting where the coffee machine was discussed by **moving between the tabs using the blue up and down arrows** above and below the tabs. The tab that corresponds to the page that you are currently on is marked in dark yellow.

3. Now, imagine that you wanted to know whether they talked about a Nespresso machine. Select the *Content* predefined criteria button, and then Keyword. Scroll through the list of keywords until you find Nespresso, and then select it. You will notice that 'Nespresso' has been added as a keyword in the Current Search Criteria list. Select meeting book 'Furniture 4' in the bookcase. You will see that the **keyword 'Nespresso' is marked in orange in the book**, and the **sentence during which it was used (and is within the topic 'coffee machine') is marked in yellow**. To undo this change, select 'Nespresso' in the Current Search Criteria list. You will see that a Delete button becomes active. Select the *Delete* button. You will see that 'Nespresso' is no longer one of your search criteria and the **whole section of the meeting which talks about coffee**

machines has become highlighted, not just the part that talks about coffee machines and Nespresso.

4. To get a better idea of why the coffee machine was mentioned you can **read through the pages of the book**. To move between the pages **select the little next-**  **or previous-page**  **icons** at the corner of each page. You can also **watch the video** from that part of the meeting. To watch the video, select the little camera icon  below the book. A media player will appear with controls like those on your video machine. Close the media player by selecting the stop button  .

5. To see what other things were **talked about in the same meeting**, select the **Table of Contents** (ToC) general tab  on the right side of the book. You can turn to the part of the meeting mentioned in the Table of Contents by selecting the corresponding number.

6. Once you have finished looking at the information, don't forget to reset the system by selecting ***Task Finished***  .

Tutorial 3

You can also get a quick overview of different information from the meetings by rearranging the books in the bookcase. For example, **you want to know which meetings took place in April.**

1. Select the *Change horizontal label* icon  , located just below the bookcase. You will see a list of possible options for bookshelf labels appear in the interactive pane. Select *Month of Meeting*. You will see that the books on the bookshelf have been sorted by month. There are 5 meetings (books) that took place in April.

2. Now, if you also want to see whether all of those meetings from April took place in the same year, you can add another label to the bookcase to further sort the books. To do this, select the *Change vertical label* icon  . Again, you will see a list of options appear. This time, select *Year of Meeting*. You can see that the books in the bookcase have now been sorted by year (on the leg of the bookcase), and by month (on the bookshelf).

You can change the labels on the bookcase at any point while you are interacting with the system, even after you have already specified other search criteria. However, it is important to remember that **specifying labels on the bookcase does not add those labels as search criteria.** As you can see from the example in this tutorial, the Current Search Criteria list has remained empty.

3. Once you have finished looking at the information, don't forget to reset the system by selecting ***Task Finished***  .

You've now finished the tutorial. If the experimenter has not come back yet, you can try to use Archivus yourself a little bit.

Appendix D2: The V condition

Tutorial 1

You want to find out what questions Denis asked during the April 21st meeting.

1. You can start by saying 'Which meetings happened on April 21st?" You will notice that three things have happened. The first is that the book for the Furniture 2 meeting has **opened automatically**. This is because it is the **only meeting that occurred on April 21st**. The second is that the colours of the books in the bookcase have changed. The Furniture 2 meeting book is **light green**, which means that it **matches your search criteria**, while all of the others which don't match are dark green. The third thing is that '**April 21st'** has been **added as a criteria to the Current Search Criteria list** on the bottom left of the screen.

2. Now that you have the correct meeting, you still need to find the questions that Denis asked. You can do this by saying 'Show me all of the questions that were asked?" The criteria 'question' will be added to the Current Search Criteria list and the book will open again. This time, you will see **yellow results tabs** 

which show all of the places **where questions were asked** in the meeting. You can see the **actual questions highlighted in yellow** in the book.

3. To see only the questions that Denis asked, tell the system that this is what you want by saying 'Show me only questions by Denis'. If you look at the Current Search Criteria list, you will see that 'Denis' has been added to the list, and the book changes to show **only the places** in the meeting **in which Denis asked questions.**

4. On the right side of book, you can also see **several green/blue tabs which are the general tabs**. Access the *Documents* tab 

by saying 'Documents'**.** You will see a list of all of the documents that were used or referred to in that meeting, including one called *Denis's furniture choices.* Select it either by saying the title, or the corresponding number. You can now see the document itself, which contains the slides that Denis presented during that meeting. You can **browse through the document by telling Archivus that you want to go up or down**, just as if you were scrolling. To close the document, ask Archivus to close it. Although an example is not shown in this tutorial, sometimes you can find references to documents in the text of the book as well. They will look like a citation (for example [3]). To access the document ask Archivus to open the document with that number, for example 'Open document 3'.

5. **Before you go on to the next example** you should reset the system by saying '***task finished'***.

Tutorial 2

Imagine that you want to find out why a coffee machine was mentioned during the meetings. You could simply ask the system 'Why was the coffee machine mentioned in the meeting?'. This is similar to how you looked for information in the previous example. There is another way of looking for information though, that we will show in this example.

1. Say 'Show me the content criteria'. This will open the *Content* predefined criteria area, which is represented by the *Content* button  near the bottom of the screen. You will see three categories appear,  ,  , and  . Say 'Show me the topics'. You will see a list of all of the topics in the meetings. Say 'Show me the topics that start with the letter C'. You will be taken to the section of the list that starts with 'C'. Ask the system to scroll down until you see 'coffee machine'. To select it, you can either say 'coffee machine' or the corresponding number. You can see that three things have happened. The first is that the books on the bookcase have changed. Only two books are now **light green**, which means that **only those books contain information about coffee machines**. Also, you will notice that 'coffee machine' has been added to the Current Search Criteria list. Finally, you will notice that the list has disappeared and been replaced with the options for the *Content* criteria button. You might also notice that these options have now changed to reflect your new criteria. Since you won't need that anymore, you can close it by saying 'done'.

2. To see the context in which the coffee machine was mentioned in each of these meetings you have to look inside the meeting book. Say 'Open the Furniture 3 book'. Archivus will open the meeting book for you, and you will see six **yellow results tabs**  that show you the **pages in the book on which the coffee machine is mentioned**. You can **scroll** through the different parts of the meeting where the coffee machine was discussed by asking Archivus to move up or down in the tabs. The tab that corresponds to the page that you are currently on is marked in dark yellow.

3. Now, imagine that you wanted to know whether they talked about a Nespresso machine. Access the *Content* predefined criteria like you did before, but this time, select the  option. Scroll through the list of keywords until you find 'Nespresso', and then select it. You will notice that 'Nespresso' has been added as a keyword to the Current Search Criteria list. Ask Archivus to open the Furniture 4 meeting book. You will see that the **keyword 'Nespresso' is marked in orange** in the book, and **the sentence during which it was used (and within the topic 'coffee machine') is marked in yellow**. To undo this change, say 'Delete Nespresso from Criteria Selection list'. You will see that 'Nespresso' is no longer one of your search criteria and **the whole section of the meeting which talks about coffee machines has become highlighted**, not just the one that talks about coffee machines and Nespresso.

4. To get a better idea of why the coffee machine was mentioned you can **read through the pages of the book**. Do this by **asking Archivus to go to the next or previous page**.

   You can also watch the video from that part of the meeting. To **watch the video**  , ask Archivus to play the video from the meeting. A media player will appear with controls like those on your video machine. You can control the player by saying 'play', 'forward', 'pause' etc. Ask Archivus to close the media player.

5. To see what other things were **talked about in the same meeting**, you can look in the **Table of Contents** which is represented by the *ToC* general tab 

on the right side of the book. Ask Archivus to open the Table of Contents. You can turn to the part of the meeting mentioned in the table of contents by saying the title of the section or the corresponding number.

6. Once you have finished looking at the information, don't forget to reset the system by saying ***Task Finished***.

Tutorial 3

You can also get a quick overview of different information from the meetings by rearranging the books in the bookcase. For example, **you want to know which meetings took place in April.**

4. Select the 

icon, located just below the bookcase, by saying '*Change horizontal label'*. You will see a list of possible options for bookshelf labels appear in the interactive pane. Select *Month of Meeting* by saying it or the corresponding number. You will see that the books on the bookshelf have been sorted by month. There are 5 meetings (books) that took place in April.

5. Now, if you also want to see whether all of those meetings from April took place in the same year, you can add another label to the bookcase to further sort the books. To do this, select the 

icon by saying *'Change vertical lablel'*. Again, you will see a list of options appear. Say *Year of Meeting* to select it. You can see that the books in the bookcase have now been sorted by year (on the leg of the bookcase), and by month (on the bookshelf).

You can change the labels on the bookcase at any point while you are interacting with the system, even after you have already specified other search criteria. However, it is important to remember that **specifying labels on the bookcase does not add those labels as search criteria.** As you can see from the example in this tutorial, the Current Search Criteria list has remained empty.

6. Don't forget to reset the system by saying ***Task Finished***.

You've now finished the tutorial. If the experimenter has not come back yet, you can try to use Archivus yourself a little bit.

Appendix D3: The MVK condition

Tutorial 1

You want to find out what questions Denis asked during the April 21st meeting.

1. You can start by saying 'Which meetings happened on April 21st?" You will notice that three things have happened. The first is that the book for the Furniture 2 meeting has **opened automatically.** This is because it is the **only meeting that occurred on April 21st**. The second is that the colours of the books in the bookcase have changed. The Furniture 2 meeting book is **light green,** which means that it **matches your search criteria**, while all of the others which don't match are dark green. The third thing is that '**April 21st**' has been **added as a criteria to the Current Search Criteria list** on the bottom left of the screen.

2. Now that you have the correct meeting, you still need to find the questions that Denis asked. Click on the *Dialogue Elements* predefined criteria button  near the bottom of the screen. You will see two options,  and  . Say 'Smaller sections of the meeting'. You will see a list of choices in the interactive pane. Click on 'Question'. The criteria 'question' will be added to the Current Search Criteria list and the book will open again. This time, you will see **yellow results tabs**  which show all of the places in the meeting **where questions were asked**. You can see the **actual questions highlighted in yellow** in the book.

3. You may have noticed that Archivus is speaking to you and trying to give you advice. If you do not want to hear what Archivus says, click on the loudspeaker icon  next to the text version of the advice. You will see a red 'X' appear over the icon . To turn the sound back on, click on that icon again.

4. To see only the questions that Denis asked, tell the system that this is what you want by typing 'Show me only questions by Denis', and pressing the return key. If you look at the Current Search Criteria list, you will see that 'Denis' has been added to the list, and the book changes to show **only the places** in the meeting **in which Denis asked questions**.

5. On the right side of book, you can also see **several green/blue tabs which are the general tabs**. Click on the tab labeled *Documents*  . You will see a list of all of the documents that were used or referred to in that meeting, including one called *Denis's furniture choices.* Choose it either by saying or typing the title or the corresponding number, or by clicking on it. You can now see the document itself, which contains the slides that Denis presented during that meeting. You can **browse through the document by telling Archivus that you want to go up or down**, or by **clicking on the blue up**  **and down**  **arrows** to the right of the document. To close the document select the  button on the bottom right. Although an example is not shown in this tutorial, sometimes you can find references to documents in the text of the book as well.

They will look like a citation (for example [3]). To access the document ask Archivus to open the document with that number, for example 'Open document 3', or click on the citation.

6. **Before you go on to the next example** you should reset the system by saying '***task finished'***.

Tutorial 2

Imagine that you want to find out why a coffee machine was mentioned during the meetings.

1. Click on the *Content* predefined criteria button  . You will see three categories appear,  ,  , and  . Click on *Topic*. You will see a list of all of the topics in the meetings. Say 'Show me the topics that start with the letter C'. You will be taken to the section of the list that starts with 'C'. Click two times on the blue arrow  that is on the bottom right side of the screen. You will see the list change. Now, towards the top of the list you will see 'coffee machine'. To choose it say 'coffee machine' or the corresponding number. You can see that three things have happened. The first is that the books on the bookcase have changed. Only two books are now **light green,** which means that **only those books contain information about coffee machines**. Also, you will notice that 'coffee machine' has been added to the Current Search Criteria list. Finally, you will notice that the list has disappeared and been replaced with the options for the *Content* criteria button. You might also notice that these options have now changed to reflect your new criteria. Since you won't need that anymore, you can close it by clicking on the  button on the bottom right.

2. To see the context in which the coffee machine was mentioned in each of these meetings you have to look inside the meeting book. Ask Archivus to open the Furniture 3 meeting book by typing 'Open Furniture 3', and pressing the return key. Archivus will open the meeting book for you, and you will see six **yellow results tabs**  that show you the **pages in the book on which the coffee machine is mentioned.** You can **scroll** through the different parts of the meeting where the coffee machine was discussed by asking Archivus to move up or down in the tabs or by **moving between the tabs using the blue up**  **and down**  **arrows** above and below the tabs. The tab that corresponds to the page that you are currently on is marked in dark yellow.

3. Now, imagine that you wanted to know whether they talked about a Nespresso machine. Say 'Show me the content criteria', and then click on  when your choices appear. Scroll through the list of keywords until you find Nespresso, and then click on it. You will notice that 'Nespresso' has been added as a keyword to the Current Search Criteria list. Open the Furniture 4 meeting book by clicking on it in the bookcase. You will see that the **keyword 'Nespresso' is marked in orange** in the book, and **the sentence during which it was used (and within the topic 'coffee machine') is marked in yellow**. To undo this change, click on 'Nespresso' in the Current Search Criteria list. You will see that a *Delete* button  becomes active. Say 'Delete Nespresso'. You will see that 'Nespresso' is no longer one of your search criteria and **the whole section of the meeting which talks about coffee machines has become highlighted**, not just the one that talks about coffee machines and Nespresso.

4. To get a better idea of why the coffee machine was mentioned you can **read through the pages of the book**. Do this by **asking Archivus to go to the next or previous page**.

   You can also watch the video from that part of the meeting. To **watch the video**  , ask Archivus to play the video from the meeting. A media player will appear with controls like those on your video machine. You can control the player by saying 'play', 'forward', 'pause' etc. Close the media player by clicking on the stop button  .

5. To see what other things were **talked about in the same meeting**, you can look in the **Table of Contents** which is represented by the *ToC* general tab 

on the right side of the book. Ask Archivus to open the Table of Contents. You can turn to the part of the meeting mentioned in the Table of Contents by clicking on the corresponding number.

6. Don't forget to reset the system by clicking ***Task Finished*** 

.

Tutorial 3

You can also get a quick overview of different information from the meetings by rearranging the books in the bookcase. For example, **you want to know which meetings took place in April.**

7. Select the 

icon, located just below the bookcase. You will see a list of possible options for bookshelf labels appear in the interactive pane. Select *Month of Meeting* by saying it or the corresponding number. You will see that the books on the bookshelf have been sorted by month. There are 5 meetings (books) that took place in April.

8. Now, if you also want to see whether all of those meetings from April took place in the same year, you can add another label to the bookcase to further sort the books. To do this, select the 

icon by typing *'Change vertical lablel'*. Again, you will see a list of options appear. Select '*Year of Meeting'* by clicking on it. You can see that the books in the bookcase have now been sorted by year (on the leg of the bookcase), and by month (on the bookshelf).

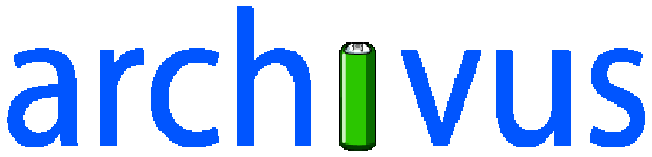
You can change the labels on the bookcase at any point while you are interacting with the system, even after you have already specified other search criteria. However, it is important to remember that **specifying labels on the bookcase does not add those labels as search criteria.** As you can see from the example in this tutorial, the Current Search Criteria list has remained empty.

9. Don't forget to reset the system by typing '*Task Finished'*.

You've now finished the tutorial. If the experimenter has not come back yet, you can try to use Archivus yourself a little bit.

Appendix E: Archivus experiment questionnaires

Appendix E1: Archivus pre-experiment (demographic) questionnaire



Questionnaire 1

Participant # _____

Personal Information:

Gender: □ Male □ Female

Age: □<18 □ 18-24 □ 25-35 □ 36-45 □ 46-55 □ 55+

Occupation: _____

Is English your native language: □ Yes □ No

Are you □ Right-handed □ Left-handed

Familiarity with computers:

Approximately how many hours per day do you use a computer:
□<1 □ 2-4 □ 5-7 □ 7-10 □ 10+

Do you use mostly a: □ a Windows system □ a Unix-based system □ an Apple system

Do you use the following types of software more than once a week?
If yes, please specify which one you use.

Internet browser □ yes □ no Which: _____
Word processor □ yes □ no Which: _____
Audio player □ yes □ no Which: _____
Video player □ yes □ no Which: _____
Instant messenger □ yes □ no Which: _____
Database tools □ yes □ no Which: _____

Have you ever:

Controlled a computer with voice: □ yes □ no
Used speech recognition software: □ yes □ no
Used a touch screen : □ yes □ no

Additional Questions:

Do you use a mobile telephone? □ Yes □ No

If you use a mobile phone, on average how often do you use the text messaging function?
- □ More than once a day
- □ A few times a week
- □ A few times a month
- □ A few times a year
- □ Never

How often do you use a library?
- □ Every day
- □ A few times a week
- □ A few times a month
- □ A few times a year
- □ Never

In general, do you think that finding what you are looking for in a library is:
- □ Very easy
- □ Easy
- □ Alright
- □ Hard
- □ Very hard

How often do you search for information using the internet?
- □ Every day
- □ A few times a week
- □ A few times a month
- □ A few times a year
- □ Never

On average, how often do you attend meetings in your everyday life?
- □ Every day
- □ A few times a week
- □ A few times a month
- □ A few times a year
- □ Never

If you never attend meetings, you do not need to answer the next two questions.

What do you use to help you remember what happened during a meeting you attended? Check all options that apply.
- □ notes you made yourself
- □ the minutes of the meeting
- □ documents referred to during the meeting (agenda, slides, etc)
- □ notes of your colleagues
- □ discussions with colleagues
- □ your memory

What do you use to find out about a meeting you missed? Check all options that apply.

- □ the minutes of the meeting
- □ documents referred to during the meeting (agenda, slides, etc)
- □ notes of your colleagues
- □ discussions with colleagues

Appendix E2: Archivus post-experiment questionnaire

This questionnaire was administered electronically during the final experiments. Spaces between questions have been removed in order to contain space.

Questionnaire 2

Participant # _____ **Experiment # _____**

Thank you for participating in the evaluation of the Archivus system. The last part of the evaluation process is filling in this questionnaire.

Part 1 The system in general

For each of the statements below, please indicate the degree of your agreement with the statement.

	strongly disagree	disagree	neutral	agree	strongly agree
The system was easy to use.	●	●	●	●	●
I was comfortable working with the system.	●	●	●	●	●
I could understand the topic of the meetings.	●	●	●	●	●
It was hard to learn to use the system.	●	●	●	●	●
I felt in control of the system.	●	●	●	●	●
Being able to use language to interact with the system was helpful.	●	●	●	●	●
The system reacted too slowly to my requests.	●	●	●	●	●
It was easy to go back and change a criterion.	●	●	●	●	●
Representing the meetings as pages in a book made them easy to browse.	●	●	●	●	●

	strongly disagree	disagree	neutral	agree	strongly agree
The bookshelf and books helped me understand how to use the system.	●	●	●	●	●
I could use the system how I wanted to.	●	●	●	●	●

Please complete the following sentences by indicating the most appropriate choice.

	not used	not at all useful	not very useful	no opinion	somewhat useful	very useful
The meeting books were …	●	●	●	●	●	●
The bookcase was …	●	●	●	●	●	●
The predefined criteria buttons were …	●	●	●	●	●	●
The current search criteria list was …	●	●	●	●	●	●
The advice given by the system was …	●	●	●	●	●	●
The user input area was …	●	●	●	●	●	●
The help button was …	●	●	●	●	●	●

Would you use the system again in the future? □ yes □ no □ maybe, if it was improved

Please answer the following questions in the space provided:

What did you like most about the system?

What did you like least about the system?

What did you like most about the way you could interact with the system?

What did you like least about the way you could interact with the system?

Part 2 Individual components of the system

Bookcase 

Please rank the modalities listed below in the order in which you thought that they were the most useful for interacting with the bookcase. Give the most useful modality a ranking of 1, the next most useful 2 etc. and 0 if a modality was not useful at all. If you think that two or more modalities were equally useful, you can give them the same ranking.

_____ voice
_____ keyboard
_____ mouse/pen

On a scale of *very hard* to *very easy*, how would you rate the following:

	not used	very hard	hard	alright	easy	very easy
Using the bookcase	●	●	●	●	●	●
Selecting a book from the bookcase	●	●	●	●	●	●
Finding a book on the bookcase	●	●	●	●	●	●
Changing the labels on the bookcase	●	●	●	●	●	●

What did you like about the bookcase?

What would you change about the bookcase?

In what way was the bookshelf the most useful to you?

Did the bookcase react the way that you expected it to? □ yes □ no If no, why:

Meeting Books 

Please rank the modalities listed below in the order in which you thought that they were the most useful for **browsing books**. Give the most useful modality a ranking of 1, the next most useful 2 etc. and 0 if a modality was not useful at all. If you think that two or more modalities were equally useful, you can give them the same ranking.

_____ voice
_____ keyboard
_____ mouse/pen

Please rank the modalities listed below in the order in which you thought that they were the most useful for **finding specific information in a book.** Give the most useful modality a ranking of 1, the next most useful 2 etc. and 0 if a modality was not useful at all. If you think that two or more modalities were equally useful, you can give them the same ranking.

_____ voice
_____ keyboard
_____ mouse/pen

On a scale of *very hard* to *very easy*, how would you rate the following:

	not used	very hard	hard	alright	easy	very easy
Using the meeting books	●	●	●	●	●	●
Browsing a meeting book	●	●	●	●	●	●
Finding items in a meeting book	●	●	●	●	●	●
Accessing video through the book	●	●	●	●	●	●
Accessing audio through the book	●	●	●	●	●	●
Closing a book	●	●	●	●	●	●
Using the general tabs	●	●	●	●	●	●
Using the results tabs	●	●	●	●	●	●

What did you like about the books?

What would you change about the books?

In what way were the meeting books the most useful to you?

Did the meetings books react the way that you expected them to? □ yes □ no If no, why:

Predefined criteria buttons 

Please rank the modalities listed below in the order in which you thought that they were the most useful for **using the predefined criteria buttons.** Give the most useful modality a ranking of 1, the next most useful 2 etc. and 0 if a modality was not useful at all. If you think that two or more modalities were equally useful, you can give them all same ranking.

_____ voice
_____ keyboard
_____ mouse/pen

On a scale of *very hard* to *very easy*, how would you rate the following:

	very hard	hard	alright	easy	very easy
Using the predefined criteria buttons	●	●	●	●	●
Knowing which button to choose	●	●	●	●	●

Please complete the following sentences by indicating the most appropriate choice.

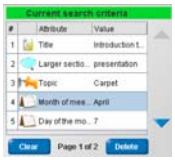
			not used	not at all useful	not very useful	no opinion	somewhat useful	very useful
The		was	●	●	●	●	●	●
The		was	●	●	●	●	●	●
The		was	●	●	●	●	●	●
The		was	●	●	●	●	●	●
The		was	●	●	●	●	●	●

What did you like about the predefined criteria buttons?

What would you change about the predefined criteria buttons?

Did the predefined criteria buttons react the way you expected them to? □ yes □ no If no, why:

Current search criteria list



Please rank the modalities listed below in the order in which you thought that they were the most useful for **accessing the current search criteria list.** Give the most useful modality a ranking of 1, the next most useful 2 etc. and 0 if a modality was not useful at all. If you think that two or more modalities were equally useful, you can give them the same ranking.

_____ voice
_____ keyboard
_____ mouse/pen

On a scale of *very hard* to *very easy*, how would you rate the following:

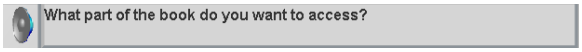
	not used	very hard	hard	alright	easy	very easy
Using the list	●	●	●	●	●	●
Removing an item from the list	●	●	●	●	●	●

What did you like about the current search criteria list?

What would you change about the current search criteria list?

In what way was the current search criteria list the most useful to you?

Did the current search criteria list react the way that you expected it to? □ yes □ no If no, why:

System advice 

Did you find it useful to hear advice from the system? □ yes □ no If no, why:

Was the content of the advice useful? □ yes □ no If no, why:

On a scale of *very hard* to *very easy*, how would you rate the following:

	very hard	hard	alright	easy	very easy
Understanding the system advice was	●	●	●	●	●

The advice was: □ too long □ alright □ too short

The advice was: □ too friendly □ friendly □ average □ formal □ too formal

How much guidance did the system give you?
□ too much □ enough to do my task □ not enough

Did the system advice react the way that you expected it to? □ yes □ no If no, why:

In what way was the system advice the most useful to you?

User input bar 

On a scale of *very hard* to *very easy*, how would you rate the following:

	not used	very hard	hard	alright	easy	very easy
Accessing the user input bar	●	●	●	●	●	●
Editing in the input bar	●	●	●	●	●	●

Help button 

On a scale of *very hard* to *very easy*, how would you rate the following:

	not used	very hard	hard	alright	easy	very easy
Asking for help	●	●	●	●	●	●

Did you find the information provided when asking for help useful? □ yes □ no If no, why:

Part 4 Interaction using language

1. Did you find the voice control of the system useful? □ yes □ no If no, why:

2. Which component did you find the most useful to access using voice?

3. Which component did you find to be the least useful to access using voice?

4. Did you find that you could talk to the system easily?

Video and Audio

On a scale of *very hard* to *very easy*, how would you rate the following:

	not used	very hard	hard	alright	easy	very easy
Starting the video	●	●	●	●	●	●
Controlling the video (start/stop etc)	●	●	●	●	●	●
Starting the audio	●	●	●	●	●	●
Controlling the audio (start/stop etc)	●	●	●	●	●	●

Part 6 The icons and graphics

If there were any graphics that you would like to comment on, please use the space below.

Part 7 Other comments

If you have any other comments about the system, or advice on how to improve it, please provide them in the space below.

Would you agree to evaluate future versions of the system once we have made improvements?

□ Yes □ No

You've now finished. Thank you for helping us to evaluate the Archivus system!

Appendix F: Example of question sheet from the Archivus pilot experiments

Tasks

Your manager has asked you to find several things. Most of these involve checking whether he has remembered certain facts from the meeting correctly, or finding bits of information that he has forgotten. He also has a copy of a part of a document and needs to know how it fit into the meetings.

All of these tasks are divided into 5 sections. Please do the sections and questions in the order in which they are given. If you feel that you really cannot find an answer to a question and you want to move on, simply select the *Task Finished* button and start with another question. Once you have found an answer, select the *Task Finished* button in the interface. You can write the answers directly on this page.

It is important to remember that we are testing how well the Archivus system helps you do your tasks and NOT your ability to answer the questions correctly. We ask that you try to find the correct answer as quickly as you can, but you should not rush to finish, or try to guess the answers.

Part 1 – Please find whether each of the statements below is true or false.

	True	False
1. The budget for the lounge furnishing was 1000CHF.	□	□
2. The dimensions of the lounge are 375cm by 477cm.	□	□
3. All of the meetings took place in Martigny.	□	□
4. Appliances were discussed in meeting ISSCO-34.	□	□
5. Denis proposed a brain-storming area.	□	□

Part 2 - Please find the answer to each question below.

1. How many photographs are there in the Google document? _____
2. How many armchairs were finally chosen for the room? _____
3. Who attended all of the meetings? _____
4. What items were on the agenda for meeting ISSCO – 36? _____

   _____
5. Who was late to the ISSCO – 35 meeting? _____

Part 3 - Find who presented the following part of a document, and in which meeting.

_____

Part 4 - Please find whether each of the statements below is true or false.

	True	False
1. They decided to put a sofa in the lounge.	□	□
2. The whiteboard was used to draw in meeting ISSCO-35.	□	□
3. Denis proposed a brain-storming area.	□	□
4. The ISSCO-37 meeting took place on March 10th, 2004.	□	□
5. Carpets were discussed in one of the meetings.	□	□
6. They decided on a relatively neutral colour scheme.	□	□

Part 5 – Please find the answer to each question below.

1. How many articles were mentioned during the meetings? _____
2. Which meeting did Susan not attend? _____
3. What colour scheme did Andrei propose? _____
4. How much did the sofa proposed by Agnes cost? _____
5. Who presented an overview of the problem of furnishing the lounge? _____

That completes the set tasks for the evaluation.

If you still have time, please feel free to browse the meetings using the interface in any way that you like until your time is finished.

Appendix G: Questions from the Archivus final experiment

Notes:

- Questions for final experiment (summer/autumn 2006)
- The questions are in the order in which they were presented to the user
- Odd numbers are true/false questions, and even numbers are short answer questions
- In most cases phase 2 started at questions 21.

1. The Furniture 4 meeting took place on March 10th, 2004.
2. Which two movies does Agnes suggest showing?
3. Appliances were discussed in the Furniture 1 meeting.
4. Where was the Design meeting held?
5. The movie club has already shown 'Lawrence of Arabia'
6. Which two participants brought Powerpoint presentations to the movie club meeting?
7. One of the meetings took place in Geneva.
8. Who attended all of the meetings?
9. Denis proposed a brain-storming area.
10. Who was the marketing expert in the Design meeting?
11. They suggested that the remote controls could be customized.
12. How many pictures are there in the Google document?
13. They decided to put a sofa in the room.
14. Who was leading the Design meeting?
15. Mirek and Andrei both suggested showing 'American Beauty'
16. What was agreed to be the minimum size of the armchairs?
17. There was disagreement about the purpose of the room
18. Which movie did they finally decide to show?
19. A prototype of the remote control was presented at the meeting.
20. What things did Susan disagree about in Agnes's presentation in Furniture3?
21. They considered adding speech recognition in the design of the remote control
22. What is the name of the company on the Design meeting slides?
23. Someone brought up the question of taking furniture outside.
24. In which movie was the colour saturation modified?
25. The budget for the room furnishing was 1000CHF.
26. How many movies does Denis suggest to the group?
27. Andrei is the president of the movie club.
28. What material did they finally decide to make the remote control out of?
29. Denis showed 4 possible versions of the movie club advertising poster at the meeting.
30. Which meetings did Susan not attend?
31. Carpets were discussed in at least one of the meetings.
32. By which other movie was the movie 'The Big Lebowski' inspired?
33. No one in the meeting has seen the movie "Usual Suspects".
34. What colour scheme did Andrei propose?

35. There is a long discussion about the movie ‘Saving Private Ryan’
36. What items were on the agenda for meeting Furniture 3?
37. It was suggested to design a remote control with a flip-up screen.
38. Which type of chip includes a sensor?
39. The date of the next movie club meeting is May 3rd
40. How many laptops were used during the Design meeting?
41. Andrei and Denis talk about the awards that Steven Spielberg has won.
42. How much did the sofa proposed by Agnes cost?
43. The dimensions of the room are 375cm by 477cm

Appendix H Classification of Archivus task questions

Classification of the questions in the WOz evaluation based on what type of information the user needs to find

Marita Ailomaa 16.01.2007

1. Global information about a meeting
   - (4) Where was the Design meeting held?
   - (7) One of the meetings took place in Geneva.
   - (30) Which meetings did Susan not attend?

2. Person
   - (6) Which two participants brought Powerpoint presentations to the movie club meeting?
   - (8) Who attended all of the meetings?
   - (10) Who was the marketing expert in the Design meeting?
   - (14) Who was leading the Design meeting?
   - (27) Andrei is the president of the movie club.

3. Decision
   - (13) They decided to put a sofa in the room.
   - (16) What was agreed to be the minimum size of the armchairs?
   - (18) Which movie did they finally decide to show?
   - (28) What material did they finally decide to make the remote control out of?

4. Suggestion
   - (2) Which two movies does Agnes suggest showing?
   - (26) How many movies does Denis suggest to the group?
   - (34) What colour scheme did Andrei propose?
   - (42) How much did the sofa proposed by Agnes cost?

5. Topic of disagreement:
   - (20) What things did Susan disagree about in Agnes's presentation in Furniture3?
   - (17) There was disagreement about the purpose of the room

6. Fact (found easily with keyword or topic search)
   - (3) Appliances were discussed in the Furniture 1 meeting.
   - (5) The movie club has already shown 'Lawrence of Arabia'
   - (9) Denis proposed a brain-storming area.
   - (11) They suggested that the remote controls could be customized.
   - (15) Mirek and Andrei both suggested showing 'American Beauty'

(19) A prototype of the remote control was presented at the meeting.
(21) They considered adding speech recognition in the design of the remote control.
(23) Someone brought up the question of taking furniture outside.
(24) In which movie was the colour saturation modified?
(25) The budget for the room furnishing was 1000CHF.
(31) Carpets were discussed in one of the meetings.
(32) By which other movie was the movie 'The Big Lebowski' inspired?
(33) No one in the meeting has seen the movie "Usual Suspects".
(35) There is a long discussion about the movie 'Saving Private Ryan'
(37) It was suggested to design a remote control with a flip-up screen.
(38) Which type of chip includes a sensor?
(41) Andrei and Denis talk about the awards that Steven Spielberg has won.

7. Item in a document or video (e.g. picture, logo)
   (12) How many pictures are there in the Google document?
   (22) What is the name of the company on the Design meeting slides?
   (29) Denis showed 4 possible versions of the movie club advertising poster at the meeting.
   (36) What items were on the agenda for meeting Furniture 3?
   (40) How many laptops were used during the Design meeting?
   (43) The dimensions of the room are 375cm by 477cm

List of Figures

Figure 1: The Archivus interface 73
Figure 2: The Archivus book 75
Figure 3: User's room during Wizard of Oz experiments 80
Figure 4: Wizard's room during Wizard of Oz experiments 81
Figure 5: Change in voice use over time 105
Figure 6: Change in keyboard use over time 106
Figure 7: Change in mouse use over time 107
Figure 8: Change in pen use over time 108
Figure 9: Modality change over time for the V condition 109
Figure 10: Modality change over time for the VK condition 110
Figure 11: Modality change over time for the M condition 111
Figure 12: Modality change over time for the P condition 111
Figure 13: Modality change over time for the MV condition 112
Figure 14: Modality change over time for the PV condition 113
Figure 15: Modality change over time for the MK condition 114
Figure 16: Modality change over time for the PK condition 115
Figure 17: Modality change over time for the MVK condition 116
Figure 18: Modality change over time for the PVK condition 116
Figure 19: Number of instances of top and bottom 20 success scores per condition 138
Figure 20: Frequency of top and bottom 20 for # of questions answered per condition 139

List of Tables

Table 1: Characterization of input devices 25
Table 2: Comparison of GUI and multimodal interface architectures (from Oviatt [62]) 40
Table 3: All possible modality combinations for WOz Archivus experiments 82
Table 4: Acronyms for modality conditions and phases 96
Table 5: User responses to post-experiment questionnaire 97
Table 6: Usefulness rankings of modalities for different Archivus functionalities 98
Table 7: Proportions of pointing and language used in each condition 99
Table 8: Proportions of use for each modality per phase and overall 101
Table 9: Voice and keyboard use proportions during P2 102
Table 10: Voice and keyboard use, in proportions, in P1 and P2 102
Table 11: Proportions of pointing and voice in both phases 103
Table 12: Total number of problems induced by each modality in P1 and P2 119
Table 13: Proportions of smooth interactions, misses and scope problems 120
Table 14: Proportion of smooth and problem switches in P1, P2 and overall 121
Table 15: Proportion of switches in problem and smooth interactions 122
Table 16: Modalities used to resolve problems in P2 for each condition 123
Table 17: Proportion of switches within and between categories 124
Table 18: Average number of interactions per modality 126
Table 19: Proportions of pointing and language (MV/PV & MK/PK) in P1 127
Table 20: Number of questions completed per condition and phase 129
Table 21: Success scores (averaged over a condition) in each phase and overall 133
Table 22: Average per question correctness score ranked by overall score 136

The author is grateful to the following for funding this research:

The University of Geneva
(Ecole de Traduction et d'Interprétation)

and

The Swiss National Science Foundation
for funding
the Swiss National Center of Competence in Research (NCCR)
on Interactive Multimodal Information Management (IM2)
which provided the research community and funds that made this thesis possible.