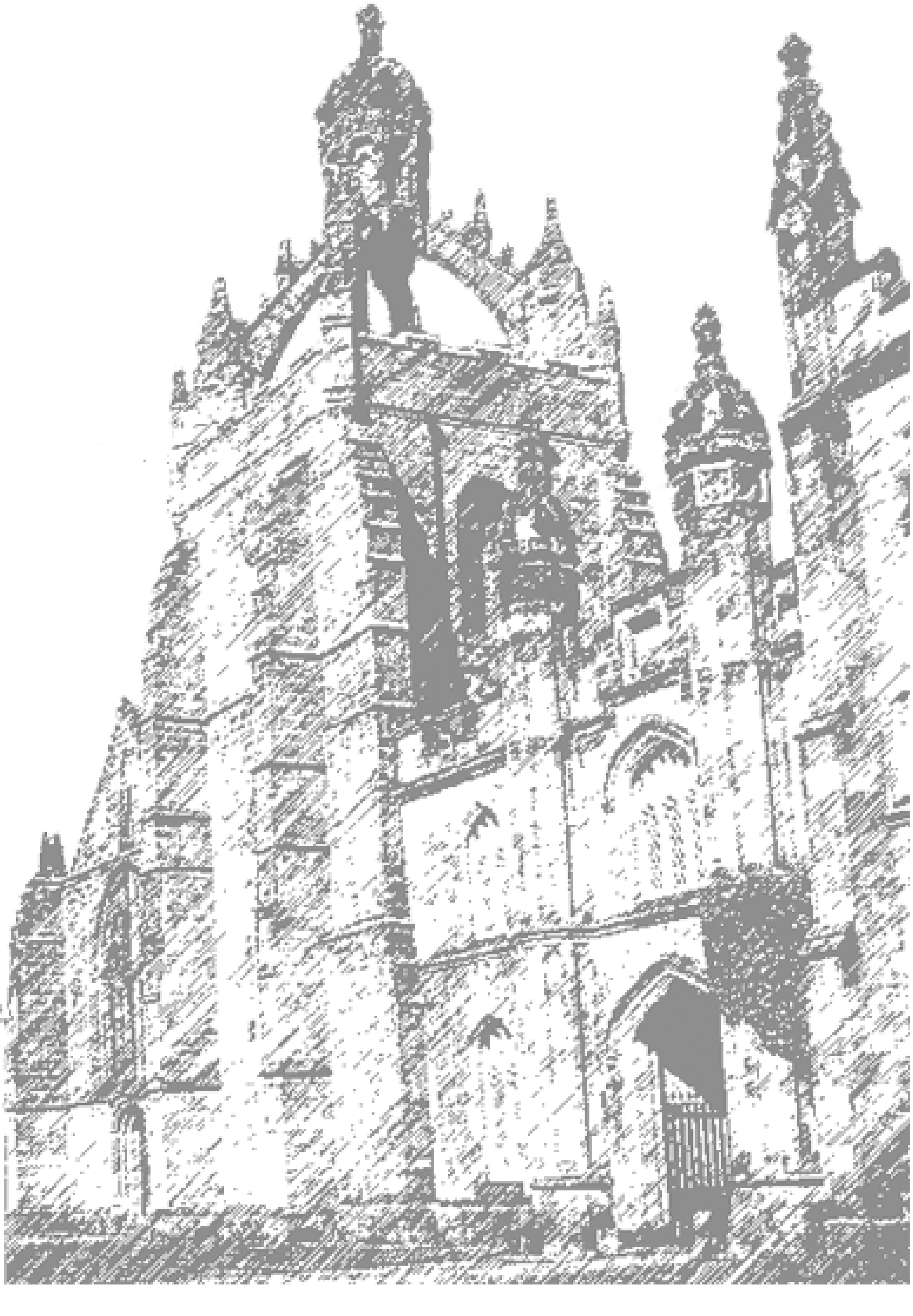
MOG 2008

Multimodal Output Generation

AISB '08

AISB 2008 Proceedings Volume 10



GLoRiClass

Sixth Framework Programme

aisb

aberdeen - 2008

1495

UNIVERSITY OF ABERDEEN

AISB 2008 Convention
Communication, Interaction and Social Intelligence

1st-4th April 2008
University of Aberdeen

Volume 10:
Proceedings of the
AISB 2008 Symposium on Multimodal Output Generation (MOG 2008)

Published by
**The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour**

http://www.aisb.org.uk/convention/aisb08/

ISBN 1 902956 69 9

Contents

The AISB'08 Convention: Communication, Interaction and Social Intelligence

As the field of Artificial Intelligence matures, AI systems begin to take their place in human society as our helpers. Thus it becomes essential for AI systems to have sophisticated social abilities, to communicate and interact. Some systems support us in our activities, while others take on tasks on our behalf. For those systems directly supporting human activities, advances in human-computer interaction become crucial. The bottleneck in such systems is often not the ability to find and process information; the bottleneck is often the inability to have natural (human) communication between computer and user. Clearly such AI research can benefit greatly from interaction with other disciplines such as linguistics and psychology. For those systems to which we delegate tasks: they become our electronic counterparts, or agents, and they need to communicate with the delegates of other humans (or organisations) to complete their tasks. Thus research on the social abilities of agents becomes central, and to this end multi-agent systems have had to borrow concepts from human societies. This interdisciplinary work borrows results from areas such as sociology and legal systems. An exciting recent development is the use of AI techniques to support and shed new light on interactions in human social networks, thus supporting effective collaboration in human societies. The research then has come full circle: techniques which were inspired by human abilities, with the original aim of enhancing AI, are now being applied to enhance those human abilities themselves. All of this underscores the importance of communication, interaction and social intelligence in current Artificial Intelligence and Cognitive Science research.

In addition to providing a home for state-of-the-art research in specialist areas, the convention also aimed to provide a fertile ground for new collaborations to be forged between complementary areas. Furthermore the 2008 Convention encouraged contributions that were not directly related to the theme, notable examples being the symposia on "Swarm Intelligence" and "Computing and Philosophy".

The invited speakers were chosen to fit with the major themes being represented in the symposia, and also to give a cross-disciplinary flavour to the event; thus speakers with Cognitive Science interests were chosen, rather than those with purely Computer Science interests. Prof. Jon Oberlander represented the themes of affective language, and multimodal communication; Prof. Rosaria Conte represented the themes of social interaction in agent systems, including behaviour regulation and emergence; Prof. Justine Cassell represented the themes of multimodal communication and embodied agents; Prof. Luciano Floridi represented the philosophical themes, in particular the impact on society. In addition there were many renowned international speakers invited to the individual symposia and workshops. Finally the public lecture was chosen to fit the broad theme of the convention – addressing the challenges of developing AI systems that could take their place in human society (Prof. Aaron Sloman) and the possible implications for humanity (Prof. Luciano Floridi).

The organisers would like to thank the University of Aberdeen for supporting the event. Special thanks are also due to the volunteers from Aberdeen University who did substantial additional local organising: Graeme Ritchie, Judith Masthoff, Joey Lam, and the student volunteers. Our sincerest thanks also go out to the symposium chairs and committees, without whose hard work and careful cooperation there could have been no Convention. Finally, and by no means least, we would like to thank the authors of the contributed papers – we sincerely hope they get value from the event.

Frank Guerin & Wamberto Vasconcelos

The AISB'08 Symposium on Multimodal Output Generation (MOG 2008)

Welcome to Aberdeen at the Symposium on Multimodal Output Generation (MOG 2008)! This year MOG is held as a part of the AISB Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour that is organised around the theme Communication, Interaction and Social Intelligence. Similar to MOG 2007, this year's MOG aims to present the state of the art and identify future research needs in multimodal output generation. MOG 2008 proves again to be successful in bringing together work from different disciplines which is usually scattered across various events. Besides contributions from research fields such as multimodal language generation and embodied conversational agents (ECA's), MOG 2008 has an additional angle by investigating how research on multimodal output generation can benefit from a non-engineering perspective on multimodality. For example, how can insights from psychology and cognitive sciences, related to understanding how humans perceive and process multimodal information, be properly formalized for the purposes of intelligent multimodal output generation? And to what extent is it possible to formalize existing theories about how meaning is made in multimodal communication?

This year, we are pleased to welcome two invited speakers: Dr. Michelle Zhou (IBM T. J. Watson Research Center) and Professor Justine Cassell (Northwestern University).

In this volume the papers presented at the MOG 2008 international symposium are collected. Five papers introduce different approaches to two of the most fundamental topics for multimodal output generation-output planning and modality choice. Gerrit Kahl et al. propose three multimodal output generation strategies: user-defined (output modalities explicitly selected by the user), symmetric multimodal (using the same modalities for output as the ones used for input) and context-based (generating an optimal multimodal output based on different factors in the user's current situational context). Yujia Cao proposes a modality planning framework which can optimally allocate one or more modalities to a chunk of information and also calculate an optimal combination of modalities. Central to the framework is a computational optimization model which takes as input the information to be conveyed, modality availability, and the user profile, translates those into constraints, and outputs a modality plan. In his contribution Flavio Soares Correa da Silva proposes a knowledge-based layer which selects the output modalities that present medical information on mobile devices of a telehealth system. The proposed layer does modality selection based on device capabilities, type of interaction, type of information and features describing modalities. The enlisted knowledge sources are described by respective ontologies or reference models. Verena Rieser and Oliver Lemon adopt a machine learning approach to modality selection. They develop and evaluate adaptive multimodal dialogue strategies using simulation-based reinforcement learning showing that the latter approach, which provides additional information about the user in the reward function, overperforms supervised learning which only mimics the data. Bosma et al. present a method for automatic text illustration, based on the similarity between the text to be illustrated and picture-related text. A user study conducted to evaluate their method showed that when compared to manually selected pictures, automatically selected pictures were rated similarly to decorative pictures, but worse than informative pictures.

Four papers are dedicated to the topic of ECAs' speech and gesture. Kirsten Bergmann and Stefan Kopp present a computational perspective on the joint production process for speech and gesture. Based on empirical evidence indicating a mutual influence of speech and gesture in utterance production, they propose an interface between imagistic and propositional knowledge at the level of content representation. Beatriz López et al. explore the possibilities of using ECAs to improve the robustness and the interactive fluency in spoken dialogue systems. Results of a comparison between two interfaces, one with an ECA and one with a voice-only output, show that user frustration is lower and interaction flows more smoothly when an ECA is present in the interface. In their paper, Rieks op den Akker and Mariët Theune discuss the main literature on addressing and present both qualitative and quantitative analyses of addressing in multi-party, face-to-face conversations. Based on these findings they sketch a model for the generation of multimodal addressing behaviour. Finally, Werner Breitfuss et al. introduce a system that automatically adds gaze and gesture to a given dialogue script between two virtual embodied agents in the roles of speaker and listener. The quality of the gaze generation was empirically tested showing that the naturalness of the agent's behaviour was not increased when compared to randomly selected gaze behaviour, but the quality of the communication between the two agents was perceived as significantly enhanced.

Topics related to corpora studies are investigated in two papers. In their contribution Michael Barclay and Antony Galton describe initial steps in the design of a scene corpus for training and testing spatial communication systems. Such a scene corpus needs to allow a full range of spatial relations to be expressed over a range of scale spaces, the scenes should be sufficiently complex to allow the construction of sequential spatial descriptions and the integration of listener models and reference frame variations should be possible. Ielka van der Sluis et al. describe the experiment setup with which a transparent multimodal dialogue corpus was collected. The corpus is currently being transcribed and will be used to test hypotheses about human production as well as hypotheses about human perception of referring expressions that include pointing gestures.

And finally two papers extend the interdisciplinary approach to multimodal output generation into semiotics and pragmatics by discussing the incorporation of principles informed by these two disciplines. In his paper Frédéric Landragin proposes a set of principles for the design of multimedia presentation systems that are based on pragmatics and human factors such as the specificities of human perception, attention, memory, conceptualization and language. Arianna Maiorani applies the multimodal discourse analysis method to the study of Internet as a multimodal semiotic system. The paper tests the validity of the functional framework used in linguistics to study online multimodal communication.

Thanks are due to the programme committee members: Adrian Bangerter, Ellen Gurman Bard, John Bateman, Harry Bunt, Stephan Kopp, Emiel Krahmer, Theo van Leeuwen, Anton Nijholt, Jon Oberlander, Niels Ole Bernsen, Paul Piwek, Ehud Reiter, Jan Peter de Ruiter, Jacques Terken, Eija Ventola, Ipke Wachsmuth, and Marilyn Walker. We would also like to thank the additional reviewers: Christian Becker-Asano, Nikolaus Bee, Kirsten Bergmann, Trung Bui, Yujia Cao, Nadine Pfeiffer-Lessmann and Mannes Poel.

MOG 2007 is endorsed by SIGGEN (ACL Special Interest Group on Generation) and SIGMedia (ACL Special Interest Group on Multimedia Language Processing). The symposium is also sponsored by NWO via IMOGEN (Interactive Multimodal Output Generation), a research project within the NWO-IMIX research programme. We are grateful to all these supporting organizations. At the University of Aberdeen a smooth AISB organisation team made it possible to organise this years MOG symposium. Very special thanks are due to Frank Guerin and Wamberto Vasconcelos.

Mariët Theune
Ielka van der Sluis
Yulia Bachvarova
Elisabeth André

Symposium Organisers:

Mariët Theune, University of Twente, The Netherlands
Yulia Bachvarova, University of Twente, The Netherlands
Elisabeth André, University of Augsburg, Germany
Ielka van der Sluis, University of Aberdeen, UK

Programme Committee:

Adrian Bangerter, University of Neuchâtel, Switzerland
Ellen Gurman Bard, University of Edinburgh, UK
John Bateman, University of Bremen, Germany
Harry Bunt, Tilburg University, The Netherlands
Stephan Kopp, University of Bielefeld, Germany
Emiel Krahmer, Tilburg University, The Netherlands
Theo van Leeuwen, University of Technology Sydney, Australia
Anton Nijholt, University of Twente, The Netherlands
Jon Oberlander, University of Edinburgh, UK
Niels Ole Bernsen, University of Southern Denmark
Paul Piwek, Open University, Milton Keynes
Ehud Reiter, University of Aberdeen, UK
Jan Peter de Ruiter, MPI, The Netherlands
Jacques Terken, Eindhoven University, The Netherlands
Ipke Wachsmuth, University of Bielefeld, Germany
Marilyn Walker, University of Sheffield, UK

Automated Multimodal Generation in Context-Sensitive Information Systems

Michelle Zhou[1]

ABSTRACT

Imagine the next generation of information portals, where users are able to obtain information through an intelligent multimodal conversation that is tailored to the tasks they are performing, customized to their personal preferences, and adapted to their context and interaction devices. To realize this vision, we are building an intelligent user interaction framework that helps to bridge the gap between what users want and what a current system can provide.

Our framework encompasses two sets of core technologies: input technologies and output technologies. Our input technologies allow users to employ a combination of input modalities, including natural language and visual query, to express their information needs in context naturally and efficiently. On the other hand, our output technologies allow a system to automatically synthesize a multimedia response to a user's request, including both verbal and visual outputs, which is tailored to the user's interaction context, including the conversation flow and the user's personal interests.

In this talk, I will highlight the use of automated multimodal generation in both our input and output technologies. As part of our input technologies, I will present how we use automated multimodal generation technologies to dynamically create cross-modality confirmations during a user's input process. Specially, when a user employs one input modality like natural language to express his/her request (*"shipments containing T42p laptops"*), the system automatically creates the interactive representation of the same request in a complementary modality such as visual query. As a result, a user can easily switch the use of different modalities whenever needed in the course of interaction without losing the query context that s/he has built so far. Furthermore, cross-modality confirmations help to teach the user about the system's capability in supporting different input modalities. Besides supporting context-sensitive user input, automated multimodal generation is also the core piece of our output technologies. In this talk, I will focus on the practical issues in developing automated multimodal generation technologies for real-world applications. In particular, I will highlight our effort in developing optimization-based approaches to automated multimodal generation with a concrete example on a graph-matching approach to multimodal allocation.

[1] IBM T. J. Watson Research Center

Knowledge-based Modality Selection for Information Presentation in a Mobile System for Primary Homecare

Flavio Soares Correa da Silva[1]

Abstract. Public homecare programs have proven to be very effective for Preventive Medicine. In Brazil, the Family Health program, initiated in the late 1990s, has taken medical doctors, nurses and social workers to the homes of lower income population in underserved urban regions. This program has been developed using nearly no IT support for its operations, and we believe that this gives room to opportunities to improve its efficiency. The Borboleta project aims at providing the Brazilian Family Health program with IT support, more specifically making use of mobile computing technologies to improve the quality and reliability of services provided to the population. Many technological challenges must be solved in order to achieve the project goals, among which we highlight *information logistics* - the necessity to bring appropriate information presented at the appropriate format to nurses, medical doctors and social workers performing field work. In order to do that, we are working on a knowledge-based presentation layer that can sense the context of the interactions between field workers and information sources and semi-automatically select the most adequate output modality for information presentation.

1 INTRODUCTION

Telehealth systems are computer based systems aiming at the provision of healthcare using telecommunication technologies. In large urban areas in which a considerable proportion of the population has lower income, telehealth systems can be of particular relevance to bring healthcare to underserved communities.

In Brazil, the Family Health program, initiated in the late 1990s, aims at the provision of *preventive* healthcare to underserved communities. This program has proved to be effective, despite the extremely low level of technological sophistication it has employed so far.

The Borboleta project [2] aims at leveraging the Family Health program with mobile computing technologies. We believe that this can improve the efficiency and the reliability of this program, thus making it even more useful to the society.

Essentially, the Borboleta project provides field workers medical doctors, nurses and social workers who go to the homes of underserved citizens to collect information about them and provide them with assistance and orientation about hygiene and preventive healthcare with portable computing and communication devices (typically, PDAs and smartphones), enabled to exchange data with computer servers located at a Central Hospital.

The data exchange between the mobile devices and the servers must be done in such way as to ensure the privacy of medical records, as well as the reliability of information. Additionally, the information presented in the mobile devices must reach the field workers as effectively as possible, to ensure the effectiveness of the visits to the homes of the population.

In [1] a knowledge-based approach is proposed for the selection of the modality of information presentation, in which a Modality Ontology is introduced. We intend to found our work on the ontology introduced in [1], and build a knowledge-based system for output modality selection for the Borboleta project.

The present article is a work-in-progress report. We are at the moment working on the detailed design of the knowledge-based system for output modality selection. In future reports we shall present implementation and empirical results related to the utilization of this system.

In section 2 we briefly describe the architecture of the Borboleta project, highlighting the information that is conveyed through the mobile devices at the present implemented prototype. In section 3 we briefly review the Modality Ontology introduced in [1], and explain a little further how it can be employed in the Borboleta project. In section 4 we outline our proposed layer for knowledge-based modality selection for information presentation. Finally, in section 5 we present some discussion, preliminary conclusions and planned future work.

2 THE BORBOLETA ARCHITECTURE

The Brazilian Family Health program is managed from central hospitals, and effectively run by professionals who visit the homes of families to provide them with medicine and information to prevent health problems. These professionals are a few medical doctors and a host of nurses and social servers with some basic training on healthcare, to whom we shall refer heretofore as field workers. Due to the focused and instrumental training that field workers have, they are not entitled to make complex decisions related to medical interventions. Given that a limited number of medical doctors are available, the system must be as assistive as possible in order to provide the field workers with the necessary information to carry on their activities.

The field workers have a strict routine to follow. Based on a general work plan, each field worker starts the day with the detailed schedule of visits to be made in that day. The schedule is defined collaboratively, so that the team of field workers can ensure that all families in their region are being visited.

Once a field worker gets his/her schedule for the day, he/she gathers from file cabinets the corresponding (paper based) forms and records for the families that are going to be visited.

On the way to a family home, the field worker studies the medical records corresponding to that family and devises a visit plan. The

[1] University of Sao Paulo, BRAZIL, email: fcs@ime.usp.br

visit consists of enquiries about health conditions of the whole family living in a specific address, comparative analysis with respect to previous records and provision of advice to ensure good health conditions for all.

Typically, because of heavy workload, the field workers do not take many notes during visits. The collected information is added to the records later typically three to seven days later at the central hospital. Information can be lost and become less reliable because of this delay, since the field workers rely heavily on memory to feed information back to the system.

One of the main goals of the Borboleta project is to make the Family Health program more efficient and reliable, by provisioning field workers with mobile communications and processing technology. Our goal is to provide field workers with expert information in real time during field work, as well as to provide them with appropriate means to feed information back to the system immediately after this information is collected.

In order to do so with the required efficiency, information flow must be as unobtrusive as possible in the workflow of field workers. It is of paramount importance that information is provided to field workers at the exact time and at the most appropriate format, and that the system provides the field workers with the best possible means to input information through the mobile devices they carry with them typically, PDAs and smartphones.

To our best knowledge, the Borboleta project is a rather innovative initiative. We have found some initiatives related to the utilization of mobile devices to bring information to medical doctors and to provide them with resources to feed information back to a database, but none of them are addressed to preventive healthcare or to underserved communities. For example, the Constellation project [3] connects medical doctors with a centralized database to interact with patient records within the Womens Hospital at Harvard University, and the MEDIC project [4] connects medical doctors with a centralized database to obtain the results of laboratory exams as soon as they become available.

We already have a prototype system for the Borboleta running and being tested by field workers. This initial prototype was built to solve the fundamental communications problems between the server and the mobile devices, as well as to set up the basic information structures that shall be required to support the full fledged Borboleta system.

In this initial prototype, the information that is provided to field workers is as follows:

- **Patient Personal Data:** personal data referring to the person being visited - name, date of birth, etc.
- **Patient Caregiver Data:** many patients have movement impairment, old age or any disabilities that require that a second person most typically some relative looks after a patient; this item contains to personal data referring to those people who look after other people.
- **Patient Socioeconomic Data:** socioeconomic data that can influence specific treatments provided to the individual professional activities, educational level, nationality, religion, etc.
- **Scheduled Visits:** schedule of future visits to the patient.
- **New Visit Registration:** a form to be filled in by the field worker during the visit.
- **Visit History of the Patient:** access to historical data about previous visits to the patient.
- **Diseases Catalog:** access to the *International Diseases Classification ICD-10* [5].
- **Drugs Catalog:** access to the list of available drugs and medicaments in stock at the central hospital.

3 THE MODALITY ONTOLOGY

The Modality Ontology [1] has been developed to provide support to the selection of the most appropriate modalities to present information in a multimodal information system.

Essentially, the Modality Ontology is a hierarchy of concepts that are relevant to the selection of an output modality (or combination of modalities). These concepts characterize attributes that can be found in different modalities, that can guide the process of rendering specific pieces of information or, alternatively, determine what pieces of information *can* be rendered in a specific device considering the attributes of modalities the device is prepared to handle.

The Modality Ontology classifies the modality attributes in two large groups:

1. Content: the specific sorts of information an output modality is capable of rendering.
2. Profile: operational features of output modalities. The profile is further classified in three sub-groups:
   (a) Information Presentation: characterizes features of modalities that identify how information is presented in each modality. For example, a modality can be characterized as *linguistic* (i.e. text based) or *analogue* (i.e. image based).
   (b) Information Perception: characterizes features of modalities that identify how information is captured by end users depending on each modality. For example, a modality can be characterized as *visual* or *auditory*.
   (c) Modality Structure: characterizes representational features that are specific of each modality. For example, a modality can be *pointer based* (e.g. maps, with which the user is expected to interact by pointing to specific locations) or *annotation based* (e.g. text, with which the user is expected to interact by using words). The modality structure also characterizes dependence relations across modalities (for example, text is usually independent of other modalities, but maps require a combination of images and annotations).

Each specific modality can be characterized using the terminology provided by the Modality Ontology. As detailed in the next section, different contexts require different attributes for a modality, and therefore a context of interaction between a field worker and information sources shall determine the required attributes and, as a consequence, the most appropriate modalities to be employed for specific purposes.

4 KNOWLEDGE-BASED MODALITY SELECTION FOR THE BORBOLETA

Our goal is to enhance the Borboleta system with a knowledge-based layer to select the modality of output data in mobile devices.

The selection shall be based on three sorts of information, which shall be determined when a mobile device sends a query to the server:

1. Device capabilities: depending on the multimedia capabilities of the mobile device in use, the possibilities for effective presentation of information are determined. For example, a smartphone with relatively large display can render text and images relatively well,

but if the display is too small or has low resolution, then images may become inadequate to convey guidance to the field workers (such as, for example, a reference image of what a social worker should identify as a symptom in a patient who may present some dermatological pathology); or a PDA with large memory and fast processor can render small videos relatively well, provided that the resolution is adjusted accordingly.

Device capabilities must be available for a system to decide how to best render output information. We suggest that every mobile device in use at the Borboleta system has an entry at a table, pointing to its corresponding capabilities. This table, together with the device capabilities, shall stay in the system server, so that the device must only communicate its ID together with any query to the server.

There are some alternatives to encode device capabilities that can be found in the literature. In order to abide by widespread standards, we intend to adopt the *CC/PP* ontology to encode device capabilities[2].

2. Type of interaction that is occurring during the field work: we intend to build a small ontology of interaction patterns specifically for the Borboleta project. An interaction pattern characterizes the activities being held by a field worker during the visit to the home of a citizen. Depending on what interaction pattern is being used, different levels of detail and emergency of feedback can be characterized. For example, the emergency of feedback can depend on attribute values such as:

   - Long visit: identifies a visit with no hard constraints on duration. The field worker can stay at the visited home for as long as necessary, interacting with the citizens who live there. In this case, the system response time admits some latency and more detailed information (e.g. higher resolution images and lengthier videos) can be appropriate.
   - Short visit: identifies a visit with hard and definite constraints on duration. The field worker may have, for example, a certain number of visits that must be done in one afternoon. In this case, response time becomes more important, and output format (e.g. image resolution and video length) can be selected in order to optimize performance.
   - Emergency situation: identifies a critical situation in which decisions must be taken as rapidly as possible. In this case, evidently the response time is the most critical feature of the interaction with the server, and data formatting must be selected accordingly.

   Long visits and short visits admit different levels of privacy of information presentation. For example, if the field worker and a patient are holding a private conversation, and if the information presented to the field worker can also be presented to the patient, then we can have less stringent privacy requirements for information presentation, that can enable for example videos with audio commentaries included; if the information cannot be presented to the patient, then audio may be inadequate; finally, if the field worker and the patient are in a room with other people, then more restrictive privacy requirements may be at place.

   The type of interaction must be informed by the field worker together with any query that is addressed to the server.
3. Type of information that is being requested: depending on the specific type of medical information that must be rendered, different output modalities may be more appropriate. For example, a dermatological symptom may be most effectively presented using images; a respiratory tract problem may manifest itself most clearly as a sound; and a neurological pathology may manifest itself as a specific pattern of movements that may be best presented in a short video.

   In order to classify the medical information appropriately, we shall employ the widespread standard *HL7 Reference Information Model*[3].

Information items in the server are already being prepared with a variety of alternative presentations. For example, we can have, for a specific item, a short video, a collection of still images extracted from the video and tagged with textual information, and a text summary of the video content. The envisaged utilization of the system presented here is as a tool to select, among alternative stored presentations for an information item, which presentation shall be sent to the renderer.

We shall maintain in the server two supporting tools to provide information to the mobile devices with the appropriate format:

1. A rule-based system to perform inferences based on device capabilities, type of interaction in field work and type of requested information, that shall advise on the best output modality for the requested information; and
2. An *information renderer* that, given the response to a specific query and a selected output modality, filters out unnecessary details and renders the obtained information using the selected modality, and then sends the rendered information to the mobile device.

The best output modality shall be characterized based on terms that are found at the Modality Ontology [1]. We envisage some possible situations in which more sophisticated reasoning mechanisms may be required:

1. Device, interaction pattern and requested information are underdetermined, i.e. the information communicated to the server is not sufficient to trigger any rule that could be used to infer a modality to be employed to render the requested information. We must include context sensitive default values for all relevant variables in order to cope with this situation, so that we always have at least one suggested modality being sent to the information renderer.
2. Device, interaction pattern and requested information are overdetermined, i.e. the information communicated to the server triggers many rules, which are used to infer several different modalities that could be employed to render the same piece of information. We must include context sensitive preference relations between output modalities to cope with this situation, so that we can always have at most one suggested modality being sent to the information renderer.

This additional layer shall provide the Borboleta system with resources to optimize the presentation of information to field workers taking into account the context of interaction with information resources.

5 DISCUSSION AND FURTHER WORK

The Borboleta project aims at provisioning a primary homecare system with mobile computing technology, thus improving its overall quality by making it more efficient and more reliable.

[2] http://www.w3.org/TR/CCPP-struct-vocab/

[3] http://www.hl7.org/

Among the several technological challenges posed by this project, we have highlighted in this article the necessity to provide field workers with appropriate information at the appropriate format.

To face this challenge, we are designing a layer for the Borboleta software architecture to manage the output modality of information sent to mobile devices used by field workers. In very general terms, this layer is comprised by a classification system that receives information from the mobile devices and associates this information with entries at appropriate ontologies that characterize the features of the information request; a rule-based inference system that infers features of the most appropriate output modality for the requested information based on the features of the request itself; and an information renderer that formats the requested information in such way that it presents the inferred features for the most appropriate output modality.

Evidently, our immediate future challenge shall be to implement this layer as part of the prototypical implementation of the Borboleta system, so that we can carry on with empirical evaluation of this proposed approach to optimize the presentation of information to field workers at the Borboleta project.

The Borboleta project is an open source project, and an initial prototype (which at the moment does not contain the information presentation layer) can already be found at its code repository (http://borboleta.incubadora.fapesp.br).

ACKNOWLEDGEMENTS

The Borboleta project is funded by the Microsoft-FAPESP Virtual Research Institute. The author has benefitted from additional partial financial support from the Brazilian funding agencies FAPESP and CNPq.

REFERENCES

[1] Y. Bachvarova, B. van Dijk, and A. Nijholt, *Towards a Unified Knowledge-based Approach to Modality Choice*, 5–15, Proceedings Workshop on Multimodal Output Generation 2007, Scotland, 2007.

[2] R. Correia, F. Kon, and Kon. R., *Borboleta: A Mobile Telehealth System for Primary Homecare*, ACM Symposium on Applied Computing, Brazil, 2008.

[3] S. Labkoff, S. Shah, J. Bormel, and Y. Lee, *The Constellation Project: experience and evaluation of personal digital assistants in the clinical environment*, 678–682, Proceedings 19th Annual Symposium on Computer Applications in Medical Care, 1995.

[4] E. McLoughlin, D. OSullivan, M. Bertolotto, and D. C. Wilson, *MEDIC Mobile Diagnosis for Improved Care*, 204–208, Proceedings ACM Symposium on Applied Computing, 2006.

[5] World Health Organization, *International Statistical Classification of Diseases and Health Related Problems ICD-10 (second edition)*, WHO, 2004.

Modality planning for preventing tunnel vision in crisis management

Yujia Cao and **Anton Nijholt** [1]

Abstract. Crisis management is a time-critical task with high uncertainty and high risk. Stress and cognitive overload often result in a set of bias in crisis manager's situation understanding and information confirming processes, known as "tunnel vision". Aiming at preventing tunnel vision, we propose an information assistant system which attempts to reduce the information quantity, improve the information quality, and prevent cognitive overload of the user. The main focus of this paper is to present the design proposal of the modality planning module. It is one of the modules which play a role in the prevention of tunnel vision. The function of this module is to determine the optimum utilization of the available modalities, in order to convey information effectively and reduce the cognitive load of the perceivers. The modality planning strategies also adapt to the user's preferences and cognitive state.

1 INTRODUCTION

A crisis is generally understood as an urgent situation with a negative outcome, such as a nature disaster, transport accident, civil attack, economical crash, etc. Crisis management is a strategic management activity aiming to prevent or minimize the negative impact of a crisis. It is a time-critical task with high uncertainty and high risk [8][11][15]. The crisis managers, who are located in the crisis response center, need to react quickly to the ongoing crisis event and make quick decisions. They also typically have to deal with information overload [15]. Under stress due to information overload and a lack of time, crisis managers tend to rely on standard operating procedures and their previous experiences without reexamination. When an understanding or a solution is forming, they have the tendency to perceive and confirm only clear and familiar information which aligns with it. Correct but ambiguous or contradicting information is easily discarded. We call the above phenomena "tunnel vision". Cognitive psychology theories provide better insight into the tunnel vision phenomena. When the decision makers tend to create one coherent interpretation without reexamining their experience with the real situation, they are experiencing "framing bias" [6] [18]. When they tend to confirm their understanding by seeking only the information which falls in harmony (evidence), they are experiencing "confirmation bias" [10]. Too much information and too little time might also cause cognitive overload [7]. The lack of cognitive capacity might deepen the biases. If an improper decision making "frame" (situation understanding) is continuously confirmed, the growing bias may lead to costly delay and errors.

Considering the specific task of crisis management, the computer has several advantages over the human brain. The computer is able to continuously record data into its memory with high speed, no matter of the quality of the data. It acts only on logic, without any influence by emotions. The computer also exceeds the brain in multi-tasking, fulfilling complex calculation and rule-based tasks. These advantages have high value when we attempt to develop a multimodal information assistant system (referred as "the system" below) which serves as a platform for monitoring the on-going crisis event. Aiming to reduce the likelihood of tunnel vision, the proposed system intends to provide the users with lower quantity but higher quality information in an effective and efficient manner. As one part of the design, this research focuses on the modality planning module. It is one of the modules which contribute to the prevention of tunnel vision. The function is to determine the optimum utilization of the available modalities, in order to convey information effectively and reduce the cognitive load of the perceivers.

Section 2 briefly describes the structure and function of the proposed system with a focus on the modalities related to the prevention of tunnel vision. Section 3 introduces previous work related to modality planning. Section 4 presents the primary research on the design of the modality planning module. The design of the other modules is out of the scope of this paper.

2 THE INFORMATION ASSISTANT SYSTEM

The general function of the system, as shown in figure 1, is a platform for monitoring a crisis event. The users are crisis managers located in the crisis response center, facing a large display. Briefly speaking, the system continuously captures the real world data (speech, video and sensor signals), records them into its memory, and simultaneously presents the on-going crisis event through the large display and speech. Via an information query interface, the user is allowed to access the crisis history (e.g. events that occurred in the previous minute) or some statistics (e.g. number of victims in area A). The crisis managers don't conduct their commands via the system. However, their commands will also be captured by the system and presented.

Three modules in the system are responsible for the prevention of tunnel vision: the reasoning and filtering module, the order planning module, and the modality planning module. The reasoning and filtering module helps to improve the information quality and reduce the information quantity. It groups reduplicate data together and provides statistic analysis on incompatible information. It also applies context reasoning based on predefined guidelines which are generally valid for certain types of crisis. The order planning module determines the presentation priority for each input information unit based on the overall evaluation of the time sequence, the urgency level, and the causal relations. The aim of this module is to guarantee that the most important and urgent information arrives at the user first. The modal-

[1] Human Media Interaction Group, University of Twente, PO BOX 217, 7500 AE Enschede, The Netherlands. y.cao@utwente.nl

ity planning module intends to present the crisis scenario effectively and efficiently by calculating the optimum utilization of the available modalities. Effectiveness means that the presentation does convey the information content correctly. Efficiency indicates that the presentation manner helps to prevent cognitive overload of the user. We expect that the users of such a system have larger chance to keep aware of the actual situation, stay open-minded for all possibilities and make proper decisions.

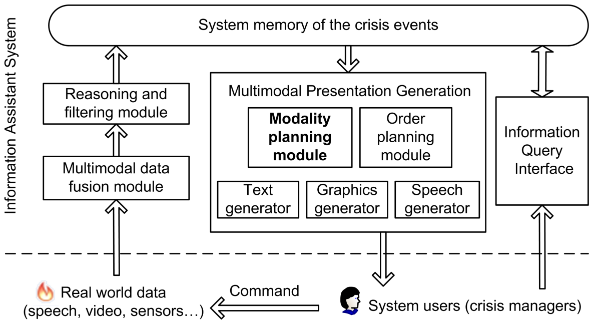

Information Assistant System
System memory of the crisis events
Reasoning and filtering module
Multimodal data fusion module
Multimodal Presentation Generation
Modality planning module
Order planning module
Text generator
Graphics generator
Speech generator
Information Query Interface
Real world data (speech, video, sensors…)
Command
System users (crisis managers)

Figure 1. Multimodal information assistant system structure

In the remainder of this paper, only the modality planning module will be discussed in more detail. There are many other on-going researches in the ICIS project (see acknowledgement) that can be applied to the design of the other modules of this system.

3 RELATED WORK

In the modality theory of Bernsen [3], a solid taxonomy of unimodal output modalities is designed to provide a theoretical foundation for understanding and generating multimodal output. Based on the observation that different modalities have different representational properties (such as linguistic, analog, arbitrary etc.), the taxonomy classifies all possible unimodalities into 4 classes in the super level, 20 classes in the generic level, and 46 classes in the atomic level. Possible extensions to a subatomic level have also been suggested. The taxonomy is claimed to be complete, unique, relevant, and intuitive. The different representational properties make different modalities suitable for expressing different types of information. For instance, linguistic modalities (e.g. text, discourse) surpass analogue modalities (e.g. images, graphics, diagrams) at explaining abstract concepts; while analogue modalities are better at expressing what things exactly look like. Their combination may have superior expressive power [2]. The combination of linguistic and analogue modalities has been adopted in many existing multimodal HCI systems, such as COMET [5], WIP [16], SmartKOM [17], and EMBASSI [4].

Concerning the cognitive load a presentation manner may place on the user, designers need to consider not only the representational properties of modalities, but also the perception properties (e.g. visual, auditory and haptic). The perception properties determine how a modality is perceived and processed by the human perceptual-sensory system [1]. The dual coding theory of Paivio [13] claims that humans possess separate information processing channels for visual and auditory material. Therefore, working memory has partially independent processors for handling visual and auditory signals. Mousavi et al. [9] suggest that the mixed use of both modalities can reduce cognitive load, because more effective cognitive capacity is available.

The modality planning task is generally agreed to be highly complex, due to many issues involved. So far, no solution in the form of a generally applicable automated modality planner has been devised yet. Instead, most of the existing approaches focus on a small set of modalities and a certain type of application. The same goes for own research. The modality planning task is commonly considered as a mapping process from the presentation task (convey certain information) domain to the modality domain, based on pre-designed rules or strategies. In the WIP system [16], a set of presentation strategies has been defined for all presentation tasks. They are represented by a name, a header, the applicability conditions and a specification of modality choice. When the presentation planner receives a presentation task, it tries to match a presentation strategy which has the corresponding effect or header. When there are more than one matches, pre-defined meta-rules are applied to make a choice. In the SmartKOM system [17], based on 121 presentation strategies, the presentation planner recursively decomposes a high-level presentation task into primitive presentation tasks and allocates different output modalities to each primitive presentation task. In the EMBASSI system [4], the combination of several unimodalities is defined as a multimodality. The model of a multimodality includes the set of unimodalities, the combination strategy, and the assignment to a physical output device. The combination strategy describes the synchronization, the necessary coordinations for multimodal references to objects, and the possible cross-modal references of the unimodalities. When receiving a presentation task, the presentation planner examines the user preference and the output device condition, and then assigns one or more multimodalities, and constructs the presentation according to the combination strategies.

4 THE MODALITY PLANNING MODULE

The design of the modality planning module aims at achieving the effectiveness and efficiency of the presentation. The modality planning process takes the following factors into account: 1) the information to be conveyed (presentation task) 2) the available modalities 3) the preferences of the crisis manager, and 4) the user's cognitive load status. Currently, we use a tunnel fire crisis scenario. The system represents the scenario by recording the actions of human actors and the state of the world. We have defined a limited set of action types, e.g. request, report, command etc. Modality strategies are designed for each action type, under several different conditions. Therefore, the modality planning approach is to use the strategy which matches the input action type, the user's profile and the user's cognitive statues.

4.1 THE PRESENTATION TASK

The system memory contains a world state database. Based on a world model (ontology), the system creates an instance, in the world state database, for each real-world entity that is involved in the crisis event (tunnel, fires, vehicles, human actors etc.) The properties of the instances (location, urgency level, stress level etc.) may change over time as the crisis event develops. We assume that the world state can only be changed by actions. If an action changes a certain property of a certain entity instance, the system records it as an "Action-StateChange" pair and makes a corresponding modification to the world state database. For example, when the system receives the report from the fire team that the fire has been put off, it creates the following action and state-change instances.

Action-1

- Type: Report

- TimeStamp: 16:45:28
- Actor: FireTeam
- Receiver: CrisisManager
- Content: State-Change-1
- UrgencyLevel: Low

StateChange-1

- TimeStamp: 16:45:28
- Object: Fire
- Property: Status
- Value: OFF

If an action (e.g. request) doesn't directly bring any change to the world state, no StateChange instance will be created. The Action-StateChange pairs or action instances are input into the modality planner as its presentation task. The planner makes the strategy match and applies the corresponding modality allocation and combination schema.

4.2 THE AVAILABLE MODALITIES

The system adopts both visual and auditory modalities. Visual modalities include map, text, image. Auditory modalities include speech and sound effects. Sutcliffe et al. [14] have introduced a set of attention effect advices for directing the user's attention to the appropriate information at the correct level of detail. Following these advices, dynamic text and dynamic image are used when extra attention is needed. Based on Bernsen's modality taxonomy [3], the properties of the available modalities are listed in table 1.

Table 1. The properties of the available modalities (based on [3])

Unimodality	Properties
Static Text	(li,-an,-ar,sta,gra)
Dynamic Text	(li,-an,-ar,dyn,gra)
Map	(-li,an,-ar,sta,gra)
Static Image	(-li,an,-ar,sta,gra)
Dynamic Image	(-li,an,-ar,dyn,gra)
Speech	(li,-an,-ar,dyn,aco)
Effect Sound	(-li,-an,ar,sta/dyn,aco)

"li": linguistic; "an": analogue; "ar": arbitrary; "sta": static; "dyn": dynamic; "gra": graphics; "aco": acoustics; "-": not

In order to specify the detailed utilization of the modalities, modality models are constructed for text, map, image, speech, and alarm sound, respectively. The modality model contains a set of parameters which describes the utilization details of the modality (see table 2). It can be viewed as a template for creating modality instances. Here, we don't separate static use and dynamic use. These properties are described by the parameter value. Therefore, static text and dynamic text share the same modality model. The same goes for static images and dynamic images.

Table 2. The modality Models

Unimodality	Model Parameters
Text	Content, ReferTo, Style, Size, Color, Blink, StartTime, Duration, DisplayArea, ScrollDirection, ScrollSpeed,
Map	Country, Province, City, InvolvedArea, DisplayedArea
Image	Source, ReferTo, DisplayArea, StartTime, Duration, Blink
Speech	Content, ReferTo, Tone, Speed, StartTime, RepeatTime
EffectSound	Source, ReferTo, StartTime, RepeatTime

When fulfilling a specific presentation task, one or more modality instances will be created. Their parameter values also indicate the combination manner. For example, the presentation task is to show the location of the policeman. The modality planner locates an image of a policeman on the map together with a text explanation. The image and text instances are created as follows. The parameter values are filled by the selected modality planning strategy (see section 4.3). The values of "StartTime", "DisplayArea", and "ReferTo" parameters indicate that the two modality instances will be shown at the same time, near to each other, and both refer to the policeman.

Text-1

- Content: In Gate Street, 550M to tunnel
- ReferTo: Policeman.Location
- Style: Arial, bold
- Size: Middle
- Color: Black
- Blink: N
- StartTime: immediate, align with Image-1
- Duration: 30 seconds
- DisplayArea: Rectangle [DisplayCoordination(300,560),
  DisplayCoordination(450,590)]
- ScrollDirection: N
- ScrollSpeed: N

Image-1

- Source: policeman.jpg
- ReferTo: policeman
- DisplayArea: Rectangle [DisplayCoordination(350,500),
  DisplayCoordination(400,550)]
- StartTime: immediate, align with Text-1
- Duration: 30 seconds
- Blink: Y

4.3 THE MODALITY PLANNING STRATEGY

In this section, we present the design proposal of the modality planning strategies. The design aims at achieving the presentation goal, i.e. effectiveness and efficiency. The desired presentation manner conveys the information content correctly and helps to prevent cognitive overload. A design proposal is presented in this section. Each presentation task has its own modality planning strategy. A strategy contains three items: 1) suitable modalities, 2) default strategy, and 3) light strategy. The choice between the default strategy and the light strategy is based on the user's cognitive state. When the user's preference is available, an adapted version of the default strategy will be generated.

4.3.1 Suitable Modalities

This item indicates which modalities are suitable for contributing to a certain type of presentation task and what each suitable modality expresses. The values of the "ReferTo" parameter will be filled in. In our crisis management application, the map is always shown as background on the display. However, it will be listed as a suitable modality only when the presentation of a action type needs to make use of it. Recall the example of showing the location of the policeman. Sound effects can do little to show a location. Image and text are selected as suitable modalities. The image refers to the policeman and the text refers to the location of the policeman.

4.3.2 Recommended Strategy and its adaption

The default strategy is designed to achieve the optimum presentation manner for a certain type of task. First, one or more suitable modalities will be selected. Based on the dual-coding theory [13], if the suitable modality list contains both visual modalities and auditory modalities, their combination owns higher priority. Second, the default strategy contains a specification of how to generate modality instances of all the selected modalities. As mentioned before, the parameters of these modality instances also indicate their combination manner. Third, following the attention effect advices in [14], this strategy also attempts to attract the user's attention to what is being presented. For instance, fire alarm (sound effect) is used for a fire report; ambulance alarm is used for a victim report. When necessary, the speech speed is increased with warning tone. Dynamic texts and dynamic images are also often used.

The default strategy will be applied when the user has no specific preference and no cognitive overload of the user is recognized. If the user especially prefers certain modalities, an adapted version of the default strategy will be made. The adaption to cognitive overload will be described in the following subsection. The adaptation to a user's preferences intends to avoid undesired annoyance for the user. If the user prefers a certain modality, it will always be selected, as long as it is on the list of suitable modalities. The user can also indicate that he prefers a less intrusive presentation manner. Then attraction efforts (e.g. using sound effect, warning tone etc.) will be reduced. The user's preferences are set up before using the system, but not during the crisis management process.

4.3.3 Light Strategy

The light strategy is less attracting. It often contains only visual modalities and it will be applied when the system recognizes that the user might be experiencing cognitive overload. When cognitive overload occurs, the user might become slow at responding to the newly-presented information. The system will notice that more requests/reports stay in pending state. For instance, a victim report stays in the pending state until the system hears a command to the doctor, addressing this victim. When possible cognitive overload is detected, the system still continues on presenting new-coming information. However, only a few most urgent tasks are presented with the default strategy, the rest will adopt the light strategy. In this way, the user's attention is drawn to only the most urgent issues. When a light-presented task becomes one of the most urgent tasks, its presentation will be refreshed with the default strategy. When the user's cognitive state recovers, all light-presented tasks will be shown in default strategy.

5 CONCLUSION AND FUTURE WORK

A multimodal information assistant system for crisis management is proposed, aiming at preventing tunnel vision - the phenomena of framing bias and confirmation bias in the crisis manager's cognitive processes. The main focus of this paper is the design of the modality planning module, which intends to achieve the effectiveness and efficiency of the presentation. The complete set of modality planning strategies is still in the design stage. Apart from the existing guidelines in the literature, we expect to find more from empirical studies that will be carried out. In these studies, some crisis scenarios will be presented in a limited amount of time. The presentation will be evaluated on two aspects, i.e. the user's level of understanding and cognitive state. The level of understanding can be estimated by interviewing the user after the presentation. The cognitive state during the presentation can be measured in several ways [12], such as by evaluating the performance of a secondary task (performed concurrently with the primary task) or by measuring physiological variables (e.g. heart activity, brain activity, and eye activity).

ACKNOWLEDGEMENTS

This research is part of the Interactive Collaborative Information System (ICIS) Project (http://www.decis.nl/html/icis.html). ICIS is sponsored by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

REFERENCES

[1] Y. Bachvarova, B. van Dijk, and A. Nijholt, 'Towards a unified knowledge-based approach to modality choice', in *Workshop on Multimodal Output Generation (MOG)*, (2007).

[2] N. O. Bernsen, 'Why are analogue graphics and natural language both needed in hci', *Interactive Systems: Design, Specification, and Verification. Focus on Computer Graphics. Springer Verlag*, 235–51, (1995).

[3] N. O. Bernsen, 'Multimodality in language and speech systems-from theory to design support tool', *Multimodality in Language and Speech Systems*, (2002).

[4] C. Elting, J. Zwickel, and R. Malaka, 'What are multimodalities made of? modeling output in a multimodal dialogue system', in *International Conference on Intelligent User Interfaces IUI*, pp. 13–16, (2002).

[5] S. K. Feiner and K. R. McKeown, 'Automating the generation of coordinated multimedia explanations', *IEEE Computer*, **24**, 33–41, (1991).

[6] G. P. Hodgkinson, N. J. Bown, A. J. Maule, K. W. Glaister, and A. D. Pearman, 'Breaking the frame: An analysis of strategic cognition and decision making under uncertainty', *Strategic management journal*, **20**(10), 977–985, (1999).

[7] D. Kirsh, 'A few thoughts on cognitive overload', *Intellectica*, **1**(30), 19–51, (2000).

[8] O. Lerbinger, *The Crisis Manager: Facing Risk and Responsibility*, Lawrence Erlbaum Associates, 1997.

[9] S. Y. Mousavi, R. Low, and J. Sweller, 'Reducing cognitive load by mixing auditory and visual presentation modes', *Journal of Educational Psychology*, **87**(2), 319–34, (1995).

[10] R. S. Nickerson, 'Confirmation bias: A ubiquitous phenomenon in many guises', *Review of General Psychology*, **2**(2), 175–220, (1998).

[11] J. F. Nunamaker Jr, E. S. Weber, C. A. P. Smith, and M. Chen, 'Crisis planning systems: tools for intelligent action', *System Sciences, 1988. Vol. III. Decision Support and Knowledge Based Systems Track, Proceedings of the Twenty-First Annual Hawaii International Conference on*, **3**, (1988).

[12] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, 'Cognitive load measurement as a means to advance cognitive load theory', *Cognitive Load Theory: A Special Issue of Educational Psychologist*, (2003).

[13] A. Paivio, *Mental Representations: A Dual Coding Approach*, Oxford University Press, USA, 1986.

[14] A. G. Sutcliffe, S. Kurniawan, and J. E. Shin, 'A method and advisor tool for multimedia user interface design', *International Journal of Human-Computer Studies*, **64**(4), 375–392, (2006).

[15] M. Turoff, M. Chumer, B. Van de Walle, and X. Yao, 'The design of a dynamic emergency response management information system (dermis)', *Journal of Information Technology Theory and Application*, **5**(4), 1–35, (2004).

[16] W. Wahlster, E. Andre, W. Finkler, H. J. Profitlich, and T. Rist, 'Plan-based integration of natural language and graphics generation', *Artificial Intelligence*, **63**(1-2), 387–427, (1993).

[17] W. Wahlster, N. Reithinger, and A. Blocher, 'Smartkom: Multimodal communication with a life-like character', *Seventh European Conference on Speech Communication and Technology*, (2001).

[18] G. Wright and P. Goodwin, 'Eliminating a framing bias by using simple instructions to'think harder' and respondents with managerial experience: Comment on 'breaking the frame'', *Strategic management journal*, **23**(11), 1059–1067, (2002).

How Do I Address You? Modelling addressing behaviour based on an analysis of multi-modal corpora of conversational discourse

Rieks op den Akker and **Mariët Theune** [1]

Abstract. Addressing is a special kind of *referring* and thus principles of multi-modal referring expression generation will also be basic for generation of address terms and addressing gestures for conversational agents. Addressing is a *special* kind of referring because of the different (second person instead of object) role that the referent has in the interaction. Based on an analysis of addressing behaviour in multi-party face-to-face conversations (meetings, TV discussions as well as theater plays), we present outlines of a model for generating multi-modal verbal and non-verbal addressing behaviour for agents in multi-party interactions.

1 INTRODUCTION

According to Gomperz [12], Hymes coined the notion of *communicative competence* to suggest that linguists, concerned as they are with communication in human groups, need to go beyond mere description of language usage patterns [15]. Gomperz then defines *communicative competence* as "the knowledge of linguistic and related communicative conventions that speakers must have to initiate and sustain conversational involvement" [12], p.326. In this study we consider the communicative competence of *addressing*, how partners in various conversational settings address each other. Addressing behaviour, i.e. behaviour that speakers show in order to make clear to others present who they address their talk to, is strongly related to, but to be distinguished from behaviour that speakers show to make clear who may or will continue talking; turn-taking is yet an other aspect of interactive behaviour. Speakers have the obligation to make clear to their listeners who they address their talk, their question for example, to and who they expect an answer from. Addressing is a special kind of *referring* and thus principles of multi-modal referring expression generation will also be basic for generation of address terms and addressing gestures for conversational agents. Addressing is a *special* kind of referring because of the different (second person instead of object) role that the referent has in the interaction.

Our analysis of addressing behaviour is partly based on the AMI meeting corpus, which provides audio and video recordings as well as handmade speech transcripts of four participants face-to-face meeting conversations [24]. But we also look at other conversational settings, such as TV discussions and theater plays. We will sketch outlines of a module for generating multi-modal addressing behaviour and addressing expressions in multi-party conversational settings.

When in a particular situation a speaker says *"Enriquez, how do you spell your name?"*, then the one who is requested to spell his name is the person *addressed by* the speaker when he performs the request. With the use of the pronouns 'you' and 'your', and the address term 'Enriquez', the speaker refers to the one he is talking to. In face to face conversations the speaker's selection of addressee is also embodied in speaker's gaze: he looks at the person he addresses. There are a number of different *categories* here:

1. the person who is expected to answer the question or to take up the request.
2. the referent of 'you',
3. the person gazed at by the speaker, and
4. the one the talk is directed to.

In a 'normal situation' they all co-refer to the same person in that situation. The distinction between 1. and 4. corresponds to Levinson's distinction between "the one the message is intended for" versus "the one the message is directed to" [23] (for a discussion of Levinson's classification of recipient roles we refer to Jovanovic' thesis [18]). In cases of *indirect* addressing in the sense of Clark and Carlson [6], these two are not the same person.

A conversational situation has a participation frame, a division of parties present in the following categories or modes of participation: speaker(s), addressee(s), co-participants and overhearers. They all stand in a different relation towards the speaker. When we talk we deal with the current participation frame – i.e. the current assignment of parties to these roles or categories – and we also have the opportunity to redesign the frame. *Recipient design* refers to "a multitude of respects in which the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are the co-participants" [26]. *Audience design* in the sense of [6] and [11] covers overhearers as well as co-participants.

In a quest for conversational rules for addressing, in this paper we discuss the main literature on addressing and present both qualitative and quantitative analyses of addressing in some multimodal corpora of multi-party, face-to-face conversations. Finally, we present a first sketch of a model for the generation of multimodal addressing behaviour.

2 ADDRESSING BEHAVIOUR

Linguists and conversational analysts describe *addressees* as those listeners who are expected (by speakers) to take up the proposed joint project [7]. The turn-taking theory of Sacks et al. has a rule saying

[1] University of Twente, the Netherlands, email: infrieks@ewi.utwente.nl, m.theune@ewi.utwente.nl

that speakers may select the next speaker by inviting them; if not, others do self-selection as next speaker [26]. Goffman's definition of addressee as "the one the speaker selects as the one he expects a response from, more than from other listeners" [11] refers to this *next speaker selecting notion of addressing*. From the point of view of addressee identification, the counterpart of generating addressing behaviour, one of the *observable* things indicative for who is being addressed is *who is talking next*. We may expect that if the speaker addresses a speech act to a *single* addressee that this addressed person will speak next. This motivates the use of the category of *next speaker* as feature (indicator) in systems for automatic addressee identification [19]. We come back to addressing and initiative when we discuss multi-party conversations in sections 3.2 and 5.

Lerner [22] distinguishes explicit methods of addressing, which are speakers' gaze and naming (the use of vocatives, address terms), from "tacit forms of addressing that call on the innumerable context-specific particulars of circumstance, content, and composition to select a next speaker" ([22], p.177). Lerner examines the context-sensitivity of addressing practices employed by a current speaker to make evident the selection of a next speaker. His discussions are restricted to those turns-at-talk that implement *sequence-initiating actions*, the first parts of adjacency pairs.

Addressing by *gaze* works only if the addressee notices the speaker's gaze and picks up the signal as a sign of addressing; moreover both have to believe that they share this common belief. Mutual gaze between speaker and addressee is basic for grounding in face-to-face conversations. Only mutual gaze between A and B is the most reliable way to establish the belief of A a) that A sees B, b) that A sees that B sees that A sees B, and c) that both share this belief. Accompanied with other messages sent by A (an utterance of a question for example, or a gesture) this may lead B to believe that A's gazing at her means that B is being addressed by A. By looking at B, A checks whether B is ready to receive his message. Others also have to understand that they are *not* selected as next speaker. Thus, "gaze is an explicit form of addressing, but its success is contingent on the separate gazing practices of co-participants" ([22], p.180).

According to Lerner there is one form of address that always has the property of indicating addressing, but that does not itself uniquely specify *who* is being addressed: the *recipient reference term* 'you':

> *The use of 'you' as a form of person reference separates the action of addressing a recipient from the designation of just who is being addressed. In interactional terms, then, 'you' might be termed a recipient indicator, but not a recipient designator. As such, it might be thought of as an incomplete form of address.* [22], p.182.

The speaker will try to complete the addressing act by gazing at the selected recipient, a completion that needs the joint gazing of the intended recipient, and of others present as well, so that they know they are not selected. Thus, for addressing to be complete it requires the joint actions of all participants. This is illustrated by the following fragment from the AMI meeting corpus [24].[2] In the first utterance by speaker P3 *'you"* is not supported by disambiguating gaze; both conversation participants P2 and P0 are gazed at by P3. P2 feels addressed and responds, but P0 also. P2's response overlaps with the elicitation and he is interrupted by P0. It is as if P2 then recognizes that *not he* but P0 was selected as next speaker. P2 and P0's *"Uh"* may also indicate the confusion in the situation.

> P3>P0: What do you think, is it fancy?
> P2>P3: Uh, it's really
> P0>P3: Uh, I think that fancy, we can say it is fancy.

[22] discusses an example of use of referring 'you' directed to a specific individual in a multi-party conversation, where the addressing does not need the support of speaker's gaze at the intended addressee. In a situation where four people are having dinner together, and everybody knows who has prepared the dinner, and the speaker assumes that everybody knows that, the speaker asks: *"Did you cook this all the way through?"* ([22], p.192). Here, the content and context are sufficient to determine the identity of the addressee without the need for explicit addressing behaviour.

The most explicit form of addressing is by use of an address term (which may or may not take the form of a name). This is either used in pre-position, in post-position, or in mid-position, as illustrated by the following examples.

> So, *mister money*, what's your opinion according to this remote control?
> What do you think, *Ed*?
> They wake up fast, *Jessie*, if they have to.

In almost all usages of address terms in talk in face-to-face conversations their function is not purely to call the addressee's attention. If it is, the term is used in pre-position, more often than elsewhere, but most often it seems to be used to put more stress on the addressing, maybe to signal the addressing to co-participants, or to express some affective or social relation with the recipient.

3 ADDRESSING AS A FORM OF MULTIMODAL REFERENCE

3.1 'Changing ideas about reference'

In their paper 'Changing ideas about reference' Clark and Bangerter list some common assumptions concerning reference made by David Olson[3] and his contemporaries ([8], p.26):

1. Referring is an *autonomous act*. It consists of planning and producing a referring expression, which speakers do on their own.
2. Referring is a *one-step process*. It consists of the planning and uttering of a referring expression, and nothing more.
3. Referring is *addressee-blind*. It depends on the context – the set of alternatives in the situation – but doesn't otherwise depend on beliefs of the addressees.
4. Referring is *ahistorical*. It doesn't take account of past relations between speakers and their addressees.
5. The referent belongs to a *specifiable set of alternatives*.

They then proceed to show that each of these assumptions is wrong, arguing that referring (in conversation) is a *cooperative* and *interactive* process involving among other things the establishment of common ground between dialogue participants and the formation of *conceptual pacts* [4], object descriptions which participants jointly decide to use throughout the conversation.

Remarkably, the assumptions listed by Clark and Bangerter still apply to most current approaches to the generation of referring expressions (GRE), which are for the most part direct descendants of Dale and Reiter's classic Incremental Algorithm [9] (see [28] for an

[2] In this and following examples, A $>$ B indicates that A addresses B.

[3] D.R. Olson, Language and thought: Aspects of a cognitive theory of semantics. Psychological Review 77(4):257-73.

overview). These approaches focus on the generation of a definite description of a target object in a visual scene, by selecting properties that apply only to the target referent and not to any of its alternatives (the 'distractors').

So far, GRE has been mainly investigated in the context of *text* generation, where the user is a "distant reader" ([8], p.36) and not a conversation partner. As a consequence, most approaches to GRE do not take conversational factors into account. A few recent exceptions are [17] and [14], who applied GRE in a dialogue setting and took the notion of *conceptual pact* [4] into account by reusing features from earlier references to the same object. Other moves towards a more situated approach to GRE are those by [20] and [28], who developed algorithms for the generation of multimodal referring expressions that combine a verbal description with a pointing gesture, taking the relative locations of speaker, target object and distractors into account to achieve an optimal combination of verbal and nonverbal modalities.

In spite of this recent trend towards more dialogue-oriented, situated approaches to GRE, overall still not much attention is paid to social and conversational aspects of reference. However, for the generation of multimodal *addressing* references, such aspects cannot be ignored, as shown in the following section.

3.2 Addressing as an interactive process

Addressing is a special kind of *multimodal reference*, and here we show that the five assumptions listed by Clark and Bangerter [8] hold even less for addressing than for reference in general. Our examples come from a Dutch TV programme ("B&W", 2002) in which a discussion leader (W) discusses with his guests whether *foie gras* (goose liver) should be banned from restaurant menus for reasons of animal cruelty. Let us go through the assumptions one by one and show that their opposite holds true.

1. *Addressing is **not** an autonomous act.* As we have seen in section 2, for addressing to be complete it requires the joint actions of all participants. The addressee has to pick up on the fact that she is being addressed, and give evidence of this by for example returning the speaker's gaze, nodding, back-channeling, or answering the question that was addressed to her. At the same time, the other participants in the conversation need to know they are not being addressed. They can give evidence of this by for example leaning back, away from the speaker. As Figure 1a shows, body posture and gestures can clearly show that speaker and addressee are 'tuned in' on each other, whereas the co-participants are literally keeping more distance.

2. *Addressing is **not** a one-step process*. It consists of at least two phases: the addressing act by the speaker and the acknowledgement by the addressee. However, when addressing is not immediately successful, additional phases may be involved. In the following example from our TV discussion, W initially relies only on gaze and the content of the dialogue act when addressing B, and then turns to a more explicit form of addressing after this initial attempt fails.

W>B: (gazing at B) Als u nou dat van de kaart haalt, dan is er toch nog voldoende lekkers te eten bij u
If you take this off the menu, won't there still be plenty of nice things to eat at your place
M>W: Nee maar het alternatief is er niet, dat is nou wat ik
No but there is no alternative, that is just what I
W>M: Zei ik tegen meneer B
I said to Mr. B

3. *Referring is **not** addressee-blind.* Addressing involves not only the person(s) being addressed but also those who are not being addressed, and the beliefs of all these participants have to be taken into account. For example, in Lerner's dinner scene discussed in section 2, the speaker asking *"Did you cook this all the way through?"* assumes that A was the cook, and he assumes that everybody knows that A was the cook. So he expects that he can safely address A using *"you"*, without supporting gaze. In other situations, where the participants' common ground is limited, more explicit references may be necessary. In our TV discussion, for example, each time when W addresses one of the discussion participants for the first time, he does this in a particular fashion. He leans strongly toward the addressee (see Figure 1b) and explicitly mentions the addressee's name, followed by a statement or question about his or her identity:

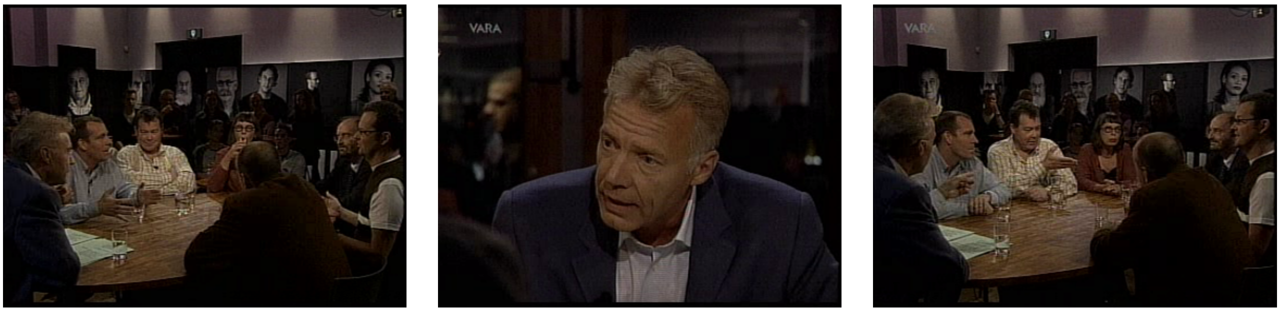
W >B: Mijnheer B, u bent van B Restaurant in O
Mr B, you are from B Restaurant in O
...
W >M: Mijnheer M, u importeert het?
Mr. M, you import it? (where it = foie gras)

W's nonverbal behaviour makes it very clear who is being addressed, also for the co-participants who might not know the addressee's name. The accompanying verbal reference has a double function: besides addressing, it also serves to inform the co-participants and overhearers (the TV audience) of the addressee's name and occupation. This information becomes part of the common ground and may be used for later reference. (The first example of this section showed an unsuccessful attempt by W to use the shared knowledge that B owns a restaurant when implicitly addressing B.)

4. *Addressing is **not** ahistorical.* As shown above, speakers make use of past information from the conversation when they are addressing. Also (as we will discuss in more detail in section 5) dialogue history plays an important role, in that the most likely addressee of the second half of an adjacency pair is the previous speaker.

Our TV discussion contains many sequences where two speakers hold the floor in the conversation, arguing with each other and taking turns in addressing each other. In these bilateral exchanges, the speakers tend to simply refer to their addressees as 'you', relying on nonverbal cues and their shared dialogue history to disambiguate this term. At some points, however, discussion leader W breaks in and assigns the turn to a non-salient participant, using the addressee's name (e.g. *"Is this how it should be done, Mr. C?"*), sometimes supported by a gesture. This is illustrated in Figure 1c, where W (on the left) explicitly assigns the next turn to M (in the middle) by pointing at him and addressing him by name. (At the same time M is pointing at D (on the right), trying to address him while D is talking to someone else.)

5. *The referent does not necessarily belong to a specifiable set of alternatives.* At first sight it seems obvious that all participants in a conversation are potential addressees. However, it is not always clear exactly who the participants are. For example, in our TV discussion the audience does not actively participate in the discussion, but they can be addressed nevertheless: at the start of the programme, discussion leader W. welcomes the viewers using an explicit address term *"Ladies and gentlemen*" and at the end he says goodbye to them. (Note that this illustrates that addressees are not necessarily the next speakers.) The borderline between addressees and non-addressees can be vague. As discussed above, speakers not only take the addressee but also the co-participants into account when designing their address terms. We also sometimes see cases of indirect addressing behaviour [6], as in the following example. Here, W interrupts an argument between animal activist D and culinary journalist S by asking D a question. D answers the question, thus 'technically' addressing



(a) Two participants in discussion (b) Discussion leader W while addressing (c) Addressing with pointing gestures

Figure 1. Screen-shots from the B&W discussion programme

W, but he keeps gazing at S while he continues accusing her:

D>S: Daar heeft u uw excuses voor gemaakt aan ons op een halve manier, eh
You apologized to us for that in a half-hearted way, eh
W>D: Maar wat heeft dat met die foie gras te maken
But what does that have to do with foie gras?
D>W: (gazing at S) Nou kijk, uhm, die mevrouw die die is niet een onafhankelijk journalist, zij is meer een promotor van dierenmishandeling als ik het zo hoor
Well, you see, uhm, that lady she she isn't an independent journalist, she's more of a proponent of animal cruelty from what I hear

In his second utterance, D refers to S in the third person, so she cannot be the 'official' addressee [22]. However, his fixed gaze on her shows that she is really the intended recipient of his message.

4 SOCIAL, AFFECTIVE AND COGNITIVE DIMENSIONS OF ADDRESSING

Addressing is an aspect of any form of communication and as such addressing has cognitive, social as well as affective implications and its realisation depends on the capacities of sender, receiver and the communication channels.

Allan Bell [2] points at the impact that the audience has on language style. Style is what an individual speaker does with a language in relation to other people. According to Bell speakers design their style primarily for and in response to their audience. Within a single language audience design does not only refer to variations in speech, it also involves choice of personal pronouns or address terms, but audience design also applies to shift of codes, dialects or languages, in bilingual situations.

4.1 Affective impact of using address terms

In Marsha Norman's play *'night, Mother* the mother has several address terms for her daughter Jessie, who in turn has only a couple of standard ones for her mother. According to Bernardy [3] Mama's use of 'Jessie' mirrors her changing mental state. When overwhelmed or dismayed and unable to articulate a more complete reply, Mama cries out the single word *"Jessie!"* at various points throughout the play.

Since *'night, Mother* has only two players, obviously the primary function of the usage of address terms by Jessie and the mother is not to select who is being addressed, as is suggested by the choice of the term *"address term"*. Most prevalent and decisive for the choice of the terms used is the *affective* value that the usage of the address terms carries and expresses. We are inclined to say that the usage of these terms of endearment in these types of affective conversations have nothing to do with addressing *in the sense of speaker's selecting the intended receiver of his speech act*.

The choice of address terms also plays an important role in expressing different forms of *politeness* [5]. Many languages make a distinction between non-honorific (T) and honorific (V) pronouns[4] where the latter are used to show respect if there is a social distance or difference in power between the speaker and the addressee. On the other hand, when addressing a non-familiar alter, T pronouns can be used to claim solidarity, similar to the use of in-group address terms such as *mate, buddy, pal, honey, dear* ([5], p.107). Such an informal way of addressing can be used as a strategy to save the addressee's positive face, i.e., his need of approval by others. In strategies aimed at saving negative face, a person's need for autonomy, we see the reverse: speakers tend to use honorific address terms in particular when carrying out face-threatening acts such as a requests. E.g., *Excuse me, sir, but would you mind...?* In contrast, the use of honorifics in a non-threatening utterance is much less natural: *Goodness, sir, that sunset is amazing* ([5], p.183).

Given the important affective and social value of address terms usage we expect that careful selection of address terms will improve the believability of synthetic conversational characters [29] and make them appear more socially intelligent [1].

4.2 Channels of communication

Addressing behaviour depends on the conversational setting a part of which is the available "hardware" for communication, the capacities of the channels, auditive, visual, and the capabilities of acting and perception. The mechanics of a conversation, and thus the mechanics of addressing behaviour, depends very much on the communication channels and information made available to the partners. Addressees, over-hearers, remote meetings participants, differ in the ways they participate and interact with other participants in a conversation. It is hard for outside observers to tell who is being addressed in a face to face conversation if they don't share with the other participants the communication channels or the knowledge about the conversation.

Gupta et al. [13] present experiments into the resolution of 'you' in multi-party face-to-face dialogue. They distinguish between *generic*

[4] From the French 'tu' (T) and 'vous' (V).

and *referential* uses of 'you'; and they try to classify the referential uses automatically by identifying the referred-to addressee(s): either one of the participants, or the group. Their results were obtained without the use of visual information; obviously, that makes a hard job because the people in conversation did not have to cope with partners that could not see them. As expected their results on face-to-face meetings were much worse than those they had obtained earlier on the Switchboard corpus, where the parties only had audio communication.

Announcing the name of the addressee helps blind people know that they are being spoken to. (See Irrizary's analysis of the Spanish play *En la ardiente oscuridad* by Buero Vallejo, which has only two sighted characters [16].) Similarly, in discussions for broadcast radio, the discussion leader who feels responsible for informing the audience about who is being addressed, will use more names as address terms than is strictly required for informing his intended addressee that she is addressed.

5 ADDRESSING IN SMALL GROUP MEETINGS

For this study we analysed a number of hand annotated conversations from the AMI meeting corpus [24]. In the scenario-based AMI meetings, design project groups of four players have the task to design a new remote TV control. Group members have roles: project manager (PM), industrial designer (ID), user interface design (UD), and marketing expert (ME). Every group has four meetings (20-30 min. each), dedicated to a subtask. In the first meeting, partners introduce themselves and they use the white board to draw their favorite animal, in the last meeting a clay prototype is presented and evaluated. Conversations in brainstorm sessions, presentations using slide show and laptop, discussions that should lead to a group decision about some detail of the design of the remote control, are embedded in non-conversational activities. Most of the time during meetings partners sit at a square table.

The meetings were recorded in a meeting room stuffed with audio and video recording devices, so that close facial views and overview video, as well as high quality audio is available. Speech was transcribed manually, and words were time aligned. The corpus has several layers of annotation and is easily extendible with new layers. The *dialogue act* (DA) layer segments speaker turns into dialogue act segments, on top of the word layer, and they are labeled with one of 15 dialogue act type labels, following an annotation procedure. A part of the corpus is also annotated with addressee information: DAs are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*). Sub-group addressing hardly occurs and was not annotated. Another layer contains *focus of attention* information (derived from head, body and gaze observations), so that for each partner, at any time instant, it is known who she is looking at; table, white board, or some other participant. In our search for patterns of addressing behaviour, that could inform better models for addressing we noticed a number of interesting fragments. We focused on those dialogue acts in which the speaker takes initiative and tries to elicit some response from some partner that he addresses. It is to be expected that the addressed person will take the turn and respond to the elicit act. The exceptional cases we encountered point at interesting addressing phenomena in small group face-to-face conversations. Recall however that a speaker addressing an individual listener does not necessarily imply that she yields turn to the addressee. For example in the following fragment B makes a proposal to A and C in (0). A addresses her objection to B in (1,3).

(0) B>A,C: I think we should ...
(1) A>B: Okay, but as Carla just said,
(2) A>C: *correct me if I'm wrong*,
(3) A>B: that is too costly for us ...

The embedded invitation to correct her is, however, addressed to C. But A does not give away the floor, instead she continues her objection towards B's proposal. While uttering (2), A gazes at C shortly to notice her non-verbal response. Speakers also invite listeners to give feedback only by briefly gazing at the listeners, especially if they refer to them in the course of their arguing, as A does to C in (1) above. It becomes clear that addressing comes in various flavors, and that conversational acts like asking for feedback can be done non-verbally as well as verbally. The annotators of the meetings were asked to tell if the speaker addressed his dialogue act in particular to some individual in the sense that it was more for her than for others present. Answers to questions are in a sense always addressed to the one who asked the question, and we see that answerers indeed gaze at the previous speaker, but if the issue is a group concern the inform act is addressed to the group.

5.1 Addressing in initiating acts

Addressing in initiating acts is more explicit than addressing in responsive acts, in as far as the speaker who takes initiative in an exchange also has to make clear whom he selects as addressee(s). Dialogue act sequences (we forget about speaker overlap, and multiple floors for a while) have a structure that reflects the fact that partners are interacting, they temporarily participate in changing participation frames around a shared task: to resolve some issue introduced by one of them. Based on conversational analysis we may indeed expect that *the structure of the dialogue gives the most indicative cues to addressee: forward-looking dialogue acts are likely to influence the addressee to speak next, while backward-looking acts might address a recent speaker*. A classical way to model the interaction is in terms of *adjacency pairs*, and Galley et al. [10] used the dialogue structure present in these smallest units of interaction as indicative for addressees: the speaker of the *a-part* of the pair would likely be the addressee of the *b-part* of the pair, and the addressee of the *a-part* would likely be the speaker of the *b-part*. In the one dimensional DA schema that is used in the AMI meeting corpus there is no clear distinction between Backward Looking and Forward Looking DA classes. However, the *elicit types* are primarily FL types of DAs. Typical BL DA types are backchannels, comments about understanding and assessments.

The total number of DAs in the addressee annotated part of the corpus is 9987, of which 6590 are real DAs (i.e. excluding stalls, fragments, backchannels, which do not have an addressee label). Of these, 2743 are addressed to some individual (*I-addressed*); the others are addressed to the Group (*G-addressed*, 3104) or the addressee label is Unknown (which means that the annotator could not tell). In 1739 (63%) of the 2743 *I-addressed* dialogue acts, the addressed person is the next speaker. In our corpus of 652 elicit acts, 236 are *G-addressed*, and 387 are *I-addressed*. Elicits are more *I-addressed* than other DAs. *I-addressed* elicits contain more referring *"you"* than G-addressed elicit acts[5]. In 302 cases (78% the addressee is the next speaker. Thus, FL DAs that are *I-addressed* are more selective for next speaker than *I-addressed* DAs in general, as we expected. When we looked at the instances in the other 22% of the cases, in

[5] If we use "more than", we always mean "significantly more than"; in this case ($\chi^2(df = 1) = 30.66, p < 0.0005$).

which the next speaker did not coincide with the addressed person, we found some interesting addressing phenomena.

Some activities center around one specific actor; a presenter, or someone drawing on the white board, or someone holding the clay prototype that is being discussed. If someone says *"is it heavy?"* it is clear who is being addressed. Or, when a person is drawing his favorite animal on the white board and the speaker makes a guess *"a horse?"*, asking the artist to reveal his secret animal. Actors of activities that are in focus are more salient than others for addressing. Moreover, these actors can tacitly be addressed by others when they comment on, or ask about, the action they perform.

Sometimes, the speaker uses the wrong name or a wrong attribute for his addressee. In such cases an unaddressed listener might feel more entitled to answer the question than the addressee. The speaker uses the referent term 'you' and gazes at P2 to make clear whose identity he is after. But the real marketing guy is called by the attributive use of *"marketing guy"*.

P3>P2: You are *the marketing guy* ? Or
P0>P3: I'm marketing .

In the following fragment it is unclear who is being addressed: a non-addressed attendant tries to answer but is interrupted by the addressee. The speaker indeed gazes at P1 at the end of his question which could easily be taken as if he has selected P1 to speak next.

P2>P0: so how many units should we sell to have a
P1: Well . Uh
P0>Group: Well each unit is is sell uh twenty five Euros .

Another example of unclear addressing: a *you*-utterance without speaker gaze to select the designated addressee

P0>P3: D D Is is there anything you want to add ?
P2>Group: Is there any fruit that is spongy ?

We have seen a number of cases that make clear that proper addressing uses beliefs that speakers have about saliency of persons because of their role in the activity that the group is busy with. Successful addressing is constrained by the general conditions about sharing beliefs about who are salient and who are gazed at as a signal of addressing.

5.2 Reliability

Are the judgments about *"what happens in these conversations"*, about who is being addressed, purely a matter of personal taste? The dialogue annotations and addressee annotations of the part of the AMI corpus we used contains parts made by three different annotators, who followed a documented annotation procedure, using dedicated annotation tools that allows listening to audio, reading the hand made transcripts, as well as looking at video recordings, showing front views of the individual participants as well as an overview of the meeting room. One meeting was annotated by all three annotators. Table 1 shows for each pair of annotators involved Krippendorff alpha values for inter-annotator agreement [21]. For the group of annotators alpha is 0.35 for addressing. The statistics are based on comparing DA-labels of completely agreed DA-segments. Most confusions in the addressing labeling are between *I-addressed* and *G-addressed*, between I and U and between G and U; there is hardly any confusion between annotators about who is addressed when they agree that the DA is *I-addressed* (see also [19]). The table shows that annotators agree more on the addressing of elicit acts than on DAs in general. For the subset of elicit acts we see hardly any U labels used, and when annotators agree that an elicit is *I-addressed* (which happens in 50-80% of the agreed elicit acts), they agree on who is addressed, without exception. Annotators agree more on the addressee of a DA in situations where the speaker clearly gazes at the addressed person. We did not find any indication that annotators systematically confused speaker's gaze with addressing. Addressing is a complex phenomenon and we believe that the low agreement between addressee annotations is due to this complexity.

Table 1. Krippendorff alpha values (and numbers of agreed DA segments) for the three pairs of annotators; for addressing, addressing of elicit acts, dialogue acts (all 15 DA classes), and elicit vs non-elicit acts.

pair	adr	adr-eli	da	da-eli
a-b	0.50(412)	0.67(31)	0.62(756)	0.69
a-c	0.37(344)	0.58(32)	0.58(735)	0.64
b-c	0.33(430)	0.62(53)	0.55(795)	0.80

Focus of attention annotation was done with high agreement, so we may conclude that the annotated data allows a good starting point for research of multi-modal conversational behaviour involved in addressing of eliciting acts and the responsive behaviour that follows in multi-party face to face conversations in general.

6 SYNTHESIS: GENERATING ADDRESSING BEHAVIOUR

We now come to the synthetical part of our project. The four aspects to be considered in the generation of referring expressions (REs) are according to Ielka van der Sluis ([28], p.21-22):

1. costs of the cooperative effort of both speaker and listener
2. accessibility of the object in its context
3. salience of objects
4. responsibility of the speaker for the effectiveness of his choice for the way of referring

All these are relevant in generating multi-modal expressions used in addressing as well. In fact the problem of finding an RE in a conversational setting (see our analyses in section 3 based on the five points made by Clark and Bangerter) is a special case of the more general problem of (language use in) communication. In [25] Prashant Parikh developes a model of communication using frameworks of situation theory and game theory in which he gives an "approximate" set of necessary and sufficient conditions for communication. The general problem of communication is "to find the necessary and sufficient conditions (which involves finding the inferential mechanism) for A to communicate some proposition p to B" ([25], p.475). We consider the *addressing problem* (AP) as an instance of this general communication problem. In a given *conversational situation* the *addressing problem* (AP) *for agent A who wants to address agent B* is to find REs so that if he *uses* them in the given situation, the effect is that he addresses B, or, more formally:

> AP is: to find REs $< \phi_i >$ such that if A uses $< \phi_i >$ it has the effect that (A and B share the belief that) A addresses B by means of $< \phi_i >$.

$ADR(A, \phi, B)$ (A addresses B by using RE ϕ) is established when (see the conditions for usage of referentials by Schegloff [27]):

$$Bel_A(Ref(A, \phi, B))$$

$$Bel_A(Bel_B(use_A(\phi) \Rightarrow Ref(A, \phi, B)))$$

$$Bel_A(SB_{<A,B>}(ADR(A, \phi, B)))$$

where $Bel_A(p)$ means: A believes that p; $Ref(A, \phi, B)$: A refers to B by ϕ; $SB_{<A,B>}(p)$: A and B have shared belief p; $use_A(\phi)$: the *act* of using RE ϕ by A.
Parikh did not take co-participants into account, but addressing only becomes a problem in case more than two partners are present. We have seen that the speaker also has to make clear to non-addressed participants who are the ones he has selected as addressee, so that they also know they are not addressed. Formally, if C is a non-addressed side-participant, the additional intention of A is that:

$$Bel_A(SB_{<A,C>}(ADR(A, \phi, B)))$$

If we restrict to single addressing, if C knows B is addressed by A then C knows that he is not addressed. As we have seen in our corpus analyses, the speaker will take this final goal into account when selecting his selection of REs to refer to his addressee(s). For example, a speaker may address by using the name of the addressee and simultaneously point at his addressee. The pointing gesture can be redundant for the addressee but not for side-participants that do not know the name of the addressed person. The intended effect is then that non-addressed partners know who the person is that is being addressed.

Corpus analysis shows that what a speaker has to do when he wants to address someone is to check if there is a communication line open. If not he has to call the intended addressee. In case the speaker does not know the name of the intended addressee, a standard method for GRE can be used to find a referring expression for addressing. For example as in the referential installment: "*The lady in the red shirt in the back*. Could you please close the door?" An RE isn't required when the conversational situation and contents of the dialogue act make it clear who is being addressed; as was the case in Lerner's 'cooking dinner' example.

The AP concerns the selection of the multi-modal means for referring (the REs) that the speaker will and can (given the communication channels available) use for this. The REs are either verbal, or non-verbal. The set of verbal REs contains special types: *proper names* and (politeness) forms of *"you"*. The latter we denote by you_T**,** you_V or simply as **you**. Proper names (PN) are special in that in any situation if an agent A has PN ϕ, when someone uses ϕ then A will be *called*. For addressing two special types of non-verbal REs are deictic pointing at (**pat**) and gaze at (**gat**). We don't go into the details of the (conditions of) usages of the various types of REs, but to recall that visual and auditive *distance* between speaker and the others is a factor that plays in selecting one of them (impact of perceptual capabilities of parties). We use $< \phi_i >$ for a set of different REs (for example a PN and **gat**), although the temporal order of use of these REs can have effects (at least side effects, i.e. not directly on addressing; recall the affective impact of using PNs in certain positions or in repetition).

A few remarks are in order. In discussions about AP, we should distinguish the *selection* of the best REs for agent A to address B in a given situation from the actual *usage* of the selected REs, and these form the beliefs that A has about the effects of using an RE. Using a name means making a sound or writing the name down in some situation. It is an action that is observable by others and thus others will make inferences about the simple fact that they observe an agent act.

If agent A is involved in an exchange with agent B then A need not use REs to address B if B may suppose that his speech act is addressed to B. For example, if A responds in a face-to-face conversation to a question asked by B, then the situation is such that B is the preferred addressee, or most salient. This motivates a special RE $<>$ (null). If an agent A chooses $<>$ to address then either *no one* is addressed in particular, *or the one who is most salient for* A to be addressed is addressed by A. Every agent has a set of beliefs about the order of saliency of other partners. The selection of REs is determined by this (shared belief of) salience order.

There will be many REs (many verbal REs in particular) that may solve AP for A. The choice is constrained by preferences related to *costs*, and *affective values* of using some RE ϕ:

- A prefers RE ϕ for addressing B above ϕ' when $cost(\phi)$ is less than $cost(\phi')$.
- A prefers RE ϕ for addressing B above ϕ' when ϕ is a better term for the affective relation of A and B than ϕ'.

We take affective values here broadly and also include social values such as politeness. Affective values constrain the selection of address terms, but affective address terms like *"friend"* can also have distinguishing value and help to rule out alternatives from being addressed.

Properties of the *cost* function are: $cost(< \phi_i >) = \sum cost(\phi_i)$, $cost(<>) = 0$. Intuitively, it holds that $cost(\textbf{gat}) > 0$, and for all REs ϕ: $cost(\text{gat}) < cost(\phi)$. (or, gazing at is cheap).

Pointing at the addressee is more precise than gazing at and thus supports the addressing act when the target is at a certain distance from the speaker and there are alternatives (i.e. non-addressed participants) in the neighborhood of the target. The cost of pointing at will be higher the larger the distance to the addressee. The cost of pointing will be larger the shorter the distance between the target and the alternatives. Let l be the line through speaker and addressee; let l' be the line through speaker and the closest non-addressed person. The cost of pointing at is higher the smaller the angle between l and l'. [6] Obviously, there is a trade-off between *saliency* of the intended addressee and the costs of the RE that should do the job: the higher the saliency, the lower the energy that needs to be put in the addressing act. We have seen that saliency of partner depends on the (beliefs the agent has about) the current participation frame; this is, regarding AP, the essential aspect of the current conversational situation.

The proposed approach to solving AP meets all the five points mentioned by [8]. There is no guarantee that the REs selected by the speaker will effectively address the intended recipient, because assumptions that the speaker makes, for example about the state of mind of the recipient, may not hold. Thus agents use repair strategies. Indeed, a solution to AP should not be seen as a one step process. Conversations are *joint projects* [7] of multiple agents, and a solution of AP will thus take the form of an *addressing strategy* that involves the joint work of multiple agents, be it that the speaker takes the initiative in this "project" and he will mostly choose an RE that he

[6] Note that in [28], p.88, the cost of pointing depends on the distance between speaker and target as well as on the size of the target. In the special case of addressing size is hardly variable, but the distance between target and non-targeted alternatives does matter.

believes to be a one step solution to the problem.[7]

The approach outlined here is also valid in situations where for example there is only audio communication. If the speaker believes that the listener can't see him he will conclude that gazing at and gestures are not effective for addressing and these acts will not be chosen. In that case the speaker will rely on vocal means or he believes that there is already a line of communication with the intended addressee so that an empty RE is sufficient for addressing his message.

7 CONCLUSION AND FURTHER WORK

Addressing has up to now not gained much attention in research devoted to the generation of multi-modal referring expressions in multi-agent systems. Addressing involves a special way of referring to partners in conversations, and the social and affective dimension is so prevalent that it strongly determines the choice of address terms and other referring expressions used in addressing.

We presented some outlines of a method for generating addressing behaviour, based on an analysis of a variety of natural multi-party conversations. We have made a first step towards extending existing methods for generating multi-modal referring expressions as in [28] so that they can be used for addressing in multi-party conversations. We formulated AP initially as a problem of a single agent, the agent in the speaker role. But we should take it as a *joint* problem, a problem that can only be solved by joint acts; by speakers as well as listeners. That an act is potentially a joint act is something however, that can only become clear after it has already been initiated, performed by an agent as an attempt, a proposal to interact, as well as taken up by the listeners, in particular the intended addressee(s). The elaboration of this is the next step in our joint project.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper. We are grateful to Dirk Heylen for providing the video and transcripts of the B&W TV discussion. "This work is supported by the European IST Programme Project FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein."

REFERENCES

[1] Elisabeth Andre, Matthias Rehm, Wolfgang Minker, and Dirk Buhler, 'Endowing spoken language dialogue systems with emotional intelligence', in *Affective Dialogue Systems*, LNCS 3068, pp. 178–187, (2004).

[2] Allan Bell, 'Language style as audience design', in *Sociolinguistics, a reader and course book*, eds., Nikolas Coupland and Adam Jaworski, Palgrave, (1997).

[3] Maria Lee Bernardy, *Beyond Intuition: Analyzing Marsha Norman's 'night, Mother; with Concordance Data and Empirical Methods.*, Ph.D. dissertation, Iowa State University, 1996.

[4] Susan E. Brennan and Herbert H. Clark, 'Conceptual pacts and lexical choice in conversation', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1482–1493, (1996).

[5] Penelope Brown and Stephen C. Levinson, *Politeness - Some universals in language usage*, Cambridge University Press, 1987.

[6] H. H. Clark and T. B. Carlson, 'Hearers and speech acts', in *Arenas of Language Use*, ed., Herbert H. Clark, 205–247, Chicago: University of Chicago Press and CSLI, (1992).

[7] Herbert H. Clark, *Using language*, Cambridge: Cambridge University Press, 1996.

[8] Herbert H. Clark and Adrian Bangerter, 'Changing ideas about reference', in *Experimental Pragmatics*, eds., Ira A. Noveck and Dan Sperber, 25–49, Palgrave Macmillan, Basingstoke, (2004).

[9] Robert Dale and Ehud Reiter, 'Computational interpretation of the Gricean maxims in the generation of referring expressions', *Cognitive Science*, **19**(2), 233–263, (1995).

[10] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg, 'Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies.', in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, (2004).

[11] Erving Goffman, 'Footing', in *Forms of Talk*, 124–159, Philadelphia: University of Pennsylvania Press, (1981).

[12] John J. Gomperz, 'The linguistic bases of communicative competence', in *Analyzing Discourse: Text and Talk*, ed., Deborah Tannen, 323–334, Georgetown University Press, (1981).

[13] Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky, 'Resolving "you" in multi-party dialog', in *Proceedings of SigDial Workshop of European Chapter of the ACL, Prague*, pp. 1–4, (2007).

[14] Surabhi Gupta and Amanda Stent, 'Automatic evaluation of referring expression generation using corpora', in *Proceedings of the Workshop on Using Corpora for Natural Language Generation (UCNLG)*, Morristown, NJ, USA, (2005). Association for Computational Linguistics.

[15] Dell H. Hymes, *Foundations in Sociolinguistics: An Ethnographic Approach*, Philadelphia: University of Pennsylvania Press, 1974.

[16] Estelle Irizarry, 'Some approaches to computer analysis of dialogue in theater: Buero vallejo's en la ardiente oscuridad', *Journal Computers and the Humanities*, **25**(1), (February 1991).

[17] Pamela W. Jordan and Marilyn A. Walker, 'Learning content selection rules for generating object descriptions in dialogue', *Journal of Artificial Intelligence Research*, **24**, 157–194, (2005).

[18] N. Jovanovic, *To whom it may concern. Addressee identification in face-to-face meetings*, Ph.D. dissertation, University of Twente, Enschede, The Netherlands, March 2007.

[19] N. Jovanovic, R. op den Akker, and A. Nijholt, 'Addressee identification in face-to-face meetings', in *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, (2006).

[20] Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth, 'Deictic object reference in task-oriented dialogue', in *Situated Communication*, eds., G. Rickheit and I. Wachsmuth, 155–207, Mouton de Gruyter, Berlin, (2006).

[21] Klaus Krippendorff, 'Reliability in content analysis: Some common misconceptions and recommendations', *Human Communication Research*, **30(3)**, 411–433, (2004).

[22] Gene H. Lerner, 'Selecting next speaker: The context-sensitive operation of a context-free organization', *Language in Society*, **32**, 177–201, (2003).

[23] S. C. Levinson, 'Putting linguistics on a proper footing: Explorations in goffman's participation framework', in *Erving Goffman: Exploring the interaction order*, eds., P. Drew and A. Wootton, 161–227, Oxford: Polity Press, (1987).

[24] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M.Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, 'The ami meeting corpus', in *Measuring Behaviour, Proceedings of 5th International Conference on Methods and Techniques in Behavioral Research*, (2005).

[25] Prashant Parikh, 'Communication and strategic inference', *Linguistics and Philosophy*, **14**, 473–514, (1991).

[26] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson, 'A simplest systematics for the organization of turn-taking for conversation', *Language*, **50**, 696–735, (1974).

[27] Emanuel A. Schegloff, 'Some practices for referring to persons in talk-in-interaction: A partial sketch of a systematics', in *Studies in Anaphora*, ed., B. Fox, 437–485, Amsterdam: John Benjamins, (1996).

[28] Ielka F. van der Sluis, *Multimodal Reference, studies in automatic generation of multi-modal referring expressions*, Ph.D. dissertation, University of Tilburg, 2005.

[29] M. Walker, J. Cahn, and S. Whittaker, 'Linguistic style improvisation for lifelike computer characters', in *Entertainment and AI / A-Life, Papers from the 1996 AAAI Workshop.*, (1996).

[7] Cost as well as expected effectiveness of actions together determine the selection of referring acts needed for addressing.

Automatic Generation of Gaze and Gestures for Dialogues between Embodied Conversational Agents: System Description and Study on Gaze Behavior

Werner Breitfuss[1] and Helmut Prendinger[2] and Mitsuru Ishizuka[1]

Abstract. In this paper we introduce a system that automatically adds different types of non-verbal behavior to a given dialogue script between two virtual embodied agents. It allows us to transform a dialogue in text format into an agent behavior script enriched by eye gaze and conversational gesture behavior. The agents' gaze behavior is informed by theories of human face-to-face gaze behavior. Gestures are generated based on the analysis of linguistic and contextual information of the input text. The resulting annotated dialogue script is then transformed into the Multimodal Presentation Markup Language for 3D agents (MPML3D), which controls the multi-modal behavior of animated life-like agents, including facial and body animation and synthetic speech. Using our system makes it very easy to add appropriate non-verbal behavior to a given dialogue text, a task that would otherwise be very cumbersome and time consuming. In order to test the quality of gaze generation, we conducted an empirical study. The results showed that by using our system, the naturalness of the agents' behavior was not increased when compared to randomly selected gaze behavior, but the quality of the communication between the two agents was perceived as significantly enhanced.

1 INTRODUCTION

Combining synthetic speech and human-like conversational behavior like gaze and gestures for virtual characters is a challenging and tedious task for human animators. As virtual characters are used in more and more applications, such as computer games, online chats or virtual worlds like Second Life, the need for automatic behavior generation becomes more pressing. Thus, there have been some attempts to generate non-verbal behavior for embodied agents automatically. Systems like the Behavior Expression Animation Toolkit (BEAT) allow one to generate a behavior script for agents by just inputting text [4]. The drawback of most current systems and tools, however, is that they consider only one agent, or only suggest behaviors, such that the animator still has to select appropriate ones by him- or herself. The aim of our work is to generate all non-verbal behavior automatically for conversing agents, so that someone writing a script to be performed by two agents can focus on creating the textual dialogue script and just feed it into the system. A salient feature of our system is that we generate the behavior not only for the speaker agent but also for the listener agent that might use backchannel behavior in response to the speaker agent. Employing two presenter agents holding a dialogue is advantageous, since watching (or interacting with) a single agent can easily become boring and it also puts pressure on users, as they are the only audience. Furthermore, two agents support richer types of interactions and "social relationships" between the interlocutors. Also TV-commercials, games, or news use two presenters, because of the increased interaction possibilities and entertainment value.

In this paper, however, we confine discussion to the case where one user just watches the performance (dialogue) of two virtual agents, and does not interact with them. To assess the quality of our system we conducted an experiment. Twenty participants watched a presentation generated by our system. We randomly assigned them either to a version where the gaze behavior of the agents was informed by our gaze generator or to another version where the gaze was generated randomly. We speculated that the first (informed) version would increase the naturalness of the conversational behavior of the virtual characters and the quality of the communication between them. By "quality of the communication" we mean that the listener is paying attention to the speaker and the speaker addresses the listener in appropriate moments. In the study both versions used the same gestures, since we wanted to investigate the gaze behavior only. The dialogues where provided by a system developed at the Open University by Sandra Williams [21]. It generates a dialogue based on the medical history of a patient. While this system is designed to create shorter dialogues, for our purpose we used its original longer (unmodified) versions. The longer versions are sometimes repetitive, since patients in this database tend to have the some examinations over and over again.

The paper is organized as follows. In Section 2 we discuss related work. Section 3 describes our system and the way gaze behavior and other non-verbal behavior is generated by means of a "walk through" example. In Section 4 we describe our empirical study on gaze generation. The results are presented and discussed in the Section 5 and Section 6. Section 7 gives a short future outlook and concludes the paper.

[1] University of Tokyo, Japan, email: werner@mi.ci.i.u-tokyo.ac.jp
[2] National Institute of Informatics, Japan, email: helmut@nii.ac.jp

2 RELATED RESEARCH

Existing character agent systems already support the automated generation of some behaviors, such as automatic lip-synchronization. The next step is to automatically generate agents' conversational behavior from text. In this section, we report on some previous attempts, which combine various disciplines like computer animation, psychology, and linguistics.

2.1 Single Agent Systems

The BEAT system [4] generates synthesized speech and synchronized non-verbal behavior for a single animated agent. It uses plain text as input, which is then transformed into animated behavior. First, text is annotated with contextual and linguistic information, based on which different (possibly conflicting) gestures are suggested. Next, the suggested behaviors are processed in a 'filtering' module that eliminates gestures that are incompatible. In the final step, a set of animations is produced that can be executed, after necessary adoptions, by an animation system. The BEAT system can handle any kind of text and generate a run-able agent script automatically. The system uses a generic knowledge base where information about certain objects and actions is stored, and the selected gestures are specified in a compositional notation defining arm trajectories and hand shapes independently, which allows the animator to add new gestures easily, or adjust existing ones.

The PPP Persona [1] is a life-like interface agent that presents multimedia material to a user. The behavior of the agent during the presentation is controlled partly by a script, written by the author of the presentation and partly by the agent's self-behavior. Behavior in the case of this agent is mostly acts such as pointing, speaking and expressing emotions and the automatically generated self-behavior which includes (1) idle-time actions to increase the personas life-like qualities, for example breathing or tipping a foot, (2) reactive behavior letting the agent react to external events like user reactions immediately, and (3) so-called navigation acts which display the movement of the agent across the screen, like jumping or walking. To generate this kind of behaviors a declarative specification language was used.

[14] describes a system that converts Japanese text into an animated agent that synchronously gestures and speaks. For assigning an appropriate gesture to some phrase the authors employed communicative dynamism (CD) as introduced by McNeill [13] and results from an empirical study that identified lexical and syntactic information and their correlation with gesture occurrence. For every "bunsetsu", the Japanese equivalent for a phrase in English, the system adds a gesture at a certain possibility, which is derived from the results of the study and the CD value. Similar to our system the specific gestures are defined in a library and if no specific gesture can be found for the bunsetsu, a beat is added as default gesture.

2.2 Multi Agent Systems

Another system is the eShowroom demonstrator[11], which was developed as a part of the NECA Project. The application automatically generates dialogues in a car-sales setting between an agent who acts as a seller and a second agent acting as buyer. The user has the possibility to choose certain parameters like topic, the personality and the mood of the virtual characters, which control the automatically generated dialogues. Also the gestures and behavior of the two screen characters would be generated by the NECA eShowroom demonstrator. It uses three types of behavior: (1) turn taking signals like looking to the other interlocutor at the end of the turn, (2) discourse functional signals, which are gestures that depend on the type of the utterance (type refers to dialogue acts like inform or request), (3) feedback gestures are also generated to signal that the listener is paying attention to the speaker.

A different approach is suggested in [9]. This system supports the author in writing agent scripts by automatically generating gestures based on predefined rules, and using machine learning to create more rules from the set of predefined rules. It was used in the COHIBIT system, where the author first has to provide a script containing the actions for two virtual characters. In the next step the author writes simple gesture rules using his or her expert knowledge. Using this corpus of annotated actions the system can learn new rules. In the third step the system suggests the most appropriate gestures to the author, which are, after resolving conflicts and filtering, added to the already existing ones. Finally it produces a script with the gestural behavior of both virtual characters. Similar to our work, two agents are used, but since we want to reduce the workload to the minimum, our system does not require any input from the author except the dialogue to be presented by our characters.

2.3 Related work on eye gaze and gestures

[7] investigated the many different functions of gaze in conversation and its importance for the design of believable virtual characters. The gaze behavior of our agents is informed by empirically founded gaze models [9,16,20]. [8] analyzed gaze behavior based on two-person dialogs and found that gaze is used to regulate the exchange between the speaker and listener. In that work, different gaze patterns like the q-gaze (the speaker is looking at the person he/she is interacting with), and a-gaze (p is not looking at the interlocutor) were defined. It was found that the speaker looks at the listener while speaking fluently, but looks away when starting to speak or during hesitation (influent speech). In this way, speakers can keep the listeners attention or, by looking away, gain time to think about what to say next. Another finding is that mutual gaze can regulate the level of emotionality between interlocutors. The experiment described in [20] evaluates gaze behavior in multiparty environments, where four-person groups discussed current-affair topics in face-to-face meetings. Their results show that on average, interlocutors look about seven times more often at the speaker they listen to, than at others, and speakers looked about three times more at the addressed listener than at non-addressed listeners. Furthermore, the total amount of time spent gazing at each individual in a group of three is nearly 1.5 times higher than if the visual attention of the speaking person were simply divided by three. These results are very relevant for our gaze algorithm since they give us the basis for a 'two agents' situation. And they also provide the needed information for our gaze generation rules. The work in [16] developed a model of attention and interest based on gaze behavior. An embodied conversational agent may start, maintain, and end a conversation dependent on its perception of the interests of the other agents.

Other related research was done is [6], which introduces a behavior synthesis technique for conversational agents in order to generate expressive gestures, including a method to individualize the variability of movements using different dimensions of expression. The work described in [10] presents a gesture animation system that uses results from neurophysiologic research and generates iconic gestures from object descriptions.

3 GESTURE GENERATION SYSTEM

Our system consists of three different modules:

- Language Tagging module,
- Non-Verbal Behavior Generation module,
- Transformation to simple script or MPML3D module.

The Multimodal Presentation Markup Language is used to control the behavior of our 3D agents [15]. We choose a modular pipelined architecture to support future extensions. The code of the system is written in Java, and the XML format is used to represent and exchange data between modules.

The Language Tagging module takes the input dialogue text and uses the language module from the BEAT toolkit [4] to annotate linguistic and contextual information. Next, the Behavior Generation module adds non-verbal behavior like eye gaze and gestures to the annotated input sentence. In the final step, an agent script file is produced. In our implemented system, we can produce an MPML3D file but also a simpler script that can be used as an interface to other systems. The MPML3D player displays the embodied characters agents.

In our system, gaze patterns are generated for two different types of roles: (1) the speaker, i.e. the agent that is speaking and addresses the other agent, and (2) the listening agent. We can currently generate gaze behavior and gestures for these two roles, based on a given set of rules. Gaze directions have certain probabilities of occurrence, which we derived from existing gaze models [9,16,20]. In order to avoid conflicts between certain gaze behaviors, like looking in two different directions at the same time, we assigned priorities to them. Typically, more specific gaze behaviors (such as looking at speaker/listener) have higher priority than e.g. looking around. Moreover, we prioritize gazes that occur before starting the utterance, i.e., speakers typically look away before starting a long utterance (in order to concentrate on planning their dialogue contribution).

The rule in Figure 1 (adapted from [4]) shows one example of how the gaze behavior for the speaker is generated.

```
FOR each THEMA node in the tree
  IF at the beginning of the utterance
  Or 70% of the time
      Look away from listener
FOR each RHEMA node in the tree
  IF at the end of the utterance
  Or 73% of the time
      Look at listener
```

Figure 1. Gaze generation rule for the speaker

In addition to generating the gaze behaviour for the speaking agent we also have to consider the agent in the role of the listener. Since listeners typically look at speakers when they start an utterance (after taking the floor) to demonstrate their attentiveness, we developed rules like the one in Figure 2.

```
FOR each THEMA node in the tree
  IF at the beginning of the utterance
  Or 80% of the time
      Look at speaker
FOR each RHEMA node in the tree
  IF at the end of the utterance
  Or 47% of the time
      Look at the speaker
```

Figure 2. Gaze generation rule for the listener

We also added gaze rules for certain gestures enacted by the speaker. For instance, pointing gestures have to be accompanied by the correct gazes. In our presentation scenarios we mostly use rectangular slides in the centre between the agents and smaller objects around them. As all of those objects have a definite position either left or right to the agent, we can exploit this knowledge to add the correct gaze direction to the agents' behavior when they talk about or point at the object. However, since defining the objects' position in the scene would increase the workload of the author, we also implemented the following straightforward principle. Every time a phrase such as "on my right side" or "to the left" occurs, we add a pointing gesture to the speaker's behavior tree. When the speaker's tree is completed, we recompile the listener's tree to adopt its gaze behaviour to the pointing gestures, and add the gestures to the correct side. The gestures of our agents are generated in similar manner, broadly following rules proposed in [4].

Let us now walk through one simple example utterance and see how our system works. As input we take the sentence: "This is just a small gaze example." [2] First, the input is sent to the Language Tagger module, which annotates the sentence with linguistic and contextual tags. The output of this process is shown in Figure 3. Here, "NEW" means that the word has not yet occurred in the conversation, and is thus a candidate for being accompanied by a "beat" gesture.

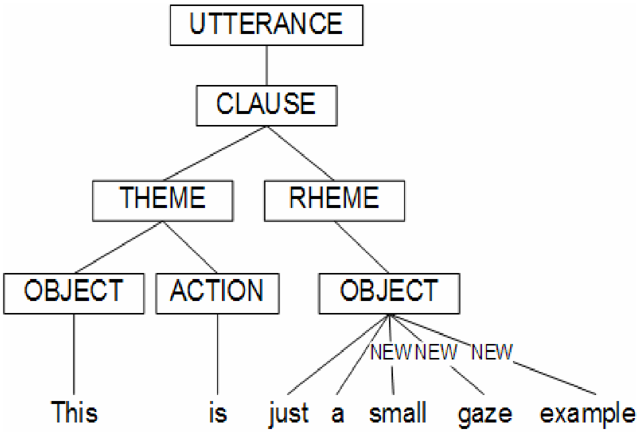


Figure 3. The output tree of the language module.

In the next step, we pass this newly constructed tree to our Behavior Generation module. It first generates a new tree with the gaze behavior and gestures (and speech parameters) for the speaker and a second tree for the listener. The tree for the listening character has the same structure as the speaker's tree, but contains the nodes for the non-verbal behavior that should be displayed by the listener agent.

Gestures are generated in two steps: first we add a beat every time some gesture is appropriate. After that the utterance is passed on to another layer that adds more specific gestures. To do this we provide a library, where we defined word bags associated with gestures. For instance, there is one word bag that contains the words "small, narrow, tiny" and the gesture for expressing something of little size. Hence, every time a word with the lemma of those words occurs in the sentence the beat gesture which has a lower priority is overwritten by the more specific gesture for small.

Figure 4 shows the speaker's tree, which was generated by our system for the sentence used in this short example. The root node of the tree is the utterance, and there is a speech pause between the theme and rheme of the sentence (see [4] for a discussion of speech parameters). The gaze behavior "Gaze away" and "Gaze at listener" is derived from the previously discussed rule (Figure 1). The gesture behavior is generated according to dedicated gesture generation rules of the Behavior Generation module. In our example, a beat gesture is selected to accompany the word "just", and an iconic gesture (for describing something small) is suggested to co-occur with the phrase "small gaze example".

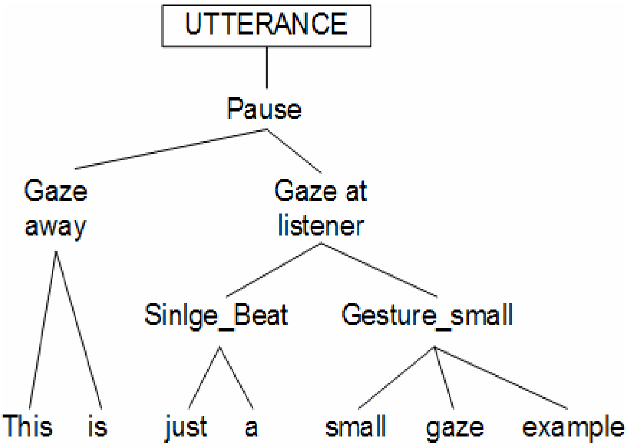


Figure 4. Tree for the speaker behaviour.

The behavior tree of the listener agent is generated similarly to that of the speaker agent (see Figure 5). It is based on the same tree that is output from the Language Tagging module of the speaker agent, but applies listener behavior generation rules instead of speaker rules. Again, we start with root node "UTTERANCE". During the speaker's speech pause, no behavior for the listener agent is defined. The listener's gaze behavior is added according to the rule in Figure 2, i.e. the listener is looking at the speaker when the utterance begins. Accordingly, our system creates the label "Gaze at speaker". Since the listener agent is paying attention to the speaker, it continues to look at the speaker also in the "rheme part" of the utterance.

Thereafter, appropriate gestures are suggested for the listener agent. Whereas no gesture is suggested for the phrase "just a", the phrase "small gaze example" is accompanied by head nods. In our system, a head nod is a basic gesture type for the listener. It is the gesture with the lowest priority and is used when no other, more specific gesture can be suggested. In the future, a dedicated "backchannel" knowledge base will be created to insert listener head nods in an informed, systematic manner.

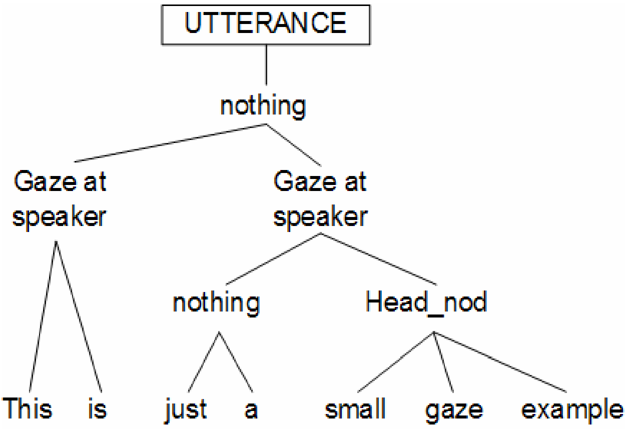


Figure 5. Tree for the speaker behaviour.

After speaker and listener behavior trees are created, they are passed to the Transformation module, which compiles them into a synchronized MPML3D XML file or a simpler XML file.

Before generating the MPML Script we have to run the two trees through a small set of filters to handle any unexpected mistakes and to make sure no errors were forwarded to the script. We also use the filters to avoid minor technical problems like certain timing issues. Currently, the MPL3D Player cannot synchronize gestures that start at the beginning of a word and stop at the end of the same word.

This last module combines the speaker and listener tree by adding the actions of both agents for every utterance into one MPML3D structure called "task". The MPML Script contains parallel and synchronized actions which can be started and ended at the beginning, middle , or end of a certain word. First we add all the actions that should occur before the speaker starts to talk, mostly gaze behavior, like looking away for the speaker and idle gestures for the listener. The next action that is added is speaking itself. In the following step, we add the gaze behavior, which has to be aligned with the appropriate words. Gaze is implemented by having the head turn to a certain direction. There is a set of parameters that can be used, like the vertical angle in which the head should be moved and the speed of the movement. As the last level we add the gesture for the speaking agent and the listening agent.

Figure 6 shows the MPML3D code, which our system generated for the sentence used in the example.

```
<Task>
  <Action>ken.turnHead(20,0.2,0.3,0.2)</Action>
 <Parallel>
  <Action name="kenspeak">
  ken.speak("This is just a small gaze example")
  </Action>
  <Action startOn="kenspeak[x0].begin"
          stopOn="kenspeak[x5].end">
  ken.turnHead(20,0.2,1,0.2)
  </Action>
  <Action startOn="kenspeak[x6].begin"
          stopOn="kenspeak[x14].end">
  ken.turnHead(0,0.2,5,0.2)</Action>
  <Action startOn="kenspeak[x0].begin"
          stopOn="kenspeak[x14].end">
  yuuki.turnHead(0,0.2,1,0.2)
  </Action>
  <Action startOn="kenspeak[x6].begin">
  ken.gesture("beat_one")
  </Action>
  <Action startOn="kenspeak[x17].begin">
  ken.gesture("showsmallvertical")
  </Action>
  <Action startOn="kenspeak[x9].begin">
  yuuki.gesture("headnod")
  </Action>
 </Parallel>
</Task>
```

Figure 6. The MPML3D code for our example.

Our System can also produce a simpler script (see Figure 7). It contains only 3 entries: (1) the text of the utterance; (2) a mood, which is generated by using the [19] system, so that a virtual character that is able to display emotions, can use this information; (3) the gesture with the highest priority.

```
<utterance>
 <text> This is just a small gaze example</text>
 <mood>neutral</mood>
 <gesture>showsmallvertical</gesture>
</utterance>
```

Figure 7. Simple XML code for our example.

The simple script is intended to be used for other agent systems, which can only display one gesture per utterance or are limited with respect to gesture and speech synchronisation like the Cantouche[3] agents.

Figure 8 shows our agents performing the example sentence.

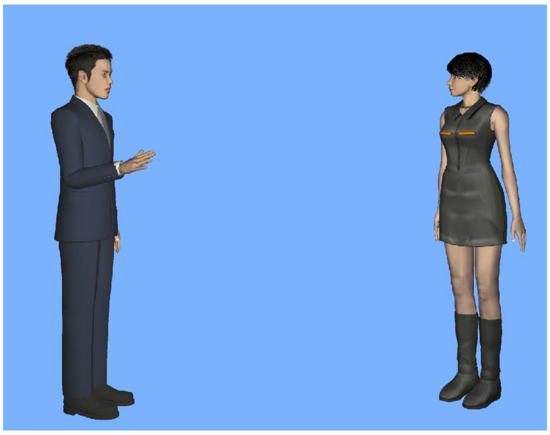


Figure 8. MPML3D Agents enacting the example sentence.

4 METHOD

4.1 Design

In the study, we compared two different versions of a presentation. In one version, gaze behavior was generated by our system (the informed version). In the control version, gaze was generated in a random manner (uninformed version). By “random” we mean that every time our system suggested a particular gaze behavior, a gaze direction was randomly chosen instead, which could be “look away” (to the left or to the right) or “look at the other agent”.

The gestures used were the same in both versions, and consisted mostly of beats in case of speaking character, and head nodes in case of listening agent. We kept the set of the gestures used very limited, since as suggested in [5] too many gestures can distract the user and consequently have a negative effect on the perception of the overall presentation and gaze behavior.

We run the study primarily to investigate the effect of our new gaze module on two dimensions: (1) the naturalness of the presentation, and (2) the perceived quality of the conversational behavior between the two agents. The dialogues where generated by the [21] system.

4.2 Participants

Twenty people participated in the study, 18 males and 2 females, their age ranged from 22 to 35 years (mean age 28.3 years). Except for two external people, subjects were students or researchers from the National Institute of Informatics, Tokyo. Subjects received 1000 Yen for participating.

4.3 Materials

The raw dialogues for the presentation were provided by an automated dialogue generation system [21], and contain the conversation between Yuuki, a female senior nurse and Ken, a male junior nurse.

The dialogue contained 106 utterances, and the duration of the presentation was around 5 minutes. The topic of the dialogue was about the medical history of a fictional patent that has breast cancer.

The following is a typical paragraph of the presentation. We wish to note again that for the purpose of the experiment (investigating gaze), we used the long, unmodified dialogue output by the system. This output was not meant to be shown to subjects when investigating e.g. the effectiveness of the dialogue.

Yuuki: For May the 24th what does the medical record say?
Ken: On May the 24th she did a self examination.
Yuuki: What did she find?
Ken: A lump.
Yuuki: What does it say next?

Ken: On May the 19st she did another self examination.
Ken: And she still had a lump.
Yuuki: And then?
Ken: On June the 7th she did another self examination.
Ken: And she still had a lump.
Ken: From May the 20th to August the 5th she had a chemotherapy course.
Ken: What is a chemotherapy course ?
Yuuki: A chemotherapy course is a treatment with drugs.
Yuuki: Is that clear ?
Ken: Uhhuh.
Yuuki: What does it say next ?
Ken: On June the 24th she had another examination.
Ken: And she still had lymphadenopathy.

4.4 Apparatus

The experiment was run on a Dell workstation with a dual-core processor. The material was presented to the subjects using a UXGA (1600 × 1200 pixels) flat screen color monitor. The speech for the agents was generated by Loquendo ([12]), a commercial text-to-speech (TTS) engine. The agents controlled by our MPML3D Player ([15]).

For videotaping the participants we used a digital camera that was positioned behind subjects and a mirror, which was fixed on the right side of the monitor, so that we could capture the face and the shoulders of the subjects. Figure 9 depicts the setup of our study.

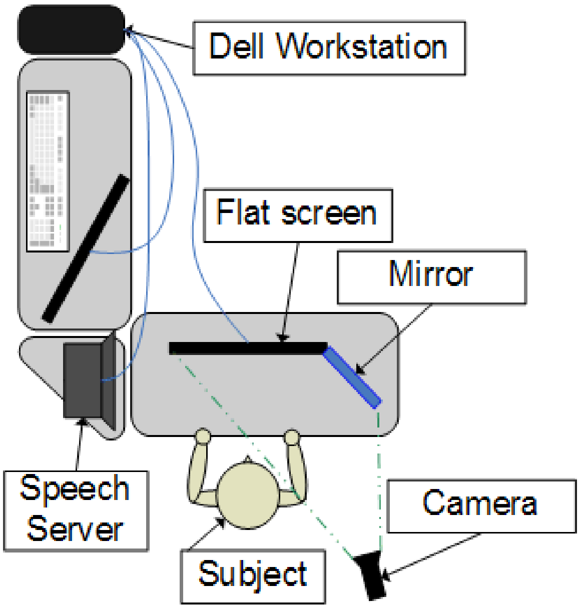


Figure 9. Experimental setup.

4.5 Procedure

Subjects entered the experiment room individually and received a written instruction about the procedure. The instruction given to the subjects was to watch the presentation as they would watch a presentation given by human presenters and they should keep an eye on the behavior of the agents.
While watching the dialogue between our two agents, the participants were videotaped for further analysis (Figure 10: screen with presentation to the left, participant to the right).

After watching the presentation, both groups of participants were asked to fill out a questionnaire with twelve questions.

1. The female agent (Yuuki) was friendly.
2. The male agent (Ken) was friendly.
3. The conversation between the two agents seemed very natural.
4. Sometimes I thought the agents react to each other in a strange way.
5. I felt that the two agents are a good team and communicate with each other well.
6. It seemed that the agents did NOT pay attention to each other.
7. I trusted the female agent (Yuuki).
8. I trusted the male agent (Ken).
9. I found the conversation easy to follow.
10. The conversation captured my attention.
11. I found that my attention wandered.
12. I found the conversation hard to understand.

The answers were based on a Likert scale, and ranged from 1 (“strongly agree”) to 7 (“strongly disagree”). At the end of the questionnaire we also provided the possibility of free text entry, so that subjects could state their comments without restrictions.
Each session of the experiment lasted around 15 minutes per person, and was conducted in our multimedia room.

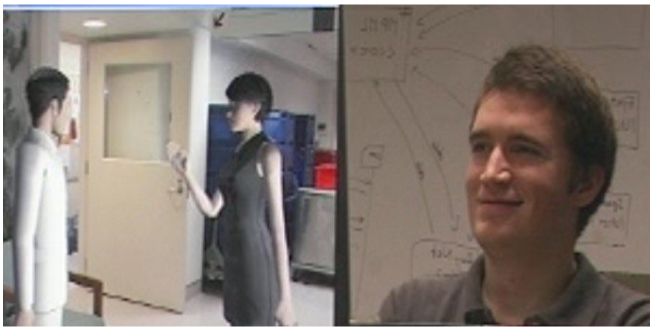


Figure 10. Screen and participant.

5 RESULTS

We performed a t-test (two-tailed) to determine the statistical significance of the differences between the averages (significance level α set to .05).

The averages of the answers to the questions in the questionnaire can be found in Figure 11 where the x-axis gives the number of the question, and the y-axis shows the value for each question. Figure 12 shows the means and standard deviations of the questions Q1 to Q12, where the first row gives the values, mean and deviation, of the uninformed version and the second row gives the values for the informed version.

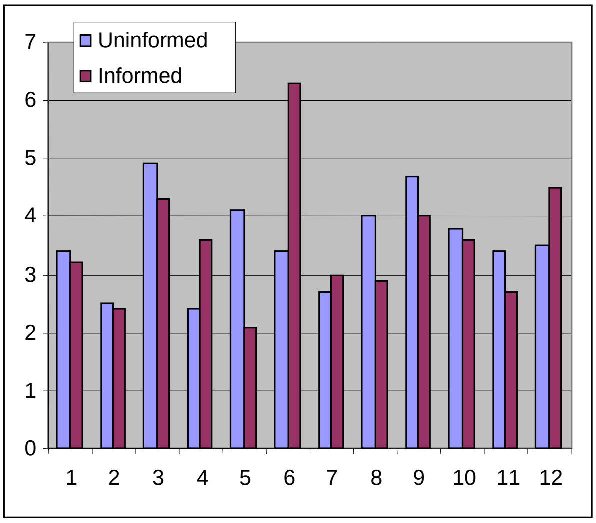
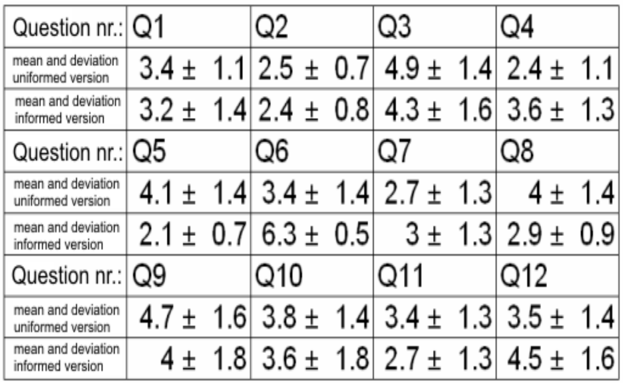


Figure 11. The means for the questions.



Question nr.:	Q1	Q2	Q3	Q4
mean and deviation uniformed version	3.4 ± 1.1	2.5 ± 0.7	4.9 ± 1.4	2.4 ± 1.1
mean and deviation informed version	3.2 ± 1.4	2.4 ± 0.8	4.3 ± 1.6	3.6 ± 1.3
Question nr.:	Q5	Q6	Q7	Q8
mean and deviation uniformed version	4.1 ± 1.4	3.4 ± 1.4	2.7 ± 1.3	4 ± 1.4
mean and deviation informed version	2.1 ± 0.7	6.3 ± 0.5	3 ± 1.3	2.9 ± 0.9
Question nr.:	Q9	Q10	Q11	Q12
mean and deviation uniformed version	4.7 ± 1.6	3.8 ± 1.4	3.4 ± 1.3	3.5 ± 1.4
mean and deviation informed version	4 ± 1.8	3.6 ± 1.8	2.7 ± 1.3	4.5 ± 1.6

Figure 12. Means and standard deviations.

We predicted that the gaze behavior generated by our system would generate a more natural dialogue and the agents would be perceived as communicating well which each other.
Regarding the first dimension (naturalness), we partly obtained significant results, while always showing the expected tendency in the answers (Questions 3 and 4). The results for the question concerning the naturalness of the agents' behavior, the results for Question 3 showed that the informed version only slightly improved the naturalness of the conversation (p = .396). The result for Question 4 though showed that the agents reacted significantly less strange (by contraposition, more natural) to each other in the informed version (p < .041).

The results for questions concerning the conversational behavior between the agents (quality of communication) are statistically significant. The results confirm the hypothesis that our system can significantly increase the level of perceived quality of conversational behaviour between the two interlocutors For Question 5, p < .0017, and for Question 6, p < .0001.

The questions regarding the friendliness of the agents (Questions 1 and 2), or about the trustworthiness of the agents (Questions 7 and 8), did not yield any important results. Note, however, that the results for Question 8 indicate that the male character was nearly significantly (p = .053) more trustworthy in the version informed than in the uninformed version.

6 DISCUSSION

The purpose of the experiment was to obtain empirical data on our newly implemented system, with a focus on the gaze behavior of the agents. The data from the questionnaires supports our expectations that the version with gaze behavior informed by our system would outperform the version with randomized gaze in terms of quality of conversational behavior between the two embodied virtual characters. In particular, the result for Question 6 provides strong evidence that the participants noticed that the agents pay more attention to each other in the informed version.

The poor results regarding the naturalness of the presented dialogues were somewhat surprising. The free-text comments we received from the participants (as part of the questionnaire) gave three different reasons why they rated the naturalness as rather poor. One issue was the beat gesture, which seemed to be irritating, and the hand movement was too fast and too wide. A second problem was the voice generation, which did not produce satisfying results for technical medical terms. (In fact, this problem could have been avoided if we had provided the correct pronunciation of rare technical terms to the TTS engine beforehand.). Third, some subjects criticized parts of the dialogue as unnatural. They noted that there are too many repetitions and some of the answers given by the junior nurse (Ken) were irritating. There is one particular part in the dialogue, where the senior nurse explains the function of auxiliary lymph nodes, and the junior nurse answers with a short "Cool". As the video analysis showed, most participants found this part rather humorous, but others stated in their comments, that it is strange to use the word "cool" in the context of cancer. In future we might also consider to use a more common topic, to enhance the naturalness of the dialogue it self. The experiences with our study provide valuable insights for designing better studies with our non-verbal behavior generation system in the future.

7 CONCLUSIONS AND FUTURE WORK

There is ample evidence that agent-based multimodal presentations can entertain and engage the user, and are also an effective way to mediate information [18]. In this paper, we described our system that automatically generates gaze and gestures for two agents, in the roles of speaker and listener. It uses a dialogue script as its only input (from the content creator), and transforms it into a run-able multimodal presentation using two highly realistic 3D character agents.

In our future work, we plan to analyze the emotional content of text based on the work described in [19], and add emotional expressions to the agents' behavior in order to improve the naturalness of the performed dialogue. The emotion expressed in a sentence will also affect voice parameters, gaze, and gesture behavior. Conversational behavior is also influenced by the social role (instructor-student, employer-employee, etc.), the cultural background, and the personality of the interlocutors.

Another venue of research relates to including a model of the user as a listener, who might be addressed by the agents.

Our next step, however, will address more feasible issues. In addition to extending the set of behavior generation rules for the listener agent, we want to align the behavior of the agents with respect to a slide show and virtual objects in a 3D environment. Here, we have to analyze phrases like “if you look at the slide” and generate appropriate behavior for the speaker and listener agent. Among others, the selected gaze behavior has to be timed and directed to specific locations in the 3D environment. In this way, “joint attention” (gaze) behavior will be implemented.

For all of our ideas, the focus will remain on the exploration of ideas that ultimately lead to a minimal workload for content creators, while ensuring high-quality, professional output in the form of natural and enjoyable multimodal presentations.

8 REFERENCES

[1] André, E., Müller, J., and Rist, T.: The PPP Persona: A Multipurpose Animated Presentation Agent In: Catarci T., Costabile M.F., Levialdi S., and Santucci G., editors, Advanced Visual Interfaces, pages 245-247. ACM Press (1996)

[2] Breitfuss W., Prendinger H., Ishizuka, M.: Automated Generation of non-verbal behavior for virtual embodied characters. In: Proceedings of the 9th international conference on Multimodal interfaces, pages 199 -202 (2007)

[3] Cantouche Inc. URL(2007) www.cantoche.com

[4] Cassell, J., Vilhjálmsson, H., and Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: Proceedings of SIGGRAPH 2001, pages 477-486 (2001)

[5] Craig, S., Gholson, B., and Driscoll, D. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundance. Journal of Educational Psychology, 94(2); pages 428 – 434, (2002)

[6] Hartmann, B., Mancini, M., Buisine, S., and Pelachaud, C.: Design and evaluation of expressive gesture synthesis for embodied conversational agents. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems. ACM Press (2005)

[7] Heylen, D.: Head gestures, gaze and the principles of conversational structure, In: International Journal of Humanoid Robotics Vol. 3 Nr. 3, pages 241-26 (2006)

[8] Kendon, A.: Some functions of gaze-direction in social interaction. In Acta Psychologica 26, pages 22-63, North-Holland Publishing Co. (1967)

[9] Kipp, M.: Creativity meets automation: Combining nonverbal action authoring with rules and machine learning, In: Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA 2006), Springer, pages 217-242 (2006)

[10] Kopp, S., Tepper, P., and Cassell, J.: Towards integrated microplanning of language and iconic gesture for multimodal output. In: Proceedings International Conference on Multimodal Interfaces 2004, ACM Press, pages 97-104 (2004)

[11] Krenn, B., Grice, M., Piwek, P., Schroeder, M., Klesen, M., Baumann, S., Pirker, H., van Deemter, K., and Gstrein, E.: Generation of multi-modal dialogue for net environments. In Proceedings of KONVENS-02, pages 91–98, (2002)

[12] Loquendo Vocal Technologies and Services, URL (2008) www.loquendo.com

[13] McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. Chicago, IL/London, UK: The University of Chicago Press. (1992)

[14] Nakano, Y., Okamoto, M., Kawahara, D., Li, Q., and NishidaT.: Converting Text into Agent Animations: Assigning Gestures to Text, In: Journal of Humanoid Robotics Vol. 3 Nr. 3, pages 241-26 (2006)

[15] Nischt M., Prendinger, H., André, E., and Ishizuka M.: MPML3D: a reactive framework for the Multimodal Presentation Markup Language. In: Proceedings 6th International Conference on Intelligent Virtual Agents, Springer, pages 218-229 (2006)

[16] Peters, C., Pelachaud, C., Bevacqua, E., and Mancini, M.: A model of attention and interest using gaze behavior. In: Proceedings of 5th International Conference on Intelligent Virtual Agents 2005, pages 229-240 (2005)

[17]Prendinger, H. and Ishizuka, M., editors. Life-Like Characters. Tools, Affective Functions, and Applications. Cognitive Technologies. Springer, Berlin Heidelberg, (2004)

[18] Rist, T., André, E., Baldes, S., Gebhard, P., Klesen, M., Kipp M., Rist, P., and Schmitt, M.: A review of the development of embodied presentation agents and their application fields. In: Prendinger and Ishizuka [17], pages 377-404

[19]Second Life Linden Research Inc., URL (2008) www.secondlife.com

[19] Shaikh, M., Prendinger, H. and Ishizuka, M.: A Cognitively Based Approach to Affect Sensing from Text. In: Proceedings 10th International Conference on Intelligent User Interfaces, ACM Press, pages 349-351 (2006) Computing Systems (CHI’03), pages 521–528, ACM Press (2003)

[20] Vertegaal, R., Weevers, I., Sohn, C., and Cheung, C.: Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems (CHI’03), pages 521–528, ACM Press (2003)

[21] Williams, S., Piwek, P., and Power, R.: Generating monologue and dialogue to present personalised medical information to patients. In Proceedings of the 11th European Workshop on Natural Language Generation, pages. 167-170 (2007)

A Scene Corpus for Training and Testing Spatial Communication Systems

Michael Barclay and **Antony Galton** [1]

Abstract. It is argued that a 'scene corpus' would be a useful tool for the training and testing of systems for grounded spatial communication in the same way that text corpora have been used for training and assessing other language processing systems. Such a scene corpus would need to allow a full range of spatial relationships to be expressed over a range of scale spaces. The scenes should be sufficiently complex to allow sequential spatial descriptions to be constructed. The integration of listener models and reference frame variation will be required. The interface design will need to allow for different implementations and corpus extensions. Initial steps in the design of a scene corpus are described.

1 INTRODUCTION

To see that spatial communication is among the most fundamental and important forms of communication in which humans engage, consider only the sentence, "There is a lion under those trees!". Before the emotional, political, financial, academic or other realms were invented as subjects for discourse the physical realm existed and required description and discussion. Spatial metaphors influence and structure more areas of human communication than any other [13].

Spatial communication is often multi-modal communication: the words 'those trees', in a typical environment, may only be meaningful if distinguished from other groups of trees by a gesture. Direction is indicated by pointing, limits on spatial location or extent may be indicated by two-handed indications (or arm sweeps) and turns or direction changes by representational hand movements.

Even the simplest of spatial phrases, containing two nouns, linked by a preposition conceals much complexity depending on the form of the objects involved [8, 10, 21]. If some functional relationship between the objects is also involved the difficulties of machine generation of spatial language become even more apparent [7, 5, 15, 9, 4]. A cup may be *usefully* described as 'on the table' even if it is actually 'on' a saucer which is 'on' a mat which is 'on' a tablecloth which is 'on' the table. The cup might not *usefully* be described as 'on' the saucer, since the saucer is as mobile as the cup and does not help a listener find the cup. In any case the cup might be 'in' the saucer (not on it) and the tablecloth 'over' the table. To form even a simple spatially locative phrase acceptably requires a broad knowledge of physical laws (in the naive sense), along with concepts of support, containment, mobility and persistence. Knowledge of how objects interact (hardness and softness, deformation, permeability etc) and how they are conventionally used also plays an important part.

This complexity has meant that the best systems to date for analysing and generating spatial language have concentrated on elements of the problem rather than on its entirety [20, 19, 11, 22, 16]. These systems have typically used their own sets of scenes, although in most cases they could have been adapted to use a standard corpus.

Currently comparison between systems and their methods and algorithms would be difficult because of this concentration on different elements of spatial language generation. Over the next few years, however, the development of more complete spatial communication systems, including those with multi-modal output, is anticipated. An agreed scene corpus which is designed to address all the elements of spatial communication will be essential.

2 CONTENT REQUIREMENTS FOR A SCENE CORPUS

2.1 Scope of the corpus

To encompass all aspects of spatial language in practice means that the scene corpus must contain examples of all expressible spatial relationships between as wide a variety of objects as possible, to enable training and testing of comprehensive spatial language systems. Given the task of desribing the location of object(s) in a scene (the 'located' object(s)), typically with respect to one or more 'reference' or 'ground' object(s), these aspects of spatial language can be summarised as follows:

1. Selection of appropriate reference object(s)
2. Adoption of appropriate reference frame
3. Use of correct spatial prepositions
4. Incorporation of gesture, emphasis or other non-verbal communication
5. Integration of listener models
6. Strategies for construction of multi-phrase descriptions

How these are to be incorporated in the corpus is discussed in the following subsections. Note the problem being addressed is not that of referring expression generation [6] in which disambiguation is the aim and for which the 'tuna' corpus [24] was designed. The located object is assumed to be unambiguously identifiable but its location must be described (relative to a reference object).

2.2 Reference object selection

The general problem of reference object selection does not seem well addressed in the literature, although there is a specific body of work on landmark selection [2, 23, 1, 18]. Generalising and extrapolating from this work, the factors influencing reference object selection, so far identified, are as follows;

1. Reference object locatability, comprising

[1] University of Exeter, UK, email: mjb231@ex.ac.uk

   (a) Visibility
   (b) Unambiguousness (Uniqueness)
   (c) Persistence (if listener not present)
   (d) User acquaintance with reference

2. Search space optimisation, comprising
   (a) Reference object location
   (b) Reference geometry
   (c) Scale of located and reference objects
   (d) Listener location

The corpus must contain scenes that allow these influences to be discriminated and their weights compared. Consider for instance a scene where a car (whose location is to be described) was parked beside a row of identical houses but across the road from a bus stop. The best located, most visible and persistent landmark would be the nearest house but this being ambiguous (there are many houses) the bus stop might be the best reference (along with the preposition 'opposite' possibly). Cases where the house might or might not be the best reference when disambiguated by a gesture accompanying the verbal description would also be needed.

2.3 Reference frame selection

The choice of reference frame is crucial to the acceptability [3] and effectiveness [17] of spatial communication. Reference frames can be briefly described as;

1. Speaker-centred (deictic). 'Left', 'right', 'in front of', 'behind', etc., are relative to the speaker.
2. Absolute (extrinsic). Typically this frame will relate to a 'previously agreed' outer reference object, such as the earth in the case of North/South etc.
3. Object-centred (intrinsic). This reference frame may be chosen when a reference object has a distinct orientation defined either by its function (e.g., a car) or by convention (e.g., some buildings).
4. Listener-centered. This is usually equivalent to either object-centered, if the listener is the reference (as in 'it's in front of you'), or speaker-centered, as in a typical route description where the speaker is talking a listener through a route as though they were together.

The scene corpus will need to include objects that have functionally or conventionally defined orientations as well as objects that do not. Currently it is thought that these orientations will need to be explicitly noted as well as indicated by features on the objects in question. Scenes will also have to have a defined external reference orientation (e.g., a North pointer) and defined positions for the listener and speaker.

2.4 Preposition assignment

The number of English spatial prepositions is small (some 70 are listed in [10]) compared to the number of expressible spatial relationships. Although some duplication is apparent (e.g., 'above' and 'over' can be interchanged in some examples), there is even more 'overloading' of meaning on some prepositions (even excluding metaphorical usage) as shown by the discussion of 'over' in [14] or the discussion of 'in' in [4].

No attempt has yet been made to devise even a partially grounded, trainable system, capable of acceptable use of all of the prepositions listed in [10] and it is an open question how large a scene corpus would need to be to enable this training. Minimally the corpus should include scenes in which some objects have spatial relationships that can unambiguously be mapped onto each of the prepositions in [10]. A wide range of representations of the common geometric prepositions will inevitably be included.

2.5 Non-verbal communication

The corpus can and should be designed to train and assess the use of gesture to distinguish and disambiguate objects in conjunction with verbal communication (as well as simply to indicate objects, locations and ranges). More work will be required to decide how far a scene corpus can be taken in this respect. For example, if the use of intonation or emphasis to indicate the degree of belief in the location of an object is required, 'partial' information from a source outside the scene corpus as currently envisaged will be needed.

2.6 Listener location and listener models

Currently three elements of a listener model are assumed to be included in the scene corpus;

1. Whether the listener is present at the scene (important to test discernment of the relevance of persistence in a reference object)
2. The listener's location in the scene if he is present (to test reference frame selection and preposition assignment)
3. Listener acquaintance with specific reference objects (to test the use of less visible references if their location is already known)

Aspects of a listener model such as preference and cognitive capacity (as discussed in [12]) are outside of the scope of the scene corpus.

The speaker model is currently limited to location which is coincident with the 'openGL camera' location for the scene.

2.7 Complex phrases and multi-phrase descriptions

At least three classes of complex description forms can be identified which are potentially important for a spatial communication system to be able to handle:

1. Complex locative statement. A locative phrase with more than one reference such as ”The vase is in the living room, on the table under the window”
2. Path and route descriptions. These are possibly the most important for multi-modal systems. Descriptions such as ”the man came from between the shops, ran along the road and disappeared down the alley by the church” are seldom unaccompanied by gestures.
3. Sequential scene descriptions. These are linked descriptive phrases such as ”Behind the shops is a church, to the left of the church is the town hall. In front of the town hall is a fountain”

Strategies for sequential scene description are discussed in [12]. It would be difficult to capture all the necessary considerations and *design* a corpus to comprehensively test these behaviours. The complexity of the scenes in the corpus, however, should enable systems to generate acceptable phrases and gesture sequences of the sort outlined in the list above.

3 USAGE ISSUES

3.1 Scene representation and partial information

The scenes in the corpus as currently envisaged will contain objects that may not be visible to an observer from a given point of view. It is not at present intended to enable the 'removal' of these objects from the scene as presented to the spatial communication generation system. Thus the information in a scene should be thought of as analogous to a 'speaker cognitive model', possibly composed from many visual images, rather than as analogous to a 'view from a point'.

How this 'cognitive model' might have been arrived at in practice or how a system would deal with partial or missing information are not addressed by the corpus at present.

3.2 Bias and information sources

It was suggested above that the corpus should be designed so that it can deliberately be made to contain all of the expressible spatial relationships, along with the other requirements in section 2. This raises the potential problem that the distribution of spatial relationships expressed in the corpus will not be representative of the real world; this could lead to bias when training a spatial communication system. It is uncertain whether any series of real world images, selected by individuals, could in practice be less skewed, however, or indeed whether a child would learn spatial language in a 'representative' environment.

Any effects of this and possible remedies will need to be the subject of further work.

4 DESIGN AND IMPLEMENTATION ISSUES

4.1 Dimension and representation

Many of the systems for spatial language generation have used 2-dimensional images, often represented as bit-maps. It is clear from the requirements however that a 3-dimensional representation will be required and this will be beyond the practical limits of bit-mapping.

A vertex-list representation that is openGL compatible has been adopted at this point. Although it is not entirely optimal, in that it is difficult to avoid line-segment duplication during analysis, it combines ease of visualisation with reasonable simplicity of analysis.

Note this does not preclude the use of 2-dimensional scenes. The representation of maps, in particular, might be a useful addition to the corpus.

Animation of scenes is also required to allow proper mapping on to motion prepositions such as 'through' and 'towards'.

The objects in the scenes as currently envisaged are defined as solid regions of arbitrary complexity immersed in a medium (assumed to be 'air'). They may be convex or concave and may entirely enclose regions of the medium. Care must be taken as currently spaces or parts of the medium cannot be named. If a ball is to go ”through a window” a window must be provided, not simply a gap in a wall.

Typical objects in a scene are constructed from primitives which can be labelled. So although geometrically a table may be treated as a single object a phrase such as ”the ball rolled between the table legs” could be correctly constructed from the information provided.

Typical scenes from the initial corpus, with example description strings, are shown in figures 1 and 2.

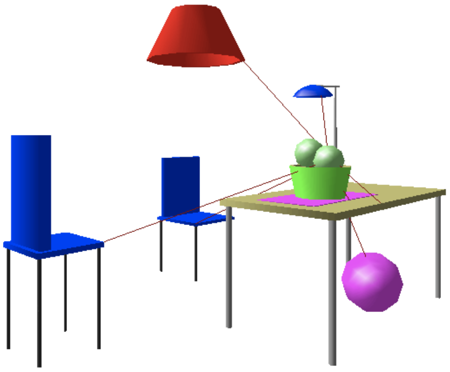


Figure 1. A table-top scene ”the apples are on the table”

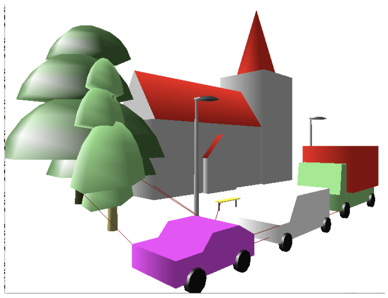


Figure 2. A street-scale scene ”the car is in front of the church”

4.2 Corpus interfaces

The key interfaces between the corpus and associated systems are shown in figure 3. A detailed description of the information presented at these interfaces is not possible here but in summary the XML file defining a scene contains the following:

1. The object list
2. Animation vectors
3. Description strings
4. OpenGL drawing information

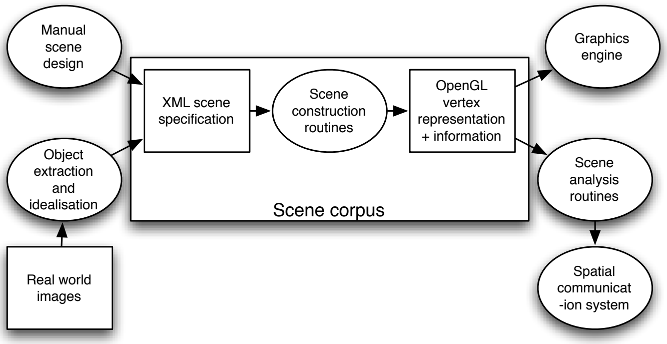

Manual scene design
Object extraction and idealisation
Real world images
XML scene specification
Scene construction routines
OpenGL vertex representation + information
Scene corpus
Graphics engine
Scene analysis routines
Spatial communicat -ion system

Figure 3. Scene corpus interfaces

The description strings contain acceptable spatial communications relating to the scene. These are in English at present but there is no reason why other languages and non-verbal comunications could not be incorporated as well. The XML files could be generated from 'real' images by image processing systems, employing object recognition and vector extraction, to provide extensions to the corpus as initially designed.

The constructed scene passed to the analysis routine contains the OpenGL representation of the object list (as a sequence of animated frames if required) and the description strings.

The designed section of the corpus will be more or less manually constructed.

4.3 Scale space coverage

To ensure coverage of a full range of spatial relationships and mappings on to prepositions in particular, a range of environments needs to be represented in the corpus. The word 'beyond', for instance, is used more frequently in large scale environments than within rooms.

As currently envisaged the corpus will include scenes at 'tabletop','room' 'building' and 'street' scales. Some scenes will need cross-scale representations for reference object selection testing in particular. Extensions to what might be termed 'landscape' scale might be required.

4.4 Complexity and computational load

Construction of the vertex representation of the scenes is relatively trivial, analysing the geometry and topology of the scene is more time-consuming. In the current implementation the complexity of analysis for a scene containing n objects each with m facets is: $O(m^2n^2)$. The most time-consuming aspects of the analysis are creation of a qualitative spatial relation matrix and calculation of closest approach vectors for all objects. Analysing a scene of 8 spheres each composed of 180 facets takes about 0.5 seconds on a 'standard' PC. A 'maximum complexity' scene of 50 objects of 180 facets each would take about 20 seconds and thus a corpus of 1000 of these scenes would require 5.5 hours (without animation). It is thought that this could be improved by a factor of 4 with more adept pruning and a further factor of 4 by attention to the algorithms.

The time taken for the spatial communication system to be trained or to produce the required descriptions is clearly system specific and not included here.

4.5 Corpus size

Considering the potentially diverse nature of systems to be tested it is not entirely glib to say the corpus should be as large as possible. With 70 prepositions, 4 reference frames and 8 complex factors influencing reference choice it is probable that a corpus of less than 1000 scenes would be inadequate even though many relationships can be expressed in a single scene.

5 NEXT STEPS

A corpus for the current research task will be constructed along the lines described but in parallel with this, input from other interested researchers, with a view to constructing a generally useful corpus, would be welcomed.

REFERENCES

[1] G. E. Burnett, D. Smith, and A. J. May, 'Supporting the navigation task: characteristics of good landmarks', in *Proceedings of the Annual Conference of the Ergonomics Society*. Taylor and Francis, (2001).

[2] G. E. Burnett, *'Turn right at the King's Head'. Drivers' requirements for route guidance information.*, Ph.D. dissertation, Loughborough University, 1998.

[3] L. A. Carlson-Radvansky and Logan G. D., 'The influence of reference frame selection on spatial template construction.', *Journal of Memory and Language*, **37**, 411–437, (1997).

[4] K.R. Coventry, 'Spatial prepositions, functional relations, and lexical specification', in *Representation and Processing of Spatial Expressions*, eds., Patrick Olivier and Klaus-Peter Gapp, 1–35, Laurence Earlbaum Associates, (1998).

[5] K. R. Coventry, M. Prat-Sala, and L Richards, 'The interplay between geometry and function in the comprehension of over, under, above, and below.', *Journal of Memory and Language*, **44**(3), 376–398, (2001).

[6] R. Dale and E. Reiter, 'Computational interpretations of the gricean maxims in the generation of referring expressions.', *Cognitive Science*, **19**, 233–263, (1995).

[7] M. I. Feist and D. Gentner, 'Factors involved in the use of in and on', in *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society.*, (2003).

[8] K.P. Gapp, 'An empirically validated model for computing spatial relations', *Künstliche Intelligenz*, 245–256, (1995).

[9] S. Garrod, G. Ferrier, and S. Campbell, 'In and on: investigating the functional geometry of spatial prepositions', *Cognition*, **72**, 167–189, (1999).

[10] A Herskovits, 'Schematization', in *Representation and Processing of Spatial Expressions*, eds., Patrick Olivier and Klaus-Peter Gapp, 149–162, Laurence Earlbaum Associates, (1998).

[11] Tanja Jording and Ipke Wachsmuth., 'An anthropomorphic agent for the use of spatial language.', in *Spatial Language. Cognitive and Computational Perspectives*, eds., K. R. Coventry and P. Olivier, 69–85, Dordrecht: Kluwer Academic Publishers, (2002).

[12] R. Klabunde and R. Porzel, 'Tailoring spatial descriptions to the addressee: a constraint-based approach.', *Linguistics*, **36**(3), 551–577, (1998).

[13] G. Lakoff and Johnson M., *Metaphors we live by*, University of Chicago Press, 1980.

[14] G. Lakoff, *Women, Fire and Dangerous Things*, University of Chicago Press, 1987.

[15] K. Lockwood, K. Forbus, D. Halstead, and J. Usher, 'Automatic categorization of spatial prepositions', *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (2006).

[16] K. Lockwood, K. Forbus, and J. Usher, 'Spacecase: A model of spatial preposition use', in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, (2005).

[17] S. D. Mainwaring, B. Tversky, Mohoto Ohgishy, and D. J. Schiano, 'Descriptions of simple spatial scenes in english and japanese', *Spatial Cognition and Computation*, **3**(1), 3–43, (2003).

[18] C. Nothegger, S. Winter, and M Raubal, 'Computation of the salience of features.', *Spatial Cognition and Computation*, **4**, 113–136, (2004).

[19] R. Porzel, M. Jansche, and R Klabunde, 'The generation of spatial descriptions from a cognitive point of view', in *Spatial Language. Cognitive and Computational Perspectives*, eds., K. R. Coventry and P. Olivier, 185–207, Dordrecht: Kluwer Academic Publishers, (2002).

[20] Terry Regier, *The human semantic potential: Spatial language and constrained connectionism.*, MIT Press, 1996.

[21] T Regier and L. Carlson, 'Grounding spatial language in perception: An empirical and computational investi- gation.', *Journal of Experimental Psychology: General*, **130**(2), 273–298, (2001).

[22] D. K. Roy, 'Learning visually-grounded words and syntax for a scene description task', *Computer Speech and Language*, **16**(3), (2002).

[23] M. Sorrows and S. Hirtle, 'The nature of landmarks for real and electronic spaces', in *Spatial Information Theory: Cognitive and Computational Foundations of GIS*, eds., C. Freska and D. Mark, Springer-Verlag, (1999).

[24] K. van Deemter, I. van der Sluis, and A. Gatt, 'Building a semantically transparent corpus for the generation of referring expressions.', (2006).

Towards a Balanced Corpus of Multimodal Referring Expressions in Dialogue

Ielka van der Sluis [1] and **Paul Piwek** [2] and **Albert Gatt** [3] and **Adrian Bangerter** [4]

Abstract. This paper describes an experiment in which dialogues are elicited through an identification task. Currently we are transcribing the collected data. The primary purpose of the experiment is to test a number of hypotheses regarding both the production and perception of multimodal referring expressions. To achieve this, the experiment was designed such that a number of factors (prior reference, focus of attention, visual attributes and cardinality) were systematically manipulated. We anticipate that the results of the experiment will yield information that can inform the construction of algorithms for the automatic generation of natural and easy-to-understand referring expressions. Moreover, the balanced corpus of multimodal referring expressions that was collected will hopefully become a resource for answering further, as yet unanticipated, questions on the nature of multimodal referring expressions.

1 Introduction

One of the fundamental tasks of Natural Language Generation (NLG) systems is the Generation of Referring Expressions (GRE). Over the past couple of decades, this has been the subject of intensive research [2, 12, 11], and is typically defined as an *identification* problem: given a domain representing entities and their properties, construct a referring expression for a target referent which singles it out from its distractors. While several recent proposals have generalised this problem definition, to deal for example with relations [10, 17], plural referents [26, 13, 14], and vague predicates [27], there has been comparatively little work on the generation of *multimodal* referring acts (but see [18, 23, 28]). Moreover, the majority of contributions have focused on monologue, with interaction between user and NLG system assumed to be absent or limited. Meanwhile, psycholinguistic work is increasingly focusing attention on the conditions governing the use of pointing gestures as part of referring acts in dialogue. Of particular relevance to the questions addressed in this paper is the interaction between the two modalities of *pointing* and *describing* [6, 4, 8, 23, 24].

This paper describes the design of an ongoing experiment on multimodal reference in two-party dialogue. Our aim is to harness the empirical evidence for the design of multimodal GRE algorithms, by studying the corpus of interactions collected in the experiment. The resulting corpus is balanced, in the sense put forward by [15], because the conditions under which references were elicited correspond to experimental variables that are counter-balanced. Moreover, the focus on dialogue permits the investigation to take both a speaker/generator's and a hearer/reader's point of view, with potentially useful data on such factors as alignment and entrainment [7], and the nature of collaboration or negotiation that is a feature of interactive referential communication [9], currently a hotly debated topic in the psycholinguistic literature [22].

Describing vs. pointing Following the influential work in [11], GRE algorithms often take into account the finding that speakers manifest *attribute preferences*, which cause them to overspecify their descriptions. For example, in experiments on reference in visual domains, colour tends to feature in speakers' descriptions irrespective of its discriminatory value, while vague properties like size are relatively dispreferred [21, 5, 3]. On the other hand, recent work on modality choice in reference suggests a potential trade-off between the use of pointing and the amount of information given in a description [28], though the use of pointing also depends on the potential ambiguity of a reference [8] and whether a change of focus is taking place [23]. Our experiment seeks to further this research in four principal directions. First, we look at modality choice as a function of the properties which are available to verbally describe a referent. Thus, if attribute preferences play a role, the possibility of describing a referent using properties like colour may reduce the likelihood of a pointing gesture. Second, we also manipulate the extent to which a referent is in (discourse) focus, that is, whether it was recently mentioned in the dialogue or not. Typically, verbal references to previously mentioned entities tend to be reduced. Does this affect the likelihood of pointing? Third, we look at both singular and plural references, the latter being references to groups of 5 entities. This may increase the visual salience of a referent, which in turn may interact with the other two factors. Finally, we examine to what extent a change of the domain focus (i.e., when the current target is distant from the previous target) affects use of pointing gestures.

Data on these questions will inform the design of multimodal GRE algorithms whose output is 1) *natural*, that is, corresponds closely to what human speakers do in comparable situations, and 2) *easy-to-understand*, i.e., allows the addressee to quickly identify the intended referent without the need for prolonged clarificatory exchanges.

2 The Experiment

2.1 Task and Setup

Figure 1 presents a bird's eye view of the experimental setup in which a director and a follower are talking about a map that is situated on the wall in front of them, henceforth the *shared map*. Both can interact freely using speech and gesture, without touching the shared map or standing up. Each also has a private copy of the map; the

[1] Computing Science, University of Aberdeen, UK
[2] Centre for Research in Computing, The Open University, UK
[3] Computing Science, University of Aberdeen, UK
[4] Institut de Psychologie du Travail et des Organisations, University of Neuchâtel, Switzerland.

director's copy has an itinerary on it, and her task is to communicate the itinerary to the follower. The follower needs to reproduce the itinerary on his private copy. The rules of the experiment were as follows:

- Since this is a conversation, the follower is free to interrupt the director and ask for any clarification s/he thinks is necessary.
- Both participants are free to indicate landmarks or parts of the map to their partner in any way they like.
- Both participants are not permitted to show their partner their private map at any point. They can only discuss the shared map.
- Both participants must remain seated throughout the experiment.

While this task resembles the MapTask experiments ([1]), the latter manipulated mismatches between features on the director and follower map, phonological properties of feature labels on maps, familiarity of participants with each other and eye contact between participants[5]. The current experiment systematically manipulates target size, colour, cardinality, prior reference and domain focus, in a balanced design. Though this arguably leads to a certain degree of artificiality in the conversational setting, the balance would not be easy to obtain in an uncontrolled setting or with off-the-shelf materials like real maps. Further properties of our experiment that distinguish it from the MapTask are: (1) objects in the visual domains are not named, so that participants need to produce their own referring expressions, (2) the participants are always able to see each other; (3) the participants are allowed to include pointing gestures in their referring expressions (a MapTask type experiment that does include non-verbal behaviour, in particular, eye gaze, is reported in [20]).

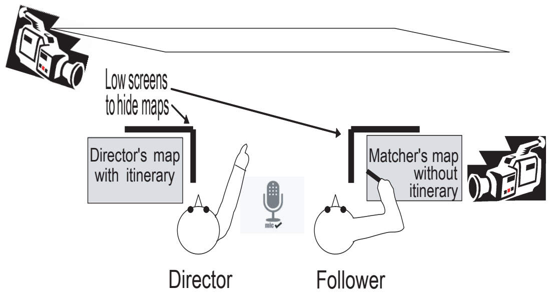


Figure 1. Bird's-eye view of the experiment setup.

2.2 Materials and Independent Variables

Four maps were constructed, consisting of simple geometrical landmarks (ovals or squares). Two of the maps (one each for ovals and squares) have *group* landmarks, whereas the other two have singletons. Objects differ in their size (large, medium, small) and colour (red, blue, green). Each dyad in the experiment discusses all four maps. Per dyad, the participants switch director/follower roles after each map. The order in which dyads discuss maps is counter balanced acrosss dyads. There are four independent variables in this experiment:

- **Cardinality** The target destinations in the itineraries are either singleton sets or sets of 5 objects that have the same attributes (e.g., all green squares)
- **Visual Attributes:** Targets on the itinerary differ from their distractors – the objects in their immediate vicinity (the 'focus area') – in colour, or in size, or in both colour and size. The focus area is defined as the set of objects immediately surrounding a target.
- **Prior reference:** Some of the targets are visited twice in the itinerary.
- **Shift of domain focus:** Targets are located near to or far away from the previous target. If two targets t_1 and t_2 are in the *near* condition, then t_1 is one of the distractors of t_2 and vice versa.

The overall set up of this experiment is illustrated in Table 2.3:

I	near	new	size
II	near	new	colour
III	near	new	both
IV	near	old	size
V	near	old	colour
VI	near	old	both
VII	far	new	size
VIII	far	new	colour
IX	far	new	both
X	far	old	size
XI	far	old	colour
XII	far	old	both

Table 1. Overview of the experimental design. Each different type of target corresponds to a Roman numeral. The types are a function of *Focus* × *Prior* reference × *Attributes*, yielding a total of $2 \times 2 \times 3 = 12$ types of targets that appear on each map. Moreover, half of the maps is populated only with singleton targets, whereas the other half is populated with target sets whose cardinality is 5. Taking this into account, overall we have $2 \times 12 = 24$ types of target in our experiment.

2.3 Current Status and Annotation Plans

Using the maps described in Section 2.2, a pilot of the experiment was carried out in Aberdeen (see Figure 3 for an impression). The pilot led to a few minor adjustments in the setup (e.g., we moved from a projected to a printed shared map), and subsequently data was collected from 24 dyads with the validated setup. Currently, the data is being transcribed, See Figure 2 for an example.

128	D	Uh and if you *go straight up* from that you've got five blue ones	D points at the map and moves his finger upwards
129	F	Yeah [*there*?]	D is still pointing F points
130	D	[There] yeah	D is still pointing F is still pointing
131	F	one two three four five	D is still pointing M is still pointing
132	D	Yeah. They're all number three	D is still pointing
133	F	Right. Right.	
134	D	And the five reds just *to the right over*	D points and moves his finger to the right
135	F	And like a kind of *downwards* arrow	D is still pointing F moves his hand upwards
136	D	Arrow yeah they're all number four. Number five. Uh and five is paired with one *with these ones*.	D stops pointing D points
137	F	All right.	

Figure 2. Excerpt from dialogue O17-S33-S34, where *D* = director, *F* = follower and where the brackets indicate overlapping speech and the text in italics indicates approximately the co-duration of gesture and speech

Our next task will be to annotate the data. For this, we will build on existing guidelines and best practice, e.g., use of stand-off XML, for annotation of multimodal data (see [19, 16]). Our main annotation tasks will be: identification of multimodal referring expressions, linking of referring expressions with domain objects (i.e., intended referents) and segmentation of dialogue into episodes spanning the point in time from initiation to successful completion of a target identification.

[5] http://www.hcrc.ed.ac.uk/maptask/maptask-description.html

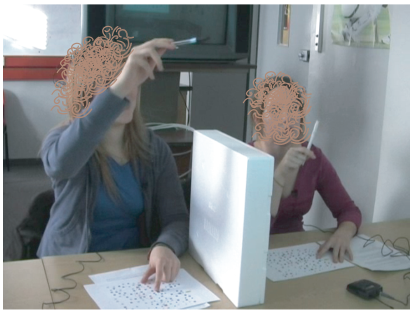


Figure 3. Two participants at work in the pilot of the experiment.

3 Research Questions and Hypotheses

3.1 Production: The Director

The main distinctive feature of the current experiment is that we rigorously controlled for a significant number of features of the referents. The experimental design allows us to both address new questions, and validate existing findings from previous observational studies that were made in more natural and less controlled settings. For example, in an observational study [23] found that some speakers used a lot of gestural information, while others did not at all. The current study will help us to answer the question whether such different styles and strategies are tied to particular features of the communicative situation, or are really a result of individual differences. Other findings include whether speakers use extensive pointing gestures or keep their gestures close to their body depends on the communicative function of the message they want to get across (c.f. [8]). Also, the linguistic information that speakers use varies considerably depending on how difficult it is to describe an object as a function of the number of relevant attributes [25]. In addition, speakers display different approaches in conveying the distinguishing properties of an object to the addressee. For instance, in the world map study discussed in [28] speakers used different strategies to indicate a country. Some used prominent objects on the map and others used the map itself as a point of reference, some used the visible properties of the objects (e.g. size, color, shape) and others traveled through landscapes, politics and economics. In the current experiment, the following **research questions**, some new and some closely related to the aforementioned conjectures and findings, will be addressed:

- **Use of Pointing Gestures:** When are pointing gestures used and in which cases are they omitted. How do the duration of the pointing gesture and the extension of the pointing device (in this case a human arm) relate to the object that is indicated?
- **Use of Linguistic Material:** Are the linguistic descriptions minimal, underspecified or overspecified? What information (e.g. preferred, absolute or relative attributes) is included in the description?
- **Interaction of Pointing Gestures and Linguistic Material** How is linguistic and gestural information combined in multimodal referring expressions? What linguistic information is left out or added if a pointing gesture is included in the referring expression?
- **Speaker's Strategies:** Which strategies for referring to objects are used (e.g. describing targets by their global position on the map, or in relation to other salient targets etc.)? How does the speaker relate a target description to the dialogue context? Are speakers consistent in their use and composition of referring expressions throughout the dialogue (e.g. entrainment)? Do they adapt their strategies to the addressee?

The experiment will directly address the following **hypotheses on production**, where we denote a target referent as t_n, where n represents the order in which the targets are referred to:

- If t_1 and t_2 are far away from each other, a reference to t_2 is more likely to include a pointing gesture, compared to the case where t_1 and t_2 are near.
- If t_1 and t_2 are far away from each other, a description of t_2 is expected to include more linguistic information compared to the case where they are near.
- If t is discourse-old, then there is less likelihood of a pointing gesture, compared to the case where t has not been referred to earlier. The amplitude of such pointing gestures is expected to be smaller.
- If t is discourse-old, then a description is expected to include less linguistic information compared to a discourse-new reference.
- If t is distinguishable only by *size* (a dispreferred property), then the descriptions is likely to include more linguistic and gestural material than descriptions of targets that are distinguishable by their colour.
- A referring expression for identifying a singleton is expected to include more linguistic and gestural material than a referring expression for identifying a target group.

3.2 Perception: The Follower

In addition to the production perspective, our experiment will also shed new light on the interpretation of multimodal referring expressions. We are particularly interested in the conditions that influence whether and how quickly an addressee successfully interpreted a referring expression. One way to measure successful reference is to take as indicative the point when the interlocutors move on from one target to the next in an itinerary. This allows one to count the number of turns or measure the time it takes from the first reference to a target to the first reference to the next target in the itinerary; the shorter the time, or the number of turns needed for identification, the easier the identification. There is, however, a danger that such a way of measuring success overestimates the time it takes to arrive at an identification, since this identification will always take place prior to moving to the next target. Moreover, how can we know that the addressee has actually identified the correct target? In our experiment this problem is addressed because we ask the follower to indicate the itinerary on his private map. Thus, the use of a camera that tracks the status of the follower's map (see Figure 1), might enable us to get a better estimate of when identification of the target takes place. In summary, our experiment will help us explore features that facilitate easy identification of targets,[6] and this will involve the following **research questions**:

- **Use of Pointing Gestures:** Are targets more easily identified, when a referring expression includes a pointing gesture? What effects does the amplitude (e.g. duration, extension) of the pointing gesture have on identification of the target by the addressee?

[6] Note that the value of these features may differ per person

- **Use of Linguistic Material:** Are targets more easily identified when the linguistic descriptions are minimal, underspecified or overspecified? Does it matter which information (e.g. preferred, absolute or relative attributes) is included or left out in the description?
- **Interaction of Pointing Gestures and Linguistic Material** What linguistic information is best combined with pointing gestures to facilitate identification? What linguistic information can be left out when a pointing gesture is included in the referring expression?
- **Addressee's Strategies:** How does the addressee check for success? By repeating or rephrasing the information that is provided by the speaker (alignment of speech, gesture or both?), or by adding extra material (e.g. relata, properties), or otherwise?

Hypotheses on perception that will be tested with this experiment:

- Target groups (consisting of 5 objects with the same features) are more easy to identify than single targets (need less time and less extensive identification by the director).
- Targets that have a prior reference in the dialogue are more easy to identify.
- Targets that are located near to the previous target are more easy to identify than targets that are located far away from the previous target.
- Ease of recognition is expected to be related to the visual attributes of the targets: Targets that differ in color and size $\leq$ Targets that differ only in color $\leq$ targets that differ only in size from their distractors.

4 Conclusion

In order to build language generation systems that produce natural and effective multimodal behaviour, a deep understanding is needed of the way human speakers choose what to say and gesture, and the impact of their choices on the hearer's ability to understand the message. This requires corpora of human–human dialogue which are annotated not just with information on the linguistic and non-linguistic realization of the speakers' utterances and non-verbal behaviour, but which also lay bare the underlying communicative situation, including the attributes of the objects that speakers refer to, and provide information on success or failure of communicative acts. The current paper reports on an effort to produce such a corpus, focussing on multimodal referring expressions. Though it is intended primarily to address a number of specific hypotheses on production and perception of multimodal referring expressions, we are also taking care to package it as a resource that might prove useful for the exploration of yet unanticipated research questions on multimodal behaviour.

Acknowledgements This work was carried out under the EPSRC platform grant 'Affecting people with natural language' (EP/E011764/1).

REFERENCES

[1] A. Anderson, M. Bader, E. Bard, E. Boyle, G.M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, and R. Weinert, 'The HCRC Map Task Corpus', *Language and Speech*, **34**, 351–366, (1991).

[2] D. Appelt, 'Planning english referring expressions.', *Artificial Intelligence*, **26**(1), 1–33, (1985).

[3] A. Arts, *Overspecification in Instructive Texts*, Ph.D. dissertation, Univiersity of Tilburg, 2004.

[4] A. Bangerter, 'Using pointing and describing to achieve joint focus of attention in dialogue.', *Psychological Science*, **15**, 415–419, (2004).

[5] E. Belke and A. Meyer, 'Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions', *European Journal of Cognitive Psychology*, **14**(2), 237–266, (2002).

[6] R.J. Beun and A. Cremers, 'Multimodal reference to objects: An empirical approach', in *Proceedings of the Conference on Cooperative Multimodal Communication (CLC 1998)*, pp. 64–88, (1998).

[7] S. Brennan and H.H. Clark, 'Conceptual pacts and lexical choice in conversation', *Journal of Experimental Psychology*, **22**(6), 1482–1493, (1996).

[8] A. Bangerter & E. Chevalley, 'Pointing and describing in referential communication: When are pointing gestures used to communicate?', in *Proceedings of the workshop on multimodal output generation (MOG 2007)*, (2007).

[9] H.H. Clark and D. Wilkes-Gibbs, 'Referring as a collaborative process', *Cognition*, **22**, 1–39, (1986).

[10] R. Dale and N. Haddock, 'Generating referring expressions containing relations', in *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, (1991).

[11] R. Dale and E. Reiter, 'Computational interpretation of the Gricean maxims in the generation of referring expressions', *Cognitive Science*, **19**(8), 233–263, (1995).

[12] Robert Dale, 'Cooking up referring expressions.', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89*, (1989).

[13] C. Gardent, 'Generating minimal definite descriptions', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02*, (2002).

[14] A. Gatt and K. van Deemter, 'Incremental generation of plural descriptions: Similarity and partitioning', in *Proceedings of EMNLP'07*, (2007).

[15] A. Gatt, I. van der Sluis, and K. van Deemter, 'Evaluating algorithms for the generation of referring expressions using a balanced corpus', in *Proceedings of ENLG'07*, (2007).

[16] ISLE Natural Interactivity and Multimodality Working Group, 'Guidelines for the Creation of NIMM Data Resources', Technical Report D8.2, IST-1999-10647 Project, (2003).

[17] J. D. Kelleher and G-J Kruijff, 'Incremental generation of spatial referring expressions in situated dialog.', in *Proceedings of the joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL/COLING-06*, (2006).

[18] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, 'Deictic object reference in task-oriented dialogue', in *Situated Communication*, eds., G. Rickheit and I. Wachsmuth, 155–208, Mouton de Gruiter, (2006).

[19] P. Kühnlein and J. Stegmann, 'Empirical Issues in Deictic Gestures: Referring to Objects in Simple Identification Tasks', Technical Report 2003/3, SFB 360, Univ. Bielefeld, (2003).

[20] M. Louwerse, P. Jeuniaux, M. Hoqueand J. Wu, and G. Lewis, 'Multimodal communication in computer-mediated map task scenarios', in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (2006).

[21] Thomas Pechmann, 'Incremental speech production and referential overspecification', *Linguistics*, **27**, 89–110, (1989).

[22] M. Pickering and S. Garrod, 'Toward a Mechanistic Psychology of Dialogue', *Behavioural and Brain Sciences*, **27**(2), 169–226, (2004).

[23] P. Piwek, 'Modality choice for generation of referring acts: Pointing versus describing', in *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*, (2007).

[24] P. Piwek, R.J. Beun, and A. Cremers, ''Proximal' and 'Distal' in language and cognition: evidence from deictic demonstratives in Dutch', *Journal of Pragmatics*, (June 2007). doi: 10.1016/j.pragma.2007.05.001.

[25] I. van der Sluis, A. Gatt, and K. van Deemter, 'Evaluating algorithms for the generation of referring expressions: Going beyond toy domains', in *Proceedings of RANLP'07*, (2007).

[26] K. van Deemter, 'Generating referring expressions: Boolean extensions of the incremental algorithm', *Computational Linguistics*, **28**(1), 37–52, (2002).

[27] K. van Deemter, 'Generating referring expressions that involve gradable properties', *Computational Linguistics*, **32**(2), 195–222, (2006).

[28] I. van der Sluis and E. Krahmer, 'Generating multimodal referring expressions', *Discourse Processes*, **44**(3), 145–174, (2007).

Illustrating Answers: An Evaluation of Automatically Retrieved Illustrations of Answers to Medical Questions

Wauter Bosma and **Mariët Theune**[1]
Charlotte van Hooijdonk and **Emiel Krahmer** and **Fons Maes**[2]

Abstract. In this paper we discuss and evaluate a method for automatic text illustration, applied to answers to medical questions. Our method for selecting illustrations is based on the idea that similarities between the answers and picture-related text (the picture's caption or the section/paragraph that includes the picture) can be used as evidence that the picture would be appropriate to illustrate the answer. In a user study, participants rated answer presentations consisting of a textual component and a picture. The textual component was a manually written reference answer; the picture was automatically retrieved by measuring the similarity between the text and either the picture's caption or its section. The caption-based selection method resulted in more attractive presentations than the section-based method; the caption-based method was also more consistent in selecting informative pictures and showed a greater correlation between user-rated informativeness and the confidence of relevance of the system. When compared to manually selected pictures, we found that automatically selected pictures were rated similarly to decorative pictures, but worse than informative pictures.

1 INTRODUCTION

According to Mayer's [11] well-known *multimedia principle,* people learn better from words and pictures than from words alone. Nevertheless most question-answering (QA) systems, which can automatically answer users' questions that are posed in natural language, still present their answers using a single modality, in the form of text snippets retrieved from a document corpus. Any pictures occurring in the documents are generally ignored, since the text-oriented retrieval methods used in QA systems cannot deal with them. A solution for dealing with non-textual media that has been proposed for use in multimedia summarization and retrieval is to analyze and convert the media content to a semantic representation usable by the system [10, 12, 6, 13]. However, automatic analysis of media content is difficult and often unreliable, while manual annotation is very laborious. Another solution, which according to de Jong et al. [9] is often overlooked, is the use of related linguistic content instead of the media items themselves. If related text adequately describes a media item, text-based retrieval methods can be used to retrieve non-textual media.

Bosma [3] proposed a method for extending the answers returned by a QA-system with appropriate illustrations by searching pictures whose related text is similar to the text of the answer. Pictures are selected by taking the best match of the answer text and a text snippet automatically associated with the picture. This method has been applied in the IMIX system for answering medical questions [5]. The purpose of the IMIX system is to answer medical questions from non-expert users, of the kind to which answers can be typically found in an encyclopedia. Questions can be typed or spoken (in Dutch), and answers are presented using speech, text and pictures. Questions can be asked in isolation, but the system is also capable of engaging in dialogs and answer follow-up questions.

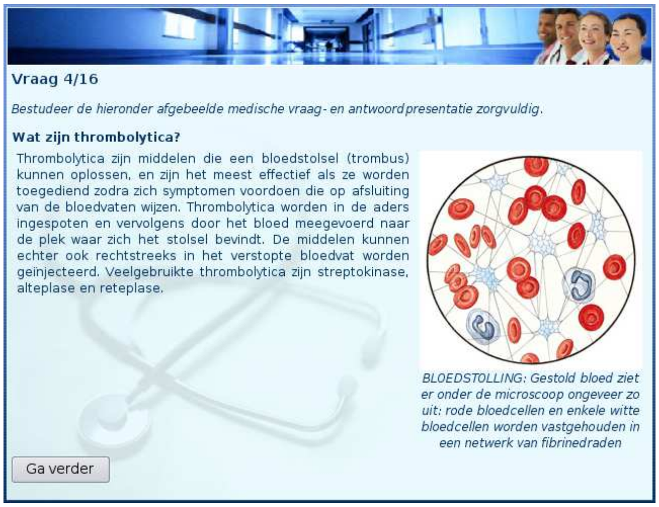
This paper presents a user evaluation of Bosma's [3] picture selection method. In the experiment, answer presentations with automatically selected pictures were rated by naive participants judging the attractiveness and informativeness of the text-picture combination. We also investigated the influence of the different presentations on learning. The experimental design was the same as that used by van Hooijdonk et al. [8], who evaluated manually created answer presentations consisting of different text-picture combinations. We repeated their experiment for answer presentations with automatically retrieved pictures, comparing two versions of the automatic picture retrieval method: one where the picture's textual annotation consists of its caption (resulting in 'caption-selected' illustrations), and one where the annotation is a part of the text near which the picture was found (resulting in 'section-selected' illustrations).

In the following sections, we first explain the picture selection method that is evaluated (Section 2). Then we describe the set-up of the evaluation experiment (Section 3) followed by a discussion of the results (Section 4). We end with some concluding remarks (Section 5).

2 AUTOMATIC TEXT ILLUSTRATION

Our picture selection method is an application of the query-based summarization framework of [4], which is applied in IMIX to generate extended answers consisting of a paragraph-sized text. In QA, the answer's content is drawn from a set of documents (the source documents) which provide an answer but were not necessarily written to answer the query. The query-based summarization approach relies on a combination of one or more feature graphs representing the source documents. The graphs express relations between the documents' content units, and are constructed using information about unit content (e.g. based on cosine similarity) or context (e.g. based on layout) to relate the units. This way, content can be presented

[1] University of Twente, The Netherlands, email: W.E.Bosma@utwente.nl, M.Theune@utwente.nl
[2] Tilburg University, The Netherlands, email: C.M.J.vanHooijdonk@uvt.nl, E.J.Krahmer@uvt.nl, Maes@uvt.nl



Vraag 4/16

Bestudeer de hieronder afgebeelde medische vraag- en antwoordpresentatie zorgvuldig.

Wat zijn thrombolytica?

Thrombolytica zijn middelen die een bloedstolsel (trombus) kunnen oplossen, en zijn het meest effectief als ze worden toegediend zodra zich symptomen voordoen die op afsluiting van de bloedvaten wijzen. Thrombolytica worden in de aders ingespoten en vervolgens door het bloed meegevoerd naar de plek waar zich het stolsel bevindt. De middelen kunnen echter ook rechtstreeks in het verstopte bloedvat worden geïnjecteerd. Veelgebruikte thrombolytica zijn streptokinase, alteplase en reteplase.

BLOEDSTOLLING: Gestold bloed ziet er onder de microscoop ongeveer zo uit: rode bloedcellen en enkele witte bloedcellen worden vastgehouden in een netwerk van fibrinedraden

Figure 1. Example of an answer presentation consisting of text and an automatically selected picture. The presentation answers the question *What are thrombolytics?* The text of the answer explains that thrombolytics are drugs used to dissolve blood clots. The picture depicts a schematic representation of clotted blood.

for which there is just indirect evidence of relevance. For instance, a sentence that is adjacent – and thus contextually related– to a sentence that is similar to the query may be included in the answer, even though it is only indirectly linked to the query.

This concept may also be applied to multimedia. A picture can be related to a piece of text by using layout information. A straightforward relatedness clue of text and picture is when the text is the picture's caption, but also if the picture belongs to a certain paragraph or section, the section and the picture may be considered related. When the relevance of the text is established, the relevance of the picture is established indirectly.

In the IMIX system, this approach is used to select the best picture to illustrate a given textual answer to a medical question. To find this picture, the illustration system compares the text of the answer with picture-associated text. The more similar the two text passages, the more likely the picture is relevant. The picture-associated text is interpreted as a textual representation of the picture. This may be either the picture's caption or the paragraph (or section if no single paragraph could be related to the picture) in which the picture was found. The relevancy of a picture for the answer is calculated as:

$$R_{picture}(i,t) = cosim(t, text(i)) \tag{1}$$

where $R_{picture}(i,t)$ is the relevancy of picture i to text t; and $text(i)$ is the text associated with picture i. The function $cosim(a,b)$ calculates the cosine similarity of a and b.

Cosine similarity is a way of determining lexical similarity of text passages. The idea behind cosine similarity is that a text's meaning is constituted by the meaning of its words. To measure cosine similarity between two passages, we represent both texts as a vector whose elements represent the contribution of a word to the meaning of the passage. Before measuring the cosine similarity, words are stemmed using Porter's stemmer [14]. The cosine similarity is calculated as follows:

$$cosim(a,b) = \frac{\sum_{k=1}^{n} a_k \cdot b_k}{|a| \cdot |b|} \tag{2}$$

where $cosim(a,b)$ is the similarity of passages a and b; n is the number of distinct words in the passages. Both passages are represented as a vector of length n, with a_k representing the contribution of word k to passage a. The denominator ensures that passage vectors are normalized by their lengths. The value $|a|$ is the length of passage vector a, measured as $\sqrt{\sum_{k=1}^{n} a_k^2}$.

Determining how much a particular word contributes to the meaning of a passage is called *term weighting*. In this paper, we use $tf \cdot idf$ term weighting, i.e. the contribution of a word to a passage is calculated as the word's occurrence frequency in the passage (term frequency, TF) multiplied by the word's inverse document frequency (IDF). IDF is a measure of how characteristic the word is for a passage. To measure the inverse document frequency, we require a large set of passages. In this paper, we use the passage vectors of picture-associated text for all pictures in the corpus, plus the passage vector of the answer text. A word occurring in few of these passages receives a high IDF value, because the low occurrence rate makes it descriptive of the few passages it appears in. Conversely, a word occurring in many passages receives a low IDF value. The contribution of word k to passage a is measured as follows:

$$a_k = tf_{a,k} \cdot idf_k \tag{3}$$

where $tf_{a,k}$ is the number of occurrences of word k in passage a; and idf_k is the IDF value of word k. The IDF value is calculated as follows:

$$idf_k = log \frac{|D|}{|\{d \mid d \in D \wedge k \in d\}|} \tag{4}$$

where $|D|$ is the number of passages in the corpus (i.e. the number of pictures plus one); and the denominator is the number of documents which contain the word k.

The final answer presentation consists of the textual answer and the most relevant picture and its caption. An example of an answer presentation containing an automatically selected picture is given in Figure 1 (this is a screen shot showing one of the answer presentations from our experiment, see the next section).

3 THE EVALUATION EXPERIMENT

We carried out an evaluation experiment in which participants evaluated a set of 16 text-picture answer presentations to medical questions. The pictures in the presentations were selected automatically using the method described above. Apart from the pictures used in the answer presentations, the study was identical to the study of manually created presentations by van Hooijdonk et al. [8]. This includes the textual component of the answers. Below we describe the creation of the stimuli used in the experiment, the participants and the experimental procedure.

3.1 Questions and textual answers

In our study, we used the same set of 16 general medical questions that had been used by [8]. Certain properties of the questions in this set were systematically varied, in order to investigate the effect of question type on the effect of the different answer presentations. Of the 16 questions, half were definition questions and half were procedural questions. Of the eight questions in both groups, half referred to body parts and half did not. Table 1 shows examples of the questions used. References to body parts may be indirect, as is the case in the first question in Table 1.

Table 1. Examples of medical questions. Questions are equally divided in the categories of *definition questions* (Def.) or *procedure questions* (Proc.); and in questions which refer to body parts and questions which do not.

Type/Bodypart	Question
Def./Yes	Where is testosterone produced?
Def./No	What does ADHD stand for?
Proc./Yes	How to apply a sling to the left arm?
Proc./No	How to organize a workspace in order to prevent RSI?

For each medical question, van Hooijdonk et al. [8] formulated a concise and an extended textual answer. A concise answer gives a direct answer to the question, and nothing more, while the extended answer also provides relevant background information (c.f. [2]). The average lengths of the concise answers and the extended answers were approximately 26 words and 66 words respectively.

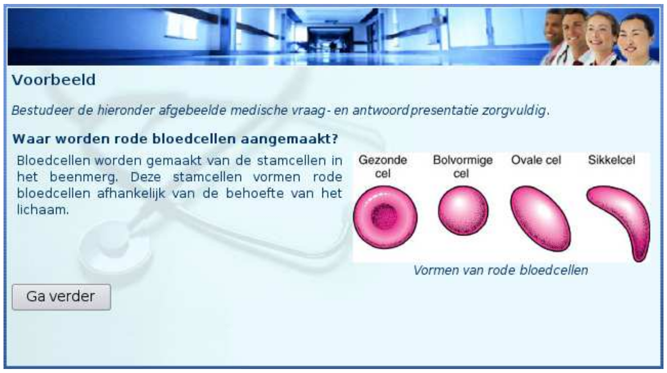
The textual component of an answer presentation was a manually written reference answer. Manual text is used in order to be able to concentrate on evaluating the multimedia aspect – the quality of the text-picture combination. In the experiment reported here, we reused the answer texts from [8] but combined them with new, automatically selected pictures as described below. The text is based on answers produced by a study in which participants answered the same questions as the ones used here, using any information available, including web search. This procedure is described in detail in [8].

3.2 Illustrating the answers

We created a corpus of annotated pictures to be used for automatically illustrating the textual answers. The pictures as well as their textual annotations were automatically extracted from two medical sources providing information about anatomy, processes, diseases, treatment and diagnosis. Both are intended for a general audience and written in Dutch. The first source, *Merck Manual medisch handboek* [1], Merck in short, contains 188 schematic illustrations of anatomy and treatment, process schema's, plots and various types of diagrams. The other source, *Winkler Prins medische encyclopedie* [7], WP in short, contains a variety of 421 pictures, including photographic pictures, schema's and diagrams. These sources were selected because they cover the popular medical domain and they are relatively structured – paragraph boundaries are marked in the text and all 609 pictures have captions.

In this experiment, for each of the textual answers, two presentations were generated by illustrating them using the algorithm described in section 2, applied to the picture corpus described above. For one of the presentations for each answer, the picture's caption was used as associated text. For the other presentation the picture was associated with the smallest unit of surrounding text from its original document; this could be a section or a paragraph. The surrounding text was extracted automatically, using meta-information in the document such as XML tags.

The average distribution of selected pictures from our two sources (Merck 33 percent; WP 66 percent) reflects the distribution in our picture corpus (Merck 31 percent; WP 69 percent). Table 2 lists the number of selected pictures from each source for the four selected conditions, with percentages given between brackets. Note that for each condition, 17 pictures were selected: 16 for the answer presentations to be evaluated, plus one for an example presentation that was presented to the participants (see Section 3.4).



Voorbeeld

Bestudeer de hieronder afgebeelde medische vraag- en antwoordpresentatie zorgvuldig.

Waar worden rode bloedcellen aangemaakt?

Bloedcellen worden gemaakt van de stamcellen in het beenmerg. Deze stamcellen vormen rode bloedcellen afhankelijk van de behoefte van het lichaam.

Figure 2. Example of a picture which is related but not complementary to the answer text. The presentation answers the question *Where are red blood cells generated?* The text explains that red blood cells are generated from stem cells in the bone marrow. Rather than illustrating this, however, the picture shows various deformations of red blood cells.

Table 2. Number of pictures (with percentages in brackets) selected from Merck [1] and WP [7].

Condition	Merck	WP
Brief text; caption-selected picture	6 (35%)	11 (65%)
Extended text; caption-selected picture	4 (24%)	13 (76%)
Brief text; section-selected picture	6 (35%)	11 (65%)
Extended text; section-selected picture	7 (41%)	10 (59%)

The corpus did not contain an appropriate picture for all answers, which forced the illustration system to select less appropriate pictures for some of the presentations. In some cases the selected picture was plain irrelevant, but in some other cases, the picture was related to the text but had a different perspective. For instance, the picture in Figure 2 addresses the deformation of red blood cells rather than their generation. This problem may have been augmented by the fact that the pictures in our corpus have a high information density; only few pictures have a decorative function only (i.e., they do not add any information to the related text). Consequently, the pictures are relatively specific to their original context, which complicates their reuse in a slightly different context.

The answer presentations were created as a web page headed by the question (in bold face), followed by the answer text on the left and the best-matching picture on the right side of the page. Regardless which method had been used to select the picture (caption-based or section-based), we considered the caption part of the picture and thus presented it along with the picture in the answer presentation. Since all pictures in our corpus had a caption, this was always included. If the text surrounding the picture had been used for its selection, this text was not included in the answer presentation.

A complicating factor here was that captions vary greatly in length, especially in the WP corpus. Table 3 shows details of the distribution of caption lengths (for comparison, details about section lengths are given in Table 4). The most extreme case was a caption as long as 428 words. Since the textual component of the answer presentations averaged only 26 or 66 words (for concise and extended presentations respectively), presenting very long captions along with the pictures would lead to an imbalance between the amount of text in the caption and the amount of text in the textual component of the answer. In order to prevent excessive caption lengths, in the answer presentations the captions were truncated to their first sentence. So only the caption's first sentence was presented along with the picture, rather than the caption as a whole. This was done *after* picture selection, so it did not affect the picture selection process.

Table 3. Caption length statistics of the Merck corpus [1] and the WP corpus [7].

	Caption length (words)	
	Average	SD
Merck	4.4	1.9
WP	39.1	42.9
Combined	28.4	39.1

Table 4. Section length statistics of the Merck corpus [1] and the WP corpus [7].

	Section length (words)		
	Average	SD	range
Merck	354	325	[30,1944]
WP	67	48	[5,336]
Combined	156	227	[5,1944]

3.3 Participants

Seventy five people participated in the experiment: 44 female and 31 male, between 18 and 55 years old. Fifty six of them (75 percent) were students recruited from Tilburg University. The remaining 25 percent were recruited from various e-mail lists. None had participated in the experiments of [8]. The participants were randomly assigned to one of the four conditions (concise or extended text, selection by means of caption or surrounding text), of which they were shown all 16 answer presentations.

3.4 Experimental procedure

The participants were invited by e-mail to participate. This e-mail shortly stated the goal of the experiment, the amount of time it would take to participate, the possibility to win a gift certificate, and the URL of the experiment. The experiment, created using WWStim [15], was entirely online.

When the participants accessed the experiment, they first received instructions about the procedure. The participants were told that they would receive the answer presentations of 16 medical questions, which they would have to study carefully and then assess their informativeness and their attractiveness. Next, the participants entered their personal data, i.e., age, gender, level of education, and optionally their e-mail to win a gift certificate.

After participants had filled out their personal data, they practiced the procedure of the actual experiment in a practice session: they were given the medical question *Where are red blood cells produced?*. First, the participants answered on a seven-point Likert scale how confident they were to know the answer to this medical question. Subsequently, the participants were shown the answer to the medical question corresponding to the condition they were assigned to. (See Figure 2 for the concise-answer, caption-selected picture condition.) The participants studied the answer presentation until they thought that they could assess its informativeness and attractiveness. Then, the participants were shown the medical question, the answer presentation, and a questionnaire. This questionnaire consisted of five questions, asking them to rate on a seven-point Likert scale:

1. the clarity of the text;
2. the informativeness of the answer presentation;
3. the attractiveness of the answer presentation;
4. the informativeness of the combination of text and picture;
5. the attractiveness of the combination of text and picture.

The participants judged the informativeness of the text-picture combination instead of directly assessing the relevance of the picture. This is because the experiment in [8] contained manually selected pictures only, for which relevance was assumed (although a distinction was made between decorative and informative pictures). In contrast, automatic pictures may be irrelevant or somewhat relevant. However, we chose not to change the design of the experiment in order to get comparable results. (See Section 4.3 for a comparison between presentations with manually and automatically selected pictures.)

After completing the practice session, the participants started with the actual experiment, proceeding in the same way as during the practice session. When they were finished with their assessment of the answer presentations to the 16 medical questions, the participants received a post test which was the same for all participants (regardless the experimental condition). In the post test, the participants had to answer the same 16 questions of which they had rated the answer presentations in the previous part of the experiment. This was done in the form of a multiple choice test, in which each medical question was provided with four textual answer possibilities. Of these four answer possibilities, one answer was correct and the other three were plausible incorrect ones. The order in which the medical questions were presented in the post test was the same as in the actual experiment. Note that – with respect to the concise textual answer – the additional information in the extended textual answers and in the pictures was not necessary to answer the question in the post test correctly.

4 RESULTS

The results of the assessments were normalized to be in the range $[0..1]$. A rating n between one and seven (inclusive) was normalized as $\frac{1}{6}(n-1)$.

For processing the results, we used the following, non-standard method. For each condition and each medical question and assessment question, we calculated the average assessment. For pair-wise significance testing of differences between two experimental conditions for a particular assessment question, we measured the percentage of answer presentations for which the rating of one condition was higher than that of another. A condition that consistently received higher average ratings than the other for each medical question got a score of 100 percent; consequently, the other condition got a relative score of 0 percent. Significance is tested by means of 10^6-fold approximate randomization. A difference is considered significant if the null hypothesis (that the sets are not different) could be rejected at a certainty greater than 95 percent ($p < 0.05$), unless stated otherwise.

The reasons for using the mutual rank instead of the average judgment are as follows. To see if one type of answer presentation is better than another, one could simply check whether the difference in average scores is significant. However, while a single average score is useful as a rough quality indication, it may not be the best method for a pairwise comparison.

If the difference in scores between two types of answer presentation does not tell anything about the difference in quality other than which one is better, a comparison can have only three possible outcomes: one is better, the other is better, or their quality is equal. If this is accepted, it remains to be seen whether the score averages are reliable for significance testing. The standard deviation of ratings of answers to some medical questions was higher than the standard deviation for answers to other medical questions. As a result, some

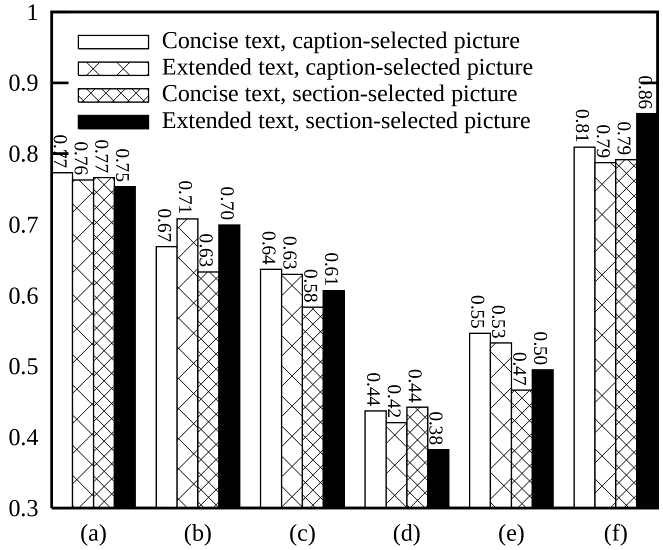


Figure 3. Average assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test.

medical questions affect the average rating more than others. This makes it less likely to find significant differences in rating. Using the mutual rank avoids this problem.

4.1 Caption or section?

Figure 3 shows an overview of the average assessments per condition. The level of clarity of the textual component of the answer (Figure 3 (a)) was judged similar. No significant differences between any two conditions were found.

Regarding the informativeness of the answer presentation as a whole (Figure 3 (b)), extended answers were rated significantly more informative than concise answers. However, for extended answers, the combination of picture and text (Figure 3 (d)) was judged less informative. This effect was the strongest for pictures that were selected using their surrounding section, although the differences were not significant.

The presentation (Figure 3 (c)) as well as the picture/text combination (Figure 3 (e)) was rated significantly more attractive if the pictures were selected based on their captions than if they were selected based on their surrounding section. The attractiveness of the presentation or the picture/text combination was not affected by the length of the textual component of the answer.

All in all, the presentations containing a section-selected picture were less informative and less attractive than the presentations containing a caption-selected picture. Apparently, captions are more representative of the content of a picture, and thus are more reliable indicators of the picture's relevance to the answer text. This is not entirely surprising, as the content of a caption generally describes (only) the picture, whereas the text surrounding a picture may also contain unrelated content.

In seeming contradiction with the good ratings of caption-selected pictures, in the post test where participants had to select the correct answer in a multiple choice test, participants who were shown section-selected pictures gave significantly more correct answers than other participants when the section-selected picture was included in a presentation with an extended textual component. This is a remarkable result because these pictures were rated least informative. A possible explanation for this is that the participants concentrated less on the picture (because they quickly dismissed it as less relevant) and more on the text. After all, the information in the picture was not required to answer the questions in the post test.

4.2 The value of confidence

The selection criterion for automatic pictures was the cosine similarity of the textual component of the answer and the text associated with the picture (a caption or a section, depending on the condition). The picture with the highest cosine similarity was selected. Because cosine similarity is used as a measure of relevance, this value can be interpreted as a *confidence value*, i.e. how confident the system is that the selected picture is actually relevant. If the cosine similarity is actually a good indicator of relevance, one would expect a high correlation between cosine similarity and relevance. In the IMIX system, in which this picture selection method is implemented, the answer is presented text-only if no picture has a confidence (cosine similarity) above a certain (configurable) threshold. Table 5 shows the averages of the cosine similarity values of the pictures selected for the answers in the experiment described in this paper.

Table 5. Statistics of the cosine similarity of the textual component of the answer and the text passage used for indexing the selected picture.

Condition	Average (standard deviation)	
Brief text; caption-selected picture	0.190	(0.00788)
Extended text; caption-selected picture	0.188	(0.00631)
Brief text; section-selected picture	0.133	(0.00501)
Extended text; section-selected picture	0.162	(0.00654)

But what is the meaning of cosine similarity as a confidence value? Cosine similarity can be used to predict the relevance of the picture if there is a correlation between the cosine similarity and the experimental participants' judgments of a presentation. Figure 4 shows the correlation of the confidence (cosine similarity) value and the participant judgments. A value of 1 (or -1) indicates a perfect increasing (or decreasing) linear correlation. This correlation was greatest for the participant judgments of the informativeness of the text-picture combination (0.51 and 0.44 with concise and extended text respectively). This is an encouraging result, given that this aspect seems to correspond most closely to picture relevance. With respect to attractiveness, the correlation with confidence was significantly greater for concise answers than for extended answers. There was only a slight difference in correlation between attractiveness and confidence for different picture selection methods.

Remarkably, participants perceived the textual component of the answer as less clear when the confidence value of the picture was greater. This puzzling result suggests that relevant pictures negatively affect the clarity of the answer text rather than enhance it. A possible explanation is that any mismatches between picture and text may be more confusing when text and picture seem closely related than when the picture obviously does not fit the text, in which case it can be easily ignored and does not influence the interpretation of the text.

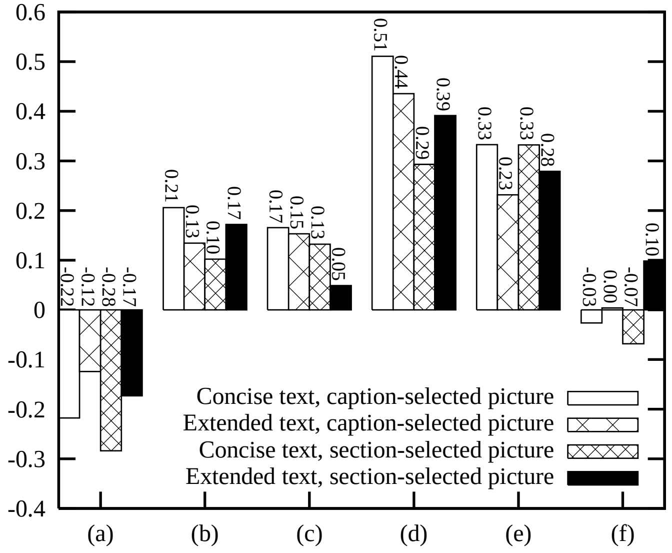


Figure 4. Pearson correlation coefficient between the confidence of picture selection and the assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test.

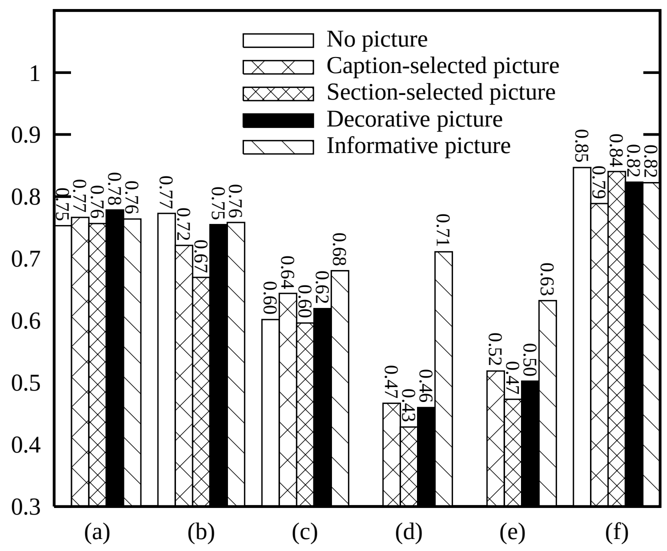


Figure 5. Average assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test. For comparability, these results include only registered students from Tilburg University. Therefore, the actual values may differ slightly from Figure 3.

4.3 Automatic or manual?

As mentioned earlier, apart from the answer presentations themselves, the design of the experiment was identical to the experiment described in [8]. This allows us to compare the evaluation results of our automatically illustrated answer presentations to those of [8], who evaluated manually created answer presentations.

In the experiment of [8], the answer presentations consisted of the same (concise or extended) textual component used in the current experiment, in combination with either no picture, a decorative picture, or an informative picture (i.e. six experimental conditions in total). These manually selected pictures can be regarded as a *gold standard* for decorative and informative pictures respectively. However, in practice, it is unlikely that this gold standard can be achieved with the set of 609 medical pictures used for automatic picture selection in our experiment, because the picture sources used by [8] were unrestricted and thus offered far more opportunities to find a suitable illustration for a given answer text.

A large portion of participants in both experiments were students from Tilburg University. Because these students received course credits for participation, they filled in their student registration number, which made it possible to distinguish them from other participants. However, in both experiments, other participants took part from outside this community, and we found significant differences between the registered students and the other participants with respect to their answers to some of the assessment questions. On average, for 65 percent ($p < 0.001$) of the answer presentations of [8], the informativeness of the presentation was rated higher by student participants than by other participants. In the same experiment, students rated the text-picture combinations more informative (60 percent, $p < 0.001$) and less attractive (58 percent, $p < 0.01$) than other participants. The answers to other assessment questions were similar for both groups, or slightly different.

The results of two experiments are comparable only if the group of participants in one experiment is similar to the participants of the other experiment. However, students and non-students are shown to produce different results, rendering the participant groups as a whole dissimilar. Therefore we filtered the non-students out from our comparison between automatically and manually selected illustrations, to ensure that the experimental conditions are the only variables over both experiments. Since the students participating in both experiments were recruited within a short time frame using the same communication channels, we consider both groups as fully comparable.

In total, 98 participants (70 female, 28 male) in both experiments were registered students. Of them, 42 contributed to the experimental conditions of [8] and 56 contributed to the conditions from our experiment, described in section 3. No one participated twice. The average assessments of the 98 participants are shown in Figure 5. These results combine the 16 concise and the 16 extended answer presentations, comprising 32 data points for each condition and assessment question.

The informativeness of text-picture combinations as well as the attractiveness of the presentation was similar when the answer contained an automatically selected picture, a manually selected decorative picture, or no picture at all. No significant differences were found. However, the text-picture combination of manually selected informative pictures was rated significantly more informative than the text-picture combination of manually selected decorative pictures and automatically selected pictures. Answer presentations were rated significantly less informative if the presentation contained a section-selected picture than if the answer contained an informative picture, a decorative picture, or no picture at all. Presentations containing caption-selected pictures are not significantly less informative than presentations with informative pictures.

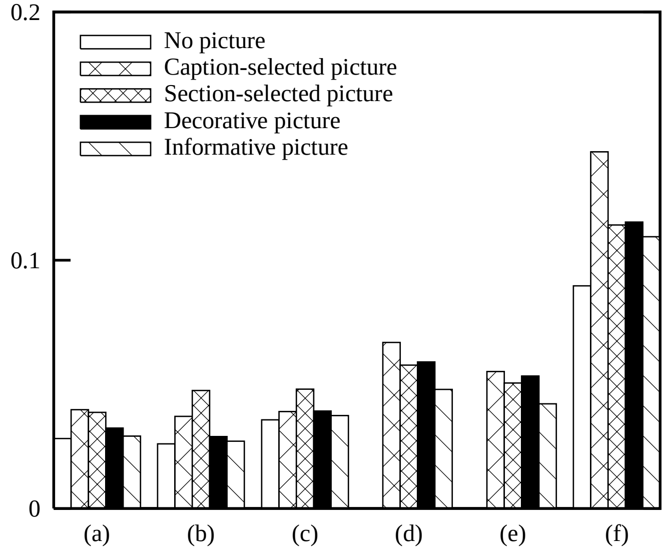


Figure 6. Standard deviations per answer presentation in the assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test. For comparability, these results include only registered students from Tilburg University.

Average ratings of automatic presentations may have been negatively affected by inconsistent performance of the picture selection algorithm. In some cases, the algorithm selected an irrelevant or a somewhat irrelevant picture because there was no appropriate picture in the database or simply because the algorithm failed to find it. If the relevance of automatic pictures is less consistent than that of manual pictures, this should reflect in the variability of the results. Figure 6 shows the standard deviations of assessments. For automatic pictures, participants indeed show greater variability than for manual pictures in their assessments of textual clarity, informativeness and attractiveness of the answer presentation. Remarkably, we found that the standard deviation of the number of correct answers in the post test was also greater for pictures which are selected by their captions.

5 CONCLUSION

This paper presented an algorithm for automatic illustration of answers to medical questions in Dutch. It is used in the IMIX question answering system to add appropriate illustrations to textual answers. To evaluate the algorithm, we conducted an experiment, following the same procedure as [8] to evaluate different types of answer presentations on informativeness, attractiveness and influence on learning.

In our experiment, the answer presentations contained a textual and a visual component, of which the text was given and the visual was automatically retrieved from an offline picture database containing 609 pictures. The pictures were automatically extracted from *Merck Manual medisch handboek* [1] and from *Winkler Prins medische encyclopedie* [7]. To find an appropriate picture, the pictures were indexed by a passage of text from the document in which they were found. Two different indexing methods were compared in the experiment, either using the picture's caption for picture selection, or using the section or paragraph that contained the picture. Both selection methods were tested in combination with a concise or an extended textual answer.

Due to limitations of the corpus (i.e. for several questions it did not contain a relevant picture at all) the standard deviations of our results are quite high, which makes it difficult to make any general claims based on them. However, some tentative conclusions can be drawn.

The results indicate that the caption-based picture selection method results in more informative and attractive presentations than the section-based method, although the difference in informativeness was not significant. Furthermore, caption-based picture selection shows a greater correlation between confidence and informativeness, which indicates that the confidence value better predicts the informativeness of the picture. A system could use this to respond by not offering any picture if no relevant picture is available (as is currently done in the IMIX system). All in all, the caption-based picture selection method offers more promising results than the section-based selection method.

An investigation of the relation between system confidence and our experimental results revealed an intriguing negative correlation between textual clarity and the predicted relevance of the selected illustration. Apparently, seeing an answer text in combination with a picture that is related to it, but not fully attuned to it, may be confusing to the user. Problems like these might be solved by the development of post-processing methods to adapt the textual and visual components of the answer presentation to each other, so that they form a more coherent whole.

When compared to manually created answer presentations, we found that answer presentations with an automatically selected picture were largely rated at the same level as presentations with a manually selected decorative picture (which did not add any information to the answer) or even no picture at all. This may be partially explained by the design of the experiment, where the visual element of the answer presentations was not needed to answer the question (since the textual element contained all the required information). Also, the results were undoubtedly influenced by the fact that our picture corpus did not contain appropriate pictures for all answers, in which case the algorithm had no choice but to select an irrelevant picture. To measure the extent of this influence, we should perform a sub-analysis on those questions for which the corpus did contain at least one appropriate picture. In general, we can say that, given the limitations of our corpus, achieving comparable ratings to manually selected decorative pictures is not a bad result.

ACKNOWLEDGEMENTS

This research was performed in the IMOGEN (Interactive Multimodal Output GENeration) project. IMOGEN is a project within the Netherlands Organisation for Scientific Research (NWO) research programme on Interactive Multimodal Information eXtraction (IMIX). The authors would like to thank Jurry de Vos and Lennard van de Laar for their help in setting up the experiment.

REFERENCES

[1] *Merck manual medisch handboek*, eds., Robert Berkow, Mark H. Beers, and Andrew J. Fletcher, Bohn Stafleu van Loghum, Houten, the Netherlands, 2nd edn., 2005.

[2] Wauter Bosma, 'Extending answers using discourse structure', in *Crossing Barriers in Text Summarization Research*, eds., Horacio Saggion and Jean-Luc Minel, pp. 2–9, Shoumen, Bulgaria, (September 2005). Incoma Ltd.

[3] Wauter Bosma, ‘Image retrieval supports multimedia authoring’, in *Linguistic Engineering meets Cognitive Engineering in Multimodal Systems*, eds., E.V. Zudilova-Seinstra and T. Adriaansen, ICMI Workshop, pp. 89–94, Trento, Italy, (October 2005). ITC-irst.

[4] Wauter Bosma, ‘Query-based summarization for question answering’, in *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*, eds., Ton van der Wouden, Michaela Poß, Hilke Reckman, and Crit Cremers, number 4 in LOT Occasional Series, pp. 29–44, Utrecht, the Netherlands, (2005). LOT.

[5] Lou Boves and Els den Os, ‘Interactivity and multimodality in the IMIX demonstrator’, in *IEEE International Conference on Multimedia and Expo*, pp. 1578–1581, Amsterdam, (July 2005). IEEE Computer Society.

[6] Kees van Deemter and Richard Power, ‘High-level authoring of illustrated documents’, *Natural Language Engineering*, **2**(9), 101–126, (2003).

[7] *Winkler Prins medische encyclopedie*, eds., Peter Fiedeldij Dop and Simon Vermeent, Spectrum, 3rd edn., 1974.

[8] Charlotte van Hooijdonk, Jurry de Vos, Emiel Krahmer, Alfons Maes, Mariët Theune, and Wauter Bosma, ‘On the role of visuals in multimodal answers to medical questions’, in *Proceedings of the 2007 Conference of the IEEE Professional Communication Society*. IEEE, (2007).

[9] Franciska de Jong, Thijs Westerveld, and Arjen de Vries, ‘Multimedia search without visual analysis: the value of linguistic and contextual information’, *IEEE Transactions on Circuits and Systems for Video Technology*, **17**(3), 365–371, (2007).

[10] Mark Maybury and A. E. Merlino, ‘Multimedia summaries of broadcast news’, in *1997 IASTED International Conference on Intelligent Information Systems*. IEEE, (1997).

[11] Richard Mayer, *The Cambridge handbook of multimedia learning*, Cambridge University Press, Cambridge, 2005.

[12] Katashi Nagao, Shigeki Ohira, and Mitsuhiro Yoneoka, ‘Annotation-based multimedia summarization and translation’, in *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7, Morristown, NJ, USA, (2002). Association for Computational Linguistics.

[13] Valery Petrushin, *Introduction into Multimedia Data Mining and Knowledge Discovery*, 3–13, Springer London, 2007.

[14] M.F. Porter, ‘An algorithm for suffix stripping’, *Readings in information retrieval*, 313–316, (1997).

[15] Theo Veenker. WWStim: A CGI script for presenting webbased questionnaires and experiments, 2005. Website: http://www.let.uu.nl/Theo.Veenker/personal/projects/wwstim/doc/en/.

Simulation-based Learning of Optimal Multimodal Presentation Strategies from Wizard-of-Oz data

Verena Rieser and **Oliver Lemon**[1]

Abstract. We address two problems in the field of automatic optimization of dialogue strategies: learning effective dialogue strategies when no initial data or system exists, and optimising dialogue management (DM) and Natural Language Generation (NLG) decisions in an integrated fashion. We use Reinforcement Learning (RL) to learn multimodal information presentation strategies through interaction with a simulated environment which is "bootstrapped" from small amounts of Wizard-of-Oz (WOZ) data. This use of WOZ data allows development of optimal strategies for domains where no working prototype is available. For information seeking dialogues, Dialogue Management and NLG are two closely interrelated problems: the decision of *when* to present information depends on the available options for *how* to present them, and vice versa. We therefore formulate the problem as a hierarchy of joint learning decisions which are optimised together. To evaluate, we compare the RL-based strategy against a supervised learning (SL) strategy which mimics the (human) wizards' policies from the original data. This comparison allows us to measure relative improvement over the training data. Our results show that RL significantly outperforms SL: the RL-based policy gains on average 50-times more reward when tested in simulation. In related work we evaluate the strategies with real users [16].

1 Introduction

One of the key advantages of statistical optimisation methods (e.g. Reinforcement Learning (RL)) for dialogue strategy design is that the problem can be formulated as a precise mathematical model which can be trained on real data [5]. In cases where a system is designed from scratch, however, there is often no suitable in-domain data. Collecting dialogue data without a working prototype is problematic, leaving the developer with a classic chicken-and-egg problem.

Here, we learn dialogue strategies by simulation-based RL [20], where the simulated environment is learned from small amounts of Wizard-of-Oz (WOZ) data. Using WOZ data rather than data from real Human-Computer Interaction (HCI) allows us to learn optimal strategies for domains where no working dialogue system exists. To date, automatic strategy learning has been applied to dialogue systems which have already been deployed in the real world using handcrafted strategies. In such work, strategy learning was performed based on already present extensive online-operation experience, e.g. [19, 3]. In contrast to this preceding work, our approach enables strategy learning in domains where no prior system is available. Optimised learned strategies are then available from the first moment of online-operation, and tedious handcrafting of dialogue strategies is avoided. This independence from large amounts of in-domain dialogue data allows researchers to apply RL to new application areas beyond the scope of existing dialogue systems. We call this method 'bootstrapping'.

The use of WOZ data has earlier been proposed in the context of RL. [23] utilise WOZ data to discover the state and action space for MDP design. [9] use WOZ data to build a simulated user and noise model for simulation-based RL. While both studies show promising first results, their simulated environments still contain many handcrafted aspects, which makes it hard to evaluate whether the success of the learned strategy indeed originates from the WOZ data. In addition, [17] propose to 'bootstrap' with a simulated user which is entirely hand-crafted. In the following we propose an entirely data-driven approach, where all components of the simulated learning environment are learned from WOZ data.

2 Wizard-of-Oz data collection

Our domains of interest are information-seeking dialogues, for example a multimodal in-car interface to a large database of MP3 files. The corpus we use for learning was collected in a multimodal study of German task-oriented dialogues for an in-car music player application by [12]. This study provides insights into natural methods of information presentation as performed by human wizards. 6 people played the role of an intelligent interface (the "wizards"). The wizards were able to speak freely and display search results on the screen by clicking on pre-computed templates. Wizards' outputs were not restricted, in order to explore the different ways they intuitively chose to present search results. Wizard's utterances were immediately transcribed and played back to the user with Text-To-Speech. 21 subjects (11 female, 10 male) were given a set of predefined tasks to perform, as well as a primary driving task, using a driving simulator. The users were able to speak, as well as make selections on the screen. Please see [12] for further detail.

The corpus gathered with this setup comprises 21 sessions and over 16K turns. Example 1 shows a typical multimodal presentation sub-dialogue (translated from German). Note that the wizard displays quite a long list of possible candidates on an (average sized) computer screen, while the user is driving. This example illustrates that even for humans it is difficult to find an "optimal" solution to the problem we are trying to solve.

(1) **User:** Please search for music by Madonna .

Wizard: I found seventeen hundred and eleven items. The items are displayed on the screen. *[displays list]*

User: Please select 'Secret'.

For each session information was logged, e.g. the transcriptions of the spoken utterances, the wizard's database query and the number of results, the screen option chosen by the wizard, and a rich set of contextual dialogue features was also annotated, see [12]. Of the 793 wizard turns 22.3% were annotated as presentation strategies, result-

[1] School of Informatics, University of Edinburgh, UK, email: {vrieser,olemon}@inf.ed.ac.uk

ing in 177 instances for learning, where the six wizards contributed about equal proportions.

Information about user preferences was obtained, using a questionnaire containing similar questions to the PARADISE study [22]. In general, users report that they get distracted from driving if too much information is presented. On the other hand, users prefer shorter dialogues (most of the user ratings are negatively related with dialogue length). These results indicate that we need to find a strategy given the competing trade-offs between the number of results (large lists are difficult for users to process), the length of the dialogue (long dialogues are tiring, but collecting more information can result in more precise results), and the noise in the speech recognition environment (in high noise conditions accurate information is difficult to obtain). In the following we utilise the ratings from the user questionnaires to optimise a presentation strategy using simulation-based RL.

3 Simulated Learning Environment

Simulation-based RL (aka model-free RL) learns by interaction with a simulated environment. We obtain the simulated components from the WOZ corpus using data-driven methods. The employed database contains 438 items and is similar in retrieval ambiguity and structure to the one used in the WOZ experiment. The dialogue system used for learning comprises some low level constraints reflecting the system logic (e.g. that only filled slots can be confirmed), implemented as Information State Update (ISU) rules. The higher level actions are left for optimisation.

3.1 MDP and problem representation

The structure of an information seeking dialogue system consists of an information acquisition phase, and an information presentation phase. For information acquisition the task of the dialogue manager is to gather 'enough' search constraints from the user, and then, 'at the right time', to start the information presentation phase where the Natural Language Generation task is to present 'the right amount' of information – either on the screen or listing the items verbally. What this actually means depends on the application, the dialogue context, and the preferences of users. For optimising dialogue strategies information acquisition and presentation are two closely interrelated problems and need to be optimised simultaneously: *when* to present information depends on the available options for *how* to present them, and vice versa. We therefore formulate the problem as a Markov Decision Process (MDP), relating states to actions in a hierarchical manner (see Figure 1): 4 actions are available for the information acquisition phase; once the action `presentInfo` is chosen, the information presentation phase is entered, where 2 different actions for output realisation are available. The state-space comprises 8 binary features representing the task for a 4 slot problem: `filledSlot` indicates whether a slots is filled, `confirmedSlot` indicates whether a slot is confirmed. We also add features human wizards pay attention to, using the feature selection techniques of [14]. Our results indicate that wizards only pay attention to the number of retrieved items (`DB`). We therefore add the feature `DB` to the state space, which takes integer values between 1 and 438, resulting in $\underline{2^8 \times 438 = 112,128}$ distinct dialogue states. In total there are $\underline{4^{112,128}}$ theoretically possible policies for information acquisition. [2] For the presentation phase the `DB` feature is discretised, as we will further discuss in Section 3.6. For the information presentation phase there are $\underline{2^{2^3} = 256}$ theoretically possible policies.

3.2 Supervised Baseline

We create a baseline by applying Supervised Learning (SL). This baseline mimics the average wizard behaviour and allows us to measure the relative improvements over the training data (cf. [3]). For our experiments we use the WEKA toolkit [24]. We learn with the decision tree J4.8 classifier, WEKA's implementation of the C4.5 system [10], and rule induction JRIP, the WEKA implementation of RIPPER [1]. We learn models which predict the following wizard actions:

- Presentation timing: *when* the 'average' wizard starts the presentation phase
- Presentation modality: in *which modality* the list is presented.

	baseline	JRip	J48
timing	52.0(±2.2)	50.2(±9.7)	53.5(±11.7)
modality	51.0(±7.0)	93.5(±11.5)*	94.6(±10.0)*

Table 1. Predicted accuracy for presentation timing and modality (with standard devaition ±), * statistically significant improvement, $p < .05$

As input features we use annotated dialogue context features, see [14]. Both models are trained using 10-fold cross validation. Table 1 presents the results for comparing the accuracy of the learned classifiers against the majority baseline. For presentation timing, none of the classifiers produces significantly improved results. Hence, we conclude that there is no distinctive pattern the wizards follow for *when* to present information. For strategy implementation we scale back to a frequency-based approach following the distribution in the WOZ data: in 0.48 of the times the baseline policy decides to present the retrieved items; for the rest of the time the system follows a hand-coded strategy. For learning presentation modality, both classifiers significantly outperform the baseline. The learned models can be rewritten as in Algorithm 1. Note that this rather simple algorithm is meant to represent the average strategy as present in the initial data (which then allows us to measure the relative improvements of the RL-based strategy).

Algorithm 1 $SupervisedStrategy$

```
if $DB \leq 3$ then
    return presentInfoVerbal
else
    return presentInfoMM
end if
```

3.3 Noise simulation

One of the fundamental characteristics of HCI is an error prone communication channel. Therefore, the simulation of channel noise is an important aspect of the learning environment. Previous work uses data-intensive simulations of ASR errors, e.g. [8]. We use a simple model simulating the effects of non- and misunderstanding on the interaction, rather than the noise itself. This method is especially suited to learning from small data sets. From our data we estimate a 30% chance of user utterances to be misunderstood, and 4% to be complete non-understandings. We simulate the effects noise has on the user behaviour, as well as for the task accuracy. For the user side, the noise model defines the likelihood of the user accepting or rejecting the system's hypothesis, i.e. in 30% of the cases the user rejects, in 70% the user agrees. These probabilities are combined with the probabilities for user actions from the user simulation, as described in the next section. For non-understandings we have the user simulation generating Out-of-Vocabulary utterances with a chance of 4%.

[2] In practise, the policy space is smaller, as some of combinations are not possible, e.g. a slot cannot be confirmed before being filled. Furthermore, some action choices are excluded by the basic system logic.

$$\left[\begin{array}{llll} \textit{acquisition action:} & \begin{bmatrix}\texttt{askASlot}\\ \texttt{implConfAskASlot}\\ \texttt{explConf}\\ \texttt{presentInfo}\end{bmatrix} & \textit{state:} & \begin{bmatrix}\texttt{filledSlot 1|2|3|4|:}\left\{\texttt{0,1}\right\}\\ \texttt{confirmedSlot 1|2|3|4|:}\left\{\texttt{0,1}\right\}\\ \texttt{DB:}\left\{\texttt{1--438}\right\}\end{bmatrix}\\ \textit{presentation action:} & \begin{bmatrix}\texttt{presentInfoVerbal}\\ \texttt{presentInfoMM}\end{bmatrix} & \textit{state:} & \begin{bmatrix}\texttt{DB low:}\left\{\texttt{0,1}\right\}\\ \texttt{DB med:}\left\{\texttt{0,1}\right\}\\ \texttt{DB high }\left\{\texttt{0,1}\right\}\end{bmatrix}\end{array}\right]$$

Figure 1. State-Action space for hierarchical Reinforcement Learning

Furthermore, the noise model determines the likelihood of task accuracy as calculated in the reward function for learning. A filled slot which is not confirmed by the user has a 30% chance of having been mis-recognised.

3.4 User simulation

A user simulation is a predictive model of real user behaviour used for automatic dialogue strategy development. For our domain, the user can either add information (`add`), repeat or paraphrase information which was already provided at an earlier stage (`repeat`), give a simple yes-no answer (`y/n`), or change to a different topic by providing a different slot value than the one asked for (`change`). These actions are annotated manually ($\kappa = .7$). We build two different types of user simulations, one is used for strategy training, one for testing. Both are simple bi-gram models which predict the next user action based on the previous system action ($P(a_{user}|a_{system})$). We face the problem of learning such models when training data is sparse. For training, we therefore use a cluster-based user simulation method, see [13]. For testing, we apply smoothing to the bi-gram model. The simulations are evaluated using the SUPER metric proposed earlier [13],which measures variance and consistency of the simulated behaviour with respect to the observed behaviour in the original data set. This technique is used because for training we need more variance to facilitate the exploration of large state-action spaces, whereas for testing we need simulations which are more realistic. Both user simulations significantly outperform random and majority class baselines.

3.5 Reward modelling

The reward function defines the goal of the overall dialogue. For example, if it is most important for the dialogue to be efficient, the reward penalises dialogue length, while rewarding task success. In most previous work the reward function is manually set, which makes it "the most hand-crafted aspect" of RL [7]. In contrast, we learn the reward model from data, using a modified version of the PARADISE framework [22], following pioneering work by [21]. In PARADISE multiple linear regression is used to build a predictive model of subjective user ratings (from questionnaires) from objective dialogue performance measures (such as dialogue length). We use PARADISE to predict Task Ease (a variable obtained by taking the average of two user ratings from the questionnaire) from various input variables, via stepwise regression. The chosen model comprises dialogue length in turns, task completion (as manually annotated in the WOZ data), and the multimodal user score from the user questionnaire, as shown in Equation 2.

$$\begin{aligned} TaskEase = & -\mathbf{20.2} * dialogueLength + \\ \mathbf{11.8} * taskCompletion & + \mathbf{8.7} * multimodalScore; \end{aligned} \tag{2}$$

This equation is used to calculate the overall reward for the information acquisition phase. During learning, Task Completion is calculated online according to the noise model, penalising all slots which are filled but not confirmed.

For the information presentation phase, we compute a local reward. We relate the multimodal score (a variable obtained by taking the average of 4 user ratings from the questionnaire) to the number of items presented (DB) for each modality, using curve fitting. In contrast to linear regression, curve fitting does not assume a linear inductive bias, but it selects the most likely model (given the data points) by function interpolation. The resulting models are shown in Figure 3.5. The reward for multimodal presentation is a quadratic function that assigns a maximal score to a strategy displaying 14.8 items (curve inflection point). The reward for verbal presentation is a linear function assigning negative scores to all presented items ≤ 4. The reward functions for information presentation intersect at no. items=3.

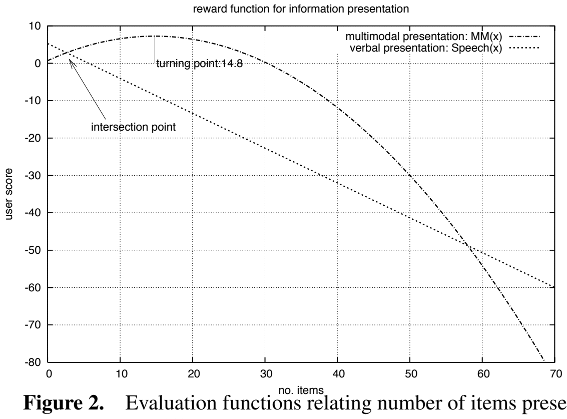


Figure 2. Evaluation functions relating number of items presented in different modalities to multimodal score

3.6 State space discretisation

We use linear function approximation in order to learn with large state-action spaces. Linear function approximation learns linear estimates for expected reward values of actions in states represented as feature vectors. This is inconsistent with the idea of non-linear reward functions (as introduced in the last section). We therefore quantise the state space for information presentation. We partition the database feature into 3 bins, taking the first intersection point between verbal and multimodal reward and the turning point of the multimodal function as discretisation boundaries. Previous work on learning with large databases commonly quantises the database feature in order to learn with large state spaces using manual heuristics, e.g. [6, 2]. Our quantisation technique is more principled as it reflects user preferences for multi-modal output. Furthermore, in previous work database items were not only quantised in the state-space, but also in the reward function, resulting in a direct mapping between quantised retrieved items and discrete reward values, whereas our reward function still operates on the continuous values. In addition, the decision *when* to present a list (information acquisition phase) is still based on continuous DB values. In future work we plan to engineer new state features in order to learn with non-linear rewards while the state space is still continuous. A continuous representation of the state space allows learning of more fine-grained local trade-offs between the parameters, as demonstrated by [15].

3.7 Testing the Learned Policies in Simulation

We now train and test the multimodal presentation strategies by interacting with the simulated learning environment. For the following RL experiments we used the REALL-DUDE toolkit of [4]. The SHARSHA algorithm is employed for training, which adds hierarchical structure to the well known SARSA algorithm [18]. The policy is trained with the cluster-based user simulation over 180k system cycles, which results in about 20k simulated dialogues. In total, the learned strategy has 371 distinct state-action pairs (see [11] for details).

We test the RL-based and supervised baseline policies by running 500 test dialogues with the smoothed user simulation. We then compare quantitative dialogue measures performing a paired t-test (with pair-wise exclusion of missing values). In particular, we compare mean values of the final rewards, number of filled and confirmed slots, dialog length, and items presented multimodally (`MM items`) and items presented verbally. (`verbal items`). RL performs significantly better ($p < .001$) than the baseline strategy. The only non-significant difference is the number of items presented verbally, where both RL and SL strategy settled on a threshold of less than 4 items. The mean performance measures for simulation-based testing are shown in Table 2. The major strength of the learned policy is that it learns to keep the dialogues reasonably short by presenting lists as soon as the number of retrieved items is within tolerance range for the respective modality (as reflected in the reward function). The SL strategy in contrast has not learned the right timing nor an upper bound for displaying items on the screen. The results show that simulation-based RL with an environment bootstrapped from WOZ data allows learning of robust strategies which significantly outperform the strategies contained in the initial data set. In contrast to SL, programming by reward allows us to provide additional information in the reward function, and therefore enables learning a policy which reflects the user preferences. In related work we evaluate the learned strategy with real users (see [11, 16]).

	SL baseline	**RL strategy**
reward	-1747.3 (±527.6)	37.3 (±54.5)***
dialog length	8.7 (±3.7)	6.3 (±3.1)***
verbal items	1.1 (±.28)	1.0 (±.31)
MM items	59.78 (±74.2)	11.5 (±2.2)***

Table 2. Comparison of mean performance for SL and RL policies (with standard deviation ±); *** denotes statistical significance at $p < .001$

4 Conclusion

We addressed two problems in the field of automatic optimization of dialogue strategies: learning effective dialogue strategies when no initial data or system exists, and optimising DM and NLG decisions in an integrated fashion. We used a simulated environment which is "bootstrapped" from small amounts of WOZ data, thus allowing strategy optimization for domains where no working prototype is available. We compared the RL-based strategy against a supervised strategy which mimics the human wizards' policy from the original data. Our results show that RL significantly outperforms Supervised Learning: the RL-based policy gains on average 50-times more reward when tested in simulation. In related work we evaluate the learned strategy with real users [16].

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (CLASSiC project `www.classic-project.org`) and from the EPSRC, project no. EP/E019501/1 and from the ITRG Saarland University.

REFERENCES

[1] W. W. Cohen, 'Fast effective rule induction', in *Proc. of the 12th ICML-95*, (1995).

[2] P. Heeman, 'Combining reinforcement learning with information-state update rules.', in *Proc. of NAACL*, pp. 268–275, (2007).

[3] J. Henderson, O. Lemon, and K. Georgila, 'Hybrid Reinforcement/Supervised Learning for Dialogue Policies from COMMUNICATOR data', in *IJCAI workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp. 68–75, (2005).

[4] O. Lemon, X. Liu, D. Shapiro, and C. Tollander, 'Hierarchical reinforcement learning of dialogue policies in a development environment for dialogue systems: REALL-DUDE', in *BRANDIAL*, (2006).

[5] O. Lemon and O. Pietquin, 'Machine learning for spoken dialogue systems', in *Proc. of Interspeech*, (2007).

[6] E. Levin, R. Pieraccini, and W. Eckert, 'A stochastic model of human-machine interaction for learning dialog strategies', *IEEE Transactions on Speech and Audio Processing*, **8**(1), (2000).

[7] T. Paek, 'Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment', in *Proc. Dialog-on-Dialog Workshop, Interspeech*, (2006).

[8] O. Pietquin and T. Dutoit, 'A probabilistic framework for dialog simulation and optimal strategy learnin', *IEEE Transactions on Audio, Speech and Language Processing*, **14**(2), 589–599, (2006).

[9] T. Prommer, H. Holzapfel, and A. Waibel, 'Rapid simulation-driven reinforcement learning of multimodal dialog strategies in human-robot interaction', in *Proc. of Interspeech/ICSLP*, (2006).

[10] R. Quinlan, *C4.5: Programs for Machine Learning.*, Morgan Kaufmann, 1993.

[11] V. Rieser, *Bootstrapping Reinforcement Learning-based Dialogue Strategies from Wizard-of-Oz data (to appear)*, Ph.D. dissertation, Saarland University, 2008.

[12] V. Rieser, I. Kruijff-Korbayová, and O. Lemon, 'A corpus collection and annotation framework for learning multimodal clarification strategies', in *Proc. of the 6th SIGdial Workshop*, (2005).

[13] V. Rieser and O. Lemon, 'Cluster-based user simulations for learning dialogue strategies', in *Proc. of Interspeech/ICSLP*, (2006).

[14] V. Rieser and O. Lemon, 'Using machine learning to explore human multimodal clarification strategies', in *Proc. of ACL*, (2006).

[15] V. Rieser and O. Lemon, 'Does this list contain what you were searching for? learning adaptive dialogue strategies for interactive question answering', *JNLE (special issue on Interactive Question answering, to appear)*, (2008).

[16] V. Rieser and O. Lemon, 'Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation (to appear)', in *Proc. of ACL*, (2008).

[17] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, 'Agenda-based user simulation for bootstrapping a POMDP dialogue system', in *Proc. of HLT/NAACL*, (2007).

[18] D. Shapiro and P. Langley, 'Separating skills from preference: Using learning to program by reward', in *Proc. of the 19th ICML*, (2002).

[19] S. Singh, D. Litman, M. Kearns, and M. Walker, 'Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system', *JAIR*, **16**, (2002).

[20] R. Sutton and A. Barto, *Reinforcement Learning*, MIT Press, 1998.

[21] M. Walker, J. Fromer, and S. Narayanan, 'Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email', in *Proceedings of ACL/COLING*, (1998).

[22] M. Walker, C. Kamm, and D. Litman, 'Towards developing general models of usability with PARADISE', *JNLE*, **6**(3), (2000).

[23] J. Williams and S. Young, 'Using Wizard-of-Oz simulations to bootstrap reinforcement-learning-based dialog management systems', in *Proc. of the 4th SIGdial Workshop*, (2004).

[24] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, 2005.

Three Output Planning Strategies for Use in Context-aware Computing Scenarios

Gerrit Kahl[1] and **Rainer Wasinger**[2] and **Tim Schwartz**[1], and **Lübomira Spassova**[1]

Abstract. In everyday life, it is useful for mobile devices like cell phones and PDAs to have an understanding of their user's surrounding context. Presentation output planning is one area where such context can be used to optimally adapt information to a user's current situational context. This paper outlines the architecture of a context-aware output planning module, as well as the design and implementation of three output generation strategies: *user-defined*, *symmetric multimodal*, and *context-based* output planning. These strategies are responsible for selecting the best suited modalities (e.g. speech, gesture, text), for presenting information to a user situated in a public environment such as a shopping mall.
A central point of this paper is the identification of context factors relevant to presentation planning on mobile devices with finite resources to obtain a private and/or public output. We show via a working demonstrator the extent to which such factors can, with readily available technology, be incorporated into a system. The paper also outlines the set of reactions that a system might take when given context information on the user and the environment.

1 Introduction

Let us consider a typical shopping scenario: a customer goes to a supermarket and works through her/his shopping list, step by step. Some products have a range of attributes influencing a purchase, for example price and brand, and the customer may expend a lot of time in searching for just the right product. This process can easily become quite tedious.

One project that deals with shopping assistance is the *Mobile ShopAssist* (MSA) - a PDA based program that acts as a mobile shopping consultant (see [10]). With the MSA software, customers can inform themselves about the properties of different products and compare them with one another. The system provides different input and output modalities that can be used for human-computer interaction, including speech, handwriting and gesture, and customers can, during the interaction process, also reference both virtual data on the PDA's display along with real objects in the surrounding instrumented environment [10].

Customers may also be shopping for different products, ranging from medication (e.g. cold and flu tablets) to electronics (e.g. a digital camera), and the product type may well affect the selection of a modality used in communicating with the customer. For example, when dealing with products of type "medication", the system should avoid revealing any medical conditions the customer might have (*private output*).

The work in this paper can be seen to extend existing work on the Mobile ShopAssist, that focused largely on the recognition and interpretation of multimodal input, as described in section 3. An important aspect that has been extended is that of the generation and presentation of system utterances (described in section 4). In particular, system utterances are dynamically created, based on pre-defined sentence templates, to suit a customer's current context. In addition to this, the presentation of such utterances has been extended such that modalities like speech, gesture, and text, are mapped at runtime to individual semantic elements in the utterance.

2 Related Work

In order to demonstrate how to combine different output modalities, Elting [3] uses a virtual character. He explains that a multimodal presentation should take the current situation and context into account. His virtual character uses graphical and acoustic output modalities, and it is able to adapt to the preferences of the user as well as to the current context.

SmartKom [9, 8] is a system bearing resemblance to this work in that it focuses on context-aware computing and presentation output planning. This system is a multilingual (English and German), multimodal dialog system. Output is done with the help of a virtual character who can use the modalities of speech, graphics, gesture, and facial expressions. The output modalities are synchronized with each other and the system can furthermore choose which situations warrant speech-only output.

Our system is designed for use on mobile PDAs with limited resources. Although SmartKom takes the current context into account for selecting the appropriate output modalities, our system is capable of much finer mappings in which any combination of modalities can be mapped to individual semantic elements in an utterance. Furthermore, this work identifies a wide range of context factors that can have an influence on output generation and also defines the types of system reactions that might be used to adapt a system's output. Some system reactions include, for example, the ability to: modify the format and tempo of the speech output; change the display duration of the text output; and decide whether the output should be presented on- and/or off-device.

[1] DFKI GmbH, email: {Gerrit.Kahl, Luebomira.Spassova, Tim.Schwartz}@dfki.de

[2] Macquarie University, Australia, email: rainer.wasinger@mq.edu.au

3 Mobile ShopAssist Demonstrator

As described in the introduction, this work is based on the Mobile ShopAssist (MSA) demonstrator [10]. In addition to the user's PDA, the MSA uses some output devices situated in the environment to cater for *off-device* communication. Visual output, for example, can be displayed either on a large plasma screen next to the product shelf, or using a steerable projector system (*Fluid Beam* [2, 7]) that allows the creation of projected displays on arbitrary flat surfaces. Acoustic output is performed using a spatial audio system (*SAFIR*) as described in [6], which allows for the creation of virtual sound sources at any location in the environment.

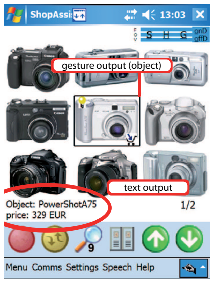
The MSA was built following the concept of symmetric multimodality, which is defined in [8] to mean that "all input modes are also available for output, and vice versa". For the MSA application, the relevant modalities are: speech, handwriting/text and gesture. Gesture can refer to pointing actions on the touch screen of the PDA (see figure 1), and also to the act of taking a physical product out of the shelf. Each product in the shelf is fitted with an RFID tag. An antenna on the back of the shelf detects when a product is taken out of it. With the handwriting mode, the MSA recognizes the user's input by pattern matching it with dynamically loaded finite-state rule grammars.

4 Output Modalities

As user input, the MSA system is able to identify semantic elements like the name of an *object* (e.g. "PowerShot Pro1") and the name of a *feature* (e.g. "mega pixels"). The generated output additionally contains the *value* (e.g. 8) corresponding to the feature of the object. We make a distinction between output on the PDA (*on-device*) and output in the environment (*off-device*).

Speech output is generated in the form of natural language. The sentences are generated using the grammar stored in an XML file and speech is synthesized using ScanSoft RealSpeak Solo[3]. For off-device output the generated sentence is transmitted via wireless LAN to a server, which controls the public speakers, as mentioned in section 3.
As an alternative to the natural language output ("*The PowerShot Pro1 has 8 mega pixels.*"), there is also the possibility to output only the semantically-rich keywords: *"<object>, <feature>, <value>"*, e.g. *"PowerShot Pro1, mega pixels, 8"*. The advantage of this output is that it takes less time. Thus, only the "important" data is presented.

Text output on a PDA is often limited by display size constraints. In the MSA, an efficient text output algorithm was developed to counter this constraint. In our system, text output is shown on the bottom of the display in two rows (see figure 1). It always has the same structure, so that the user can easily find the part of interest. Off-device text output is displayed on a public screen and/or in the space that occurs when an object is taken out of the shelf. The so called *Product Associated Displays* (PADs, see [7]) provide visual feedback to the user in the form of projected images and text.



Object: PowerShotA75
price: 329 EUR

Figure 1. Gesture and text output of the MSA

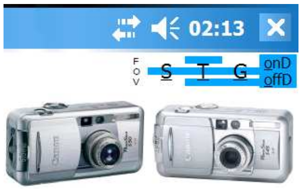


Figure 2. Visualization of the used output modalities

Gesture output for the object consists of the drawing of a border around its image (see figure 1). For the feature, gesture output is achieved by highlighting the corresponding phrase in a scrollbar at the bottom of the screen. In this way, the user gets visual feedback that the system has recognized her/his input. Off-device gesture output is implemented only for *object* (and not for *feature* or *value* attributes). It is presented as a highlighted spot that is displayed on the object in the shelf (see [2]).

5 Output Planning Strategies

Three different output planning strategies are available in the MSA, namely: *user-defined*, *symmetric multimodal*, and *context-based*. The current modality settings are visualized in the right corner of the PDA display. In this way, the user has an overview of the currently used output modalities (see figure 2).

The letters S, T and G stand here for speech, text and gesture; F, O, V, onD and $offD$ stand for feature, object, value, on-device and off-device. When for example speech output for the object is selected, a coloured bar is displayed in the middle of the S-block, i.e. in the same row as the O. In order to make clear which of the three strategies mentioned above is currently being used, there are three different colours for the bars.

5.1 Symmetric Multimodal Output

A central point of this paper is the redefinition of the scope of the well-known term "symmetric multimodality" (see [8]), which in this work refers not just to the ability of using the same modalities for input as for output, but rather to the ability of using individual semantic-element to modality mappings for output as for input. In this way, the user can control which output modalities should be used without explicitly setting them. This means that the user controls the output by applying the corresponding input modality. As there is no input for the value attribute, the output modality for it is set to the input modality of the feature.

5.2 User-defined Output

With the user-defined output strategy, the user can explicitly select the output modalities. Similar to the symmetric output planning strategy described above, output modalities can be flexibly selected for all semantic elements or for any

[3] http://www.scansoft.com/realspeak/mobility/

combination of individual semantic elements to be presented to the user. In this strategy, the used output modalities are independent of the current situation and input modalities. This allows, for example, the user to manually select her/his favoured output modalities. The user can additionally exclude the use of certain modalities by not selecting them. A disadvantage of this strategy is that each modification of the user's preferences requires manual intervention by the user.

5.3 Context-based Output

The context-based output strategy makes use of different *factors* in the user's current situational context to generate an optimal output. These factors lead to different system *reactions* that influence the generation of the output. This context-based planning method represents one of the main contributions of this paper, and is novel with respect to the type and number of context influencing factors, the type and number of resulting system reactions, and the approaches used in determining what reaction to take for a given set of identified context factors, i.e. the output planning (see figure 3).

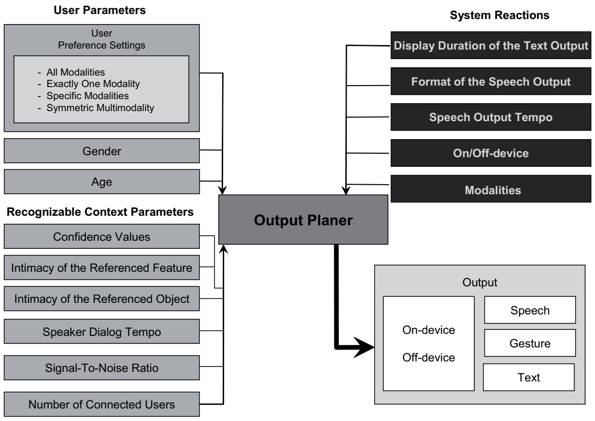


Figure 3. User parameters and system reactions for context-based output generation.

5.3.1 User and Context Parameters

This section describes different factors that can influence the final presentation of the output information. These factors can be classified into two groups: *user parameters*, which are specific to a given user, and *recognizable context parameters*, which are environment-specific.

Age: The age can be a criterion for the retentiveness of a user [5]. Elderly people might need more time for getting the presented information. Hence the text output should be visible for a longer period of time and the speech output tempo should be slower. The MSA system currently obtains the user's age from a user model managing system called UbisWorld [4].

Gender: The gender can for example influence the choice of the voice used for speech output. For instance, the system can choose a female voice for a male user and a male voice for a female user. Similar to the age parameter, gender is obtained through the UbisWorld service.

User Preference Modalities: In the context-based output planning strategy, the modalities selected by the user should be preferred to those not selected. The user can either select exactly one modality, or a combination of output modalities. The exclusive selection of only one modality is considered more important than a combined selection with only one modality. This means that the user prefers only this specific modality, so it should be used over other modalities.

Speaker Dialog Tempo: The speaker dialog tempo is calculated from the time that the user has taken to provide speech input for a pre-determined character length.

Signal-to-Noise Ratio: The signal-to-noise ratio (SNR) describes the signal strength compared to the background noise. The lower the SNR the more noise was recognized. In a noisy environment, the speech output might be hard to understand. To compensate for this, the system can raise the volume of the speech output or use an additional modality, e.g. graphical output. The SNR is detected by the speech recognizer.

Confidence Values: Confidence scoring refers to the process of attaching likelihood values to recognition results in an attempt to measure the certainty of finding a correct match to a user's input. For each of the modalities, a confidence value (Cf) is generated each time a user interacts with the system [12, 10].

Intimacy of the Object: Different objects have different intimacy levels. The intimacy level of the object is often highly user-dependent. Examples of objects with high intimacy levels might be medications, cosmetics, or contraceptives. For such products, unobtrusive output modalities (e.g. graphical output on-device) should be used.

Intimacy of the Feature: Similar to object intimacy, there are also different intimacy levels for features. An example of a feature with a high intimacy value might be the size of a particular item of clothing.

Number of Nearby People: If many people are nearby to the user, it might be undesirable to use off-device output: On the one hand, speech output of different users could overlap, and on the other hand, the user could feel uncomfortable by the speech output. The MSA system estimates the number of nearby people based on the number of people currently localized in the vicinity (see [1]).

5.3.2 Heuristically-derived System Reactions

After detecting the possible factors that might influence the system's reactions, the system uses these factors to generate an appropriate output.

Display Duration of the Text Output: As a baseline for the duration of the text output, the user can determine a preferred value. However, the ultimate display duration can be increased by the system according to the user's age and the SNR value. On the other hand, it can also be decreased depending on the level of intimacy of the object and/or the feature.

Format of the Speech Output: The decision to choose either natural language or short output depends on the one hand on the user's parameters, like preferences, age, and speech input tempo and on the other hand on environmental parameters. like SNR value and the number of nearby people. A higher age and a higher SNR value for example favour the natural language output, whereas a large number of nearby people and a high speech input tempo would rather lead to the generation of short output.

Speech Output Tempo: Similar to the choice of the speech output format, the speed of the speech output depends on the user's age and speech input tempo, and the number of nearby people.

On-/Off-device: Whether the output is presented on- and/or off-device depends on the user settings, the intimacy of the object and feature attributes, the number of nearby people, and the confidence values. The more people that are in the vicinity of the user, the more off-device output should be avoided. Information about objects or features with a high intimacy level should not be presented in the environment. Moreover, unconfident system responses should not be presented off-device. The on-device modality is selected, if it is explicitly preferred in the user settings or if no off-device output is allowed.

Modality Selection: The gesture output modality is selected if it is either explicitly preferred by the user or if gesture input for the object was used. It is difficult to foresee any reason for not selecting gesture output when the user explicitly prefers it, and hence this preference is never ignored. When the user applies gesture for input, the system responds with gesture for output.

In the case of speech output, the environmental context plays an important role. In certain situations, it seems sensible not to use acoustic output, even if it is selected as preferred in the user settings. First of all, we consider the number of modalities selected by the user: if more modalities are selected, the speech modality can be resigned more easily. If *symmetric multimodal* is selected, the used input modality is also taken into account. As we found in an empirical study, speech output is more preferred by female users. Probably the most important factors for the choice of the speech output modality are the signal-to-noise ratio, the number of nearby people, and the intimacy of the currently selected object/feature. Additionally, the certainty with which the object input was recognized also plays a role, such that a more unobtrusive modality should be used if the object might not have been recognized correctly.

Similar to with gesture output, there is no reason to deselect the text output modality if it is explicitly selected by the user. As stated in [5], elderly people can recognize graphical output better than acoustic output. Therefore, for elderly users, text output is always displayed in our system.

6 Conclusions

The Mobile ShopAssist has undergone a number of usability studies in the past, primarily concerned with user preference for modality combinations (see [13] and [11]). Current work is now focused on an additional field study aimed at determining the accuracy and suitability of the presented output strategies for mobile users in a shopping domain. From a pilot study that has recently been conducted on the relevance of individual context factors, it has already been found, for example, that the intimacy of an object's features is considered less important than the intimacy of the object itself. Most of the interviewed participants in this study also highly rated the importance that the number of nearby people would have in a system selecting the current optimal set of output modalities.

In this paper, we have presented a context-aware output planning module and three accompanying strategies used for output generation in a shopping domain, namely: *user-defined*, *symmetric multimodal*, and *context-based*. In support of these strategies a range of context parameters relating to the user and the environment were identified (e.g. the user's age and gender; signal-to-noise ratio; the number of nearby users). Additionally, a range of possible system parameters used in determining appropriate reactions to take when presenting semantic information over a given set of modalities was identified (e.g. duration in which text is displayed; the format and speed of speech output). The outlined context parameters and system reactions can be seen to provide vital insight for all systems with a research focus on context-aware computing. Future work will now entail testing the degree of suitability of the proposed output planning strategies.

REFERENCES

[1] B. Brandherm and T. Schwartz, 'Geo referenced dynamic Bayesian networks for user positioning on mobile systems', in *Proceedings of the International Workshop on Location- and Context-Awareness (LoCA)*, pp. 223–234, (2005).

[2] A. Butz, M. Schneider, and M. Spassova, 'SearchLight: A Lightweight Search Function for Pervasive Environments', in *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive)*, pp. 351–356, (2004).

[3] C. Elting, 'What are multimodalities made of? - modeling output in a multimodal dialogue system', in *Workshop on Intelligent Situation-Aware Media and Presentations ISAMP 2002*, Edmonton, Alberta, Canada, (2002).

[4] D. Heckmann, *Ubiquitous User Modeling*, Ph.D. dissertation, Department of Computer Science, Saarland University, 2005.

[5] J. Jorge. Adaptive tools for the elderly: new devices to cope with age-induced cognitive disabilities, 2001.

[6] M. Schmitz and A. Butz, 'Safir: Low-cost spatial audio for instrumented environments', in *Proceedings of the 2nd International Conference on Intelligent Environments*, pp. 427–430, (2006).

[7] L. Spassova, R. Wasinger, J. Baus, and A. Krüger, 'Product Associated Displays in a Shopping Scenario', in *Proceedings of the 4th IEEE / ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 210–211, (2005).

[8] W. Wahlster, 'Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell', in *Proceedings of the Human Computer Interaction Status Conference*, pp. 47–62, (2003).

[9] W. Wahlster, N. Reithinger, and A. Blocher, 'Smartkom: Multimodal communication with a life-like character', in *Proceedings of Eurospeech*, pp. 1547–1550, (2001).

[10] R. Wasinger, *Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations*, Ph.D. dissertation, Saarland University, Department of Computer Science, 2006.

[11] R. Wasinger, A. Krüger, and O. Jacobs, 'Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant', in *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive)*, pp. 297–314, (2005).

[12] R. Wasinger, C. Stahl, and A. Krüger, 'Robust Speech Interaction in a Mobile Environment through the use of Multiple and Different Media Input Types', in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1049–1052, (2003).

[13] R. Wasinger and W. Wahlster, 'Multi-modal Human-Environment Interaction', in *True Visions: The Emergence of Ambient Intelligence*, 291–306, (2006).

Pragmatics and Human Factors for Intelligent Multimedia Presentation: A Synthesis and a Set of Principles

Frédéric Landragin[1]

Abstract. Intelligent multimedia presentation systems (IMMPS) have to take into account pragmatics and human factors such as the specificities of human perception, attention, memory, conceptualization, and language. Using a conversational animated agent or not, some principles can be followed to increase the communicative abilities of interactive systems. In this paper we propose a set of such principles. We exploit our background in natural language processing and computational pragmatics to provide specifications for multimodal systems. The classifications and principles and architectural concerns we present are based on some experimental observations (that are not described here) and constitute a kind of white paper for future implementations.

1 INTRODUCTION

An intelligent multimedia presentation system (IMMPS, see Bernsen [4], Bordegoni et al. [5], Karagiannidis et al. [12] and others) has to translate display requests from a dialogue manager into output messages, and therefore has to take into account the particular characteristics of the information, the terminal, the physical environment, and the addressee (or user). When information has to be spread over several communication channels that correspond to different communicative modalities, we talk about multimodal fission. The term 'information' groups natural language and multimodal utterances from the user and the system as well as the associated application data.

Following this definition that characterizes our approach, we can distinguish two parts for the presentation process in a dialogue system. First, the dialogue manager takes the **decisions** on the following aspects ('WH- part' of the process):

- '**Who**' = to whom the information has to be presented,
- '**What**' = what is the information to present,
- '**Which**' = which part of the information has to be emphasized,
- '**Where**' = where can the information be displayed, i.e., on which devices,
- '**When**' = when and for how long must the information be presented.

Second, the IMMPS **realizes** these decisions ('HOW part' of the process) by: choosing the method to valorize the related piece of information (cf. 'which' in the previous list), choosing the modality or modalities and the device or devices to exploit (cf. 'where'), dividing the information to determine the related pieces of information for each modality (cf. 'where'), dividing the information to spread its presentation over time (cf. 'when'), and, possibly but not necessarily, managing a human-machine interface (HMI), for instance a graphical user interface (GUI), that is specific to the presentation, e.g., navigation buttons when information has to be split for several display steps.

With these simple items, we want to make precise the roles of intelligent multimedia presentation. Our proposal is not that different from existing ones like [19] or others, but it includes as many aspects as possible, in particular a clear separation between the dialogue concerns and the presentation concerns. These items are valid whatever the form of the IMMPS (avatar or not), whatever its communicative status, from a fully recognized interlocutor to a simple intermediary with the application. More precisely, the IMMPS can have the status of an interlocutor, i.e., can stand as a 'majordomo'. The user can interact with it, the details of the exchanged information having no interest for the application or for the dialogue manager. The advantage is that the user's actions that concern only the HMI or GUI are treated very quickly. To the contrary, the IMMPS can have no materiality for the user, who believes he/she is communicating directly with the application. The advantage here is the simplicity and transparency for the user.

After a section presenting some first principles for taking into account pragmatics and human factors in multimedia presentation, we will focus on the determination of all input parameters that a presentation system may take into account. These parameters are presented in a set of classifications. A general architecture illustrates the processes that exploit them. These processes are grouped into two main steps that are then described in details, with examples of rules and strategies for multimedia presentation.

2 FIRST PRINCIPLES

2.1 Nine principles for IMMPS

Our approach and preoccupations can also be summarized into a set of principles, which can be compared to the Grice's maxims dealing with more general conversational principles [10]. To us, designing more natural and adaptive IMMPS requires that the characteristics of the information (or message) in its context, in particular the linguistic context or dialogue history, are taken into account in a better way. This first point leads us to propose four principles:

1. "More natural IMMPS with a better repartition of information over the communication channels",
2. "More natural IMMPS with a natural rendering and valorization of the information on a communication channel",

[1] CNRS, LaTTICe Laboratory (UMR 8094), Montrouge and Paris, France. Email: `frederic.landragin@linguist.jussieu.fr`.

3. “More natural IMMPS with a better exploitation of the semantic content of the message”,
4. “More natural IMMPS by maintaining better cohesion and coherence with previous messages”.

Second, designing more natural IMMPS (more natural in the sense of more used-centric, with more naturalness and adaptative abilities) requires taking into account the characteristics of the terminal (presentation means), and the physical and situational environment (presentation conditions). Hence, we propose the two following principles:

5. “More natural IMMPS with a more refined exploitation of presentation means”,
6. “More natural IMMPS with a more refined exploitation of presentation conditions”.

Third, to provide more user-oriented IMMPS, i.e., presentation systems that are more sensitive to human abilities and behaviors, there is a particular need to take into account the addressee’s physical and cognitive abilities, as well as his role(s) in the application domain and preferences for information presentation. Three additional principles can then be expressed:

7. “More natural IMMPS with a better exploitation of the addressee’s expectations”,
8. “More natural IMMPS for a better perception of the message by the addressee”,
9. “More natural IMMPS for more relevant future reactions from the addressee”.

2.2 Multimedia information

Multimedia information can have many forms and contents. Some characteristics are essential when presenting. Among them, we can cite following Arens and Hovy [2] the urgency (‘urgent’ or ‘routine’, for instance), the transience (‘live’ or ‘dead’), the critical importance or criticality (‘nominal’, ‘critical’, ‘fatal’), the density or structure (‘continuous’, ‘discrete’), the coverage or number of simple items that are grouped into one complex structure (‘singular’, ‘low’, ‘high’, ‘total’), the volume (‘much’, ‘little’, ‘single’) and so on. Moreover, the application often consists of managing complex information such as cartography or video. A lot of work also deals with the best ways to represent such information with a particular concern on adaptation to the context. Still, choosing the relevant characteristics given a particular application remains a complex problem. We will try to extract the characteristics that are essential to our proposal from all the ones mentioned.

Another aspect of research work dealing with multimedia information is the representation of such information for communicative systems. A lot of standards or standard proposals have been designed: EMMA from the W3C [7, 25], SMIL that focuses on the synchronization problems [3, 26], MPML [17] and others. Even if studying such initiatives can provide ideas on how to represent multimodal information, our approach is at too early a phase to exploit them. Semantics, pragmatics and user’s abilities are not the main preoccupations of these initiatives, but they are ours.

2.3 Human factors for IMMPS

Whereas the term ‘adaptability’ is used for the adaptation of the interface by the user and is studied at design time, ‘adaptivity’ is used for the adaptation of the interface by the system itself, at run-time. Then, adaptivity groups all dynamic aspects of adaptation and is very close to our concerns about IMMPS and human factors. More precisely, following work such as [8], we can state that:

- adaptation to the **terminal** intervenes during the presentation because the presentation method depends on the terminal characteristics,
- adaptation to the **physical environment** intervenes during the presentation because criteria such as background noise level consist of parameters for IMMPS,
- adaptation to the **user’s preferences** intervenes during the presentation (it is of course in the interest of IMMPS to follow the display preferences),
- adaptation to the **user’s roles** (or user task) intervenes during the presentation because IMMPS can exploit its knowledge of the user’s roles for emphasizing a piece of information,
- adaptation to the **user’s access rights** (or user’s prerogatives or profile) intervenes before the presentation because the dialogue manager has to filter the information that the user must not know.

Information presentation and information adaptation are thus two similar problems. The important point is how information is really adapted to the user’s preferences and abilities. In particular, cognitive abilities such as attention, perception, immediate memory, mental representation, conceptualization, judgment, decision, and so on, can have an influence on how information could be best represented. A lot of work has been done on communicative agents and avatars, for instance the current research on emotion rendering, but there is a lack of implementation of theories dealing for instance with the Gestalt Theory [13] or salience, for instance. What we want to address here is the integration of such factors into IMMPS in order to have a certain control on the user’s behavior. For that, we will follow work such as [23] and we will extend our approach for natural language and multimodality understanding.

3 IMMPS AND HUMAN FACTORS

3.1 General presentation

Figure 1 presents the global architecture that underlies our approach. The core is the multimodality manager that treats input and output multimodality, and manages the multimodal context, i.e., the history of multimodal utterances and actions from the user and the system. The components of the output manager and the parameters they exploit are described in the following subsections.

3.2 Input parameters for IMMPS

Input parameters can be separated into three categories. The first relates to the information to be presented, and includes the structure and content of the information as well as the pragmatic force it is associated with. The second relates to the presentation means and groups the terminal and the physical environment. The third relates to the presentation addressee (the user) and groups the preferences, abilities, roles, and all human factors, with a sub-distinction between physiological factors, linguistic factors, and cognitive factors.

3.2.1 Information-related parameters

The criteria used for the repartition and valorization of the information and related to the information itself can be classified into three

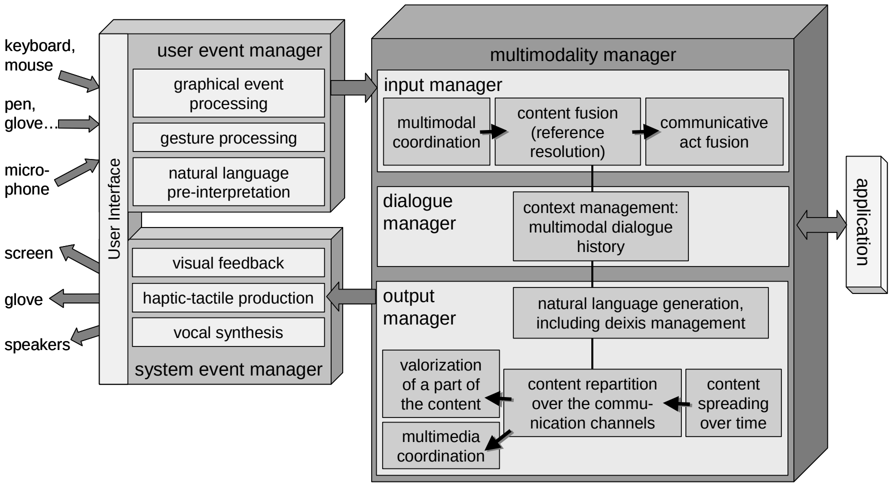

keyboard, mouse
pen, glove…
micro-phone
screen
glove
speakers
User Interface
user event manager
graphical event processing
gesture processing
natural language pre-interpretation
visual feedback
haptic-tactile production
vocal synthesis
system event manager
multimodality manager
input manager
multimodal coordination
content fusion (reference resolution)
communicative act fusion
dialogue manager
context management: multimodal dialogue history
output manager
natural language generation, including deixis management
valorization of a part of the content
content repartition over the communication channels
content spreading over time
multimedia coordination
application

Figure 1. Architecture for input and output multimodality management

categories. The first deals with the message content and includes: (a) the level of criticality, (b) the level of urgency, very important because IMMPS must be able to stop any process when an urgent information has to be displayed, (c) the information complexity, i.e., some precisions about the data structuring and the size and numbers of items, (d) the information constitution, i.e., some precisions about its density (discrete or continuous, list or table of items, timetable, etc.), (e) the information scope, for instance the fact that the information has two poles of interest, firstly the whole information and secondly a zoom in on one particular element, and (f) the presentation constraints that are inherent to the information: visual constraint for a cartography, no constraint for a linguistic message or for data that can either be displayed or verbalized. Note, the distinction between the level of criticality and the level of urgency is important because it has an influence on the IMMPS behavior. Critical information should be presented using a particular rendering to make it obvious to the user, whereas an urgency should stop all the current processes so that the user is face to face with it and only it. There is concerning the information scope an important aspect linked to the users' perceptive abilities. In the case of a huge table of numeric values as the information to present, two complementary strategies can be imagined. The aim can first be to present the information in its entirety, so that the user can apprehend its scope in one glance, even if no value can be read because of the very low font size that is required. A second aim can be to present the content of one particular cell to the user, and then to exploit a kind of magnifying glass. The method that groups both aims is sometimes called 'keeping the context'. With the information scope and the privileged aim as parameters, IMMPS should be able to choose between one strategy, the other, or both.

The second category of parameters groups pragmatic aspects with illocutionary and perlocutionary forces of the message: (a) the communicative act(s) that is/are determined by the dialogue manager, and (b) the expected reaction from the user: feedback or not, immediate action or not. Illocutionary and perlocutionary forces [21] will be described in detail in section 3.3.

The third category relates to the interaction history: (a) the history of the display actions, in order to allow the mention of a previously executed action, and (b) the stack of the displayed data, in order to allow the mention of previously displayed data.

3.2.2 Presentation means-related parameters

The characteristics of the terminal constitute a first set of parameters related to the presentation means: (a) terminal availability, (b) dimension constraints such as screen size, (c) constraints on the processing delays, and (d) constraints and preferences on output modalities.

A second set of parameters consists of the parameters that are related to the presentation environment. For these, we propose to exploit and adapt the information presentation to the three functions of gesture that were identified by Cadoz [6] for the gesture as an input in HMI: (a) **epistemic** constraints, that are linked to the 'learning from the environment' function, typically picking up and taking into account the ambient noise and the ambient luminosity, (b) **ergotic** constraints, that are linked to the 'transforming, changing the state of the environment' function, typically thresholds for ambient noise and luminosity, that must not be overstepped in order to not disturb the environment, and (c) **semiotic** constraints, that are related to the 'communicating meaningful information toward the environment' function, typically the quantity and quality of speech delivery, e.g., too loud or too fast considering the environment.

3.2.3 User-related parameters

Four categories can be distinguished here. First, the parameters that deal with the user's physical abilities: (a) constraints on the ways of working with communication channels, for instance due to a handicap, and (b) constraints and preferences on the exploitation levels of the communication channels, e.g., when the visual channel is already monopolized by another part of the ongoing task. Here, the auditory channel has a particular role, because it is sometimes the only possible modality to convey a message (the user may use his hands for the ongoing task and can only use speech to express something else). Second, the parameters related to the user's roles: (a) constraints on the access rights and bans that come from the 'user profile' (a user-related resource that is managed by the dialogue manager), and (b) constraints and preferences that come from the ongoing 'user task' (another resource managed by the system). Third, we can group all other individual preferences, particularly: (a) the preferences for linguistic terms and presentation metaphors, which were previously expressed by the user, and (b) the preferences on the dialogue management, which are detected and exploited dynamically by the dialogue manager, e.g.: the user prefers short answers to long ones; the user always prefers to conclude a sub-dialogue before going back to the main dialogue. In the last category we can group all other human factors that correspond to universal preferences, i.e., preferences that apply to everybody due to (a) human physiology, (b) linguistic abilities, and (c) cognitive abilities.

Physiologic preferences are first linked to the modality. Within the sound modality, two statements are of importance in particular for beep or horn messages (also called earcons). First, the stronger the sound is, the more powerful it is (but the more stressful it is). Second, high pitch is more strident than low pitch. Similar statements can be taken into account for visual modality. Following color theories [11], red is perceived much quicker than blue or yellow, and therefore is more often exploited for visual alerts. Blue can be perceived much easier in dark environments than in luminous environments. The center of the visual field that corresponds to the fovea is a privileged place. The notion of 'good form' from the Gestalt Theory [13], for instance a perfect and simple circle, is a privileged form. Moreover, salience and pregnance can be relevantly exploited whatever the modality. A salient element, i.e., an element that can be distinguished by singular properties (e.g., the only red element), is more easily perceived. A pregnant element, i.e., an element that has been the object of previous repetitions so that it impregnates the user's memory, is also more easily perceived. IMMPS should exploit such criteria to optimize its presentations considering the particularities of human perception.

Concerning linguistic particularities, simple statements can also be done at the different linguistic levels. At lexical and syntactic levels, IMMPS may keep the terms and syntactic constructions from the user, and may, in a general manner, use simple words and constructions. At semantic and pragmatic levels, the Grice's maxims [10] may be exploited when determining the message to generate. The risks of ambiguities could be minimized, for instance by avoiding anaphora when several potential antecedents are possible. With the same purpose, indirect and composite speech acts should be avoided. At a stylistic level, the informational or communicative structure [15] should be exploited in order to put one particular message element forward. This 'putting into salience' or 'saliencing' process is done by choosing the relevant grammatical function, thematic role, theme, focus, etc. Coherence (generating a message with a logical link with previous ones) and cohesion (generating a message whose form is in direct continuity with the form of previous messages) should be exploited.

Concerning cognitive preferences, the particularities of lower cognitive processes (perception, attention, memory) and of upper cognitive processes (mental representation, judgment, decision) should be clarified and taken into account. Then, IMMPS should be aware of the size of short-term memory (from 5 to 7 independent items, see [16]), of selective and persistent attentions, etc. More precisely, a message can have the purpose of capturing selective attention (e.g., alerts) or to request an important amount of persistent attention for a thorough treatment (e.g., presentation of an important information). IMMPS must give no opportunity for selective attention to be diverted in various directions, and should provide time to the user's persistent attention. Moreover, each message leads to a representation process whose complexity depends on the complexity of the information in its canonical form. So IMMPS should stay inside reasonable limits. Some pieces of information require a judgment. So IMMPS should not multiply such pieces of information in the same presentation act. Because of their visual characteristics, some pieces of information have an influence on the actions that can be done on them. IMMPS should manage such affordances [9] in a relevant way. In a general manner, it can be very efficient to exploit all that has already worked well. For instance, if the system noticed that a particular visual form has a positive and efficient influence on the user, it may decide to use it again in similar situations.

3.2.4 Statement on inputs and outputs

The constraints and principles we have described can be summarized in the following process:

- From the applicative domain, the user task and user profile: (a) levels of criticality and urgency, (b) self-descriptive information (organized and quantified information), and (c) presentation constraints and preferences that are specific to the task or task type;
- Computed by the dialogue manager: (a) pragmatic forces and other labels on the message, for instance an emotion to render, (b) coherence and cohesion indications, (c) linguistic valorizations, and (d) constraints and preferences on linguistic terms and dialogue management;
- Determined by IMMPS on the basis of the constraints from the previous items: (a) information ordering (e.g., depending only on urgency levels), (b) method to dissociate an information into several presentation phases, (c) method to dissociate an information over the communication channels, (d) for each piece of information, level of valorization (e.g., depending only on criticality), (e) method to valorize a piece of information, and (f) method to exploit the preferences, in particular when they contradict each other.

3.3 Pragmatics for IMMPS

Following our approach, human-machine dialogue systems should be able to communicate with their users in a spontaneous and natural way, by exploiting the main human communicative means that are language and gesture. Thus, information presentation must be linked to natural language generation. Among natural language aspects, we want to emphasize the pragmatic aspects, and, in particular, the pragmatic forces (locutionary, illocutionary and perlocutionary forces, see [20] and [21]) that are conveyed together with a message. Since multimodality includes natural language, pragmatic forces will apply on each multimodal message and multimedia presentation act. In this subsection we show how illocutionary and perlocutionary forces can be handled by IMMPS.

3.3.1 Illocutionary force

When interpreting as well as generating, the message content is associated with an illocutionary force that expresses the act that is realized by the enunciation, and that depends on an underlying intention. Following Relevance Theory [22], 'saying that', 'telling to', and 'asking' are the main illocutionary forces. By 'saying that', the speaker expresses an **assertion** in order to make the addressee know something. By 'telling to', the speaker expresses a **demand** in order to make the addressee do something. By 'asking', the speaker expresses a **question** in order to know something from the addressee, with two cases: the close question or 'asking if' whose answer is yes or no, and the open question or 'asking WH-' whose answer is an information.

Concerning multimedia presentation, the way of presenting, for instance an alert, depends on the illocutionary force. If the system just wants to **inform** (‘saying that’), it can use a certain method of presentation that totally differs from the method used to **encourage to act** (‘telling to’). Moreover, the dialogue system may need a confirmation of the message reception. Then we can distinguish a ‘saying that without feedback’ from a ‘saying that with a mandatory feedback’. The second corresponds to the use of the classical ‘OK’ or ‘OK/cancel’ dialogue boxes. As one of the contributions of our approach, we propose to model the two previous points with composite speech acts. That is not very far from the concepts of active presentation and passive presentation from [1], but here we emphasize the pragmatics of communication as it is modeled elsewhere in linguistics and computational linguistics work. One important point concerning the request for a feedback from the user is that such a behavior may be decided by the IMMPS with the help of the task manager. In fact, for some particular actions, a feedback request can be included in the task model. In such a case, IMMPS must not add an additional feedback request.

A distinction can be made between an explicit order and an implicit order. The ‘OK’ dialogue box constitutes an explicit order because it materializes the need of the system to get a confirmation of the reception of its message. On the other side, a message, for instance an alert, which aims at strongly encouraging action with no materialization of this goal constitutes an implicit order:

- Alert = “**saying that** problem” + “**telling to** react to it” (implicit order),
- Saying that with ‘OK’ feedback = “**saying that** information” + “**telling to** feedback” (‘OK’ explicit order),
- Saying that with ‘OK’ or ‘cancel’ feedback= “**saying that** information” + “**asking if** agree”.

3.3.2 Perlocutionary force

Each message also has the aim of producing an effect on its addressee, whatever this effect is (just taking the message content into account, or realizing something precise). We claim that it is the dialogue manager that must manage the perlocutionary force, in particular the expected reaction following a demand from itself (next state in the user task model). It is the IMMPS that must correctly convey the perlocutionary aim, for instance by making a waiting attitude from itself obvious. As an example, an alarm that follows the detection of an inconsistency in the application database can have two aims: informing the user, i.e., something like “be careful, there is an inconsistency”, or encouraging the user to give an information he is susceptible to know but has not yet passed on. In this case, a solution consists of opening a text box window, as an equivalent of an ‘asking WH-’ speech act. If the interface integrates an animated conversational agent, another solution consists of displaying an attitude that clearly conveys an expectation of the user’s behavior.

In the case of a GUI, the perlocutionary force is linked to the graphical metaphors. In fact, the choice of the elements of the GUI has an influence on the user’s future actions. As very simple examples, we are used to pushing buttons, and we try to write text inside each element that looks like a text box. In particular, when a table is displayed, we try to modify the content of the cells. In a general manner, we know that each displayed element has a function, and if we do not know that function we try to identify it. Consequently, the IMMPS must know the functions of all the GUI elements that it may have to present. Moreover, it must take these functions into account during the various phases of the presentation. For each GUI element, it must be aware of the input interaction possibilities. Then, it must inform the input events manager, and indeed the fusion module. To continue with the previous example, a table of numeric values can be presented using several methods depending on whether the values can be modified or not. First, the cells can be presented with a particular color or rendering, for instance with a grey tint if they are not modifiable. Second, each cell can be accompanied with a text box that makes the possibility of modification obvious.

Two additional aspects can be discussed on how the perlocutionary forces can be materialized considering the particularities of vocal interaction and natural language, for instance with mentional expressions. Considering the difficulties of speech recognition, several recognition grammars can be specified depending on the type of expected input utterances right after a multimedia presentation. Typically, a very general grammar, which by consequence is not very precise, is used when the system has to detect the theme in order to launch the related application. On the other hand, a command grammar is used when the user has the dialogue initiative. Another grammar that is specific to numbers, dates, and numeric values, is also used when the user must answer a question from the system that deals with such data. Consequently, IMMPS must take into account the type of vocal feedback that is susceptible to follow a presentation act, and must inform the recognition module. For instance, the command grammar may be activated right after an inform-like presentation, and a specific grammar after a question-like presentation.

Concerning the particularities of natural language, we claim that the method to present multimedia information has an influence on the user’s future linguistic choices. In particular, displaying pieces of information that follow an obvious order (arrival order or visual organization order) will favor **mentional expressions** such as “the first”, “the second”, “the next one”, or “the last one”. Likewise, displaying pieces of information that are obviously dissociated will favor **quantified expressions** such as “each”, “all the”. The consequences for the understanding of a linguistic message that follows a multimedia presentation are multiple. IMMPS must be aware that it may have to make obvious an ordering that was not expressed by the dialogue manager, for instance the default occidental vision order, from left to right and from up to down. IMMPS must also be aware of the way pieces of information are stuck together or not. Moreover, in such cases, IMMPS could inform not only the recognition module but also the module dedicated to natural language understanding.

3.3.3 Statement on communicative acts for IMMPS

With the four intentions that are (a) inform without feedback, (b) inform with mandatory feedback, (c) encourage to react, and (d) question, we propose the corresponding presentation acts classification:

- ‘Inform’: equivalent to the ‘saying that’ speech act, with no feedback required,
- ‘Feedback inform’: equivalent to the ‘saying that’ + ‘telling to’/‘asking if’ composite speech act,
- ‘Demand’: equivalent to the ‘telling to’ speech act,
- ‘Question’: equivalent to the corresponding speech act, with the distinction between open question and closed question.

As an example of simple processing rules for IMMPS, the presentation act could depend firstly on the level of criticality: ‘feedback inform’ for a high level and ‘inform’ for a lower level. The act could also depend on the level of urgency: ‘demand’ for a high level and ‘inform’ for a lower level. This shows how pragmatics is related to information nature and useful to multimedia information presentation.

4 A TWO-STEPS PROCESS FOR IMMPS

4.1 First step: information repartition over the communication channels

The generation of output multimodal messages as well as the presentation of multimedia information is a process that aims at repartitioning the content of the message or information over the communication channels, and at valorizing all partial contents within each communication channel. The main strategy for information repartition consists of taking into account a set of constraints and a set of preferences. Some additional strategies can then be imagined to optimize the generation of the message and favor its perception by the user.

4.1.1 Repartition by exploiting constraints

Constraints must of course be taken into account at the earliest phase of the presentation process. Some constraints are inherent to the information. As a typical example, the presentation of a map must be done graphically and not vocally (with the exception of the description of an itinerary over the phone, but in such a case the information is not really a map but an annotated description of a particular aspect of a map). Some other constraints are linked to the terminal, for instance when the terminal cannot produce vocal messages. There are also constraints that are associated with the presentation environment. In fact, environmental conditions can impose or forbid some of the communication channels. For instance, a strong ambient noise will forbid the vocal modality. Finally other constraints are linked to the user's abilities and roles. Concerning abilities, this is the case for some kinds of handicap. Concerning roles, it depends on the application and usual user profiles.

4.1.2 Repartition by exploiting preferences

Once the constraints have been taken into account, a lot of choices are possible and the IMMPS has to compute the various aspects of the interaction to make the most relevant choice. Among these aspects are: (a) the message content, with the exploitation of the communication channel that best fits the information constitution, and the preferential exploitation of several channels when the information is very complex, (b) the communicative act, with the preference of one single channel for a simple act and two channels for a composite act, (c) the interaction history, with a preference for the exploitation of the channel that has already been successfully exploited, (d) the user's preferences, which of course should be satisfied (the user can for instance prefer auditory feedbacks to visual ones), and (e) human factors. As an example of this last important aspect, displaying large information should be spread over time, in particular if reading it requires an important amount of persistent attention.

To compute the constraints and preferences and to choose the best information repartition paradigm, a lot of rule systems have been defined in the literature. Our purpose here is not to define another rule system, but to provide pragmatic and human factors related recommendations to design more natural IMMPS. These recommendations will have to be completed considering the design context. In particular, application specificities such as the data manipulated and the objectives of the user, will have to be carefully studied before defining the rule system. The rules will also be defined with preliminary consultations of ergonomics experts. As a basic example dealing with the characteristics of the message, here is a set of rules that can be considered as a basis:

- If the urgency of the message is fatal, then use both visual and auditory communication channels;
- If the urgency of the message is critical, then use the auditory communication channel as a priority;
- If the urgency of the message is nominal, then use either visual and/or auditory communication channel;
- If the information level of criticality is fatal, then use the auditory communication channel;
- Etc.

Another example of a rule system deals with the user's activity during the interaction:

- If the user is distant from the different interaction devices, then use both visual and auditory communication channels;
- If the user is close to the different interaction devices, then use either visual and/or auditory communication channels;
- If the user is attentive to the ongoing interaction, then use either visual and/or auditory communication channels;
- If the user is absent-minded, then do not use an auditory communication channel.

4.1.3 Making the link between distributed information

When a part of the information is displayed and another part is uttered, the user may not be able to make the link between the two parts and consider the information as a whole. But such a link is important because the two parts of the message must not be considered as two distinct messages, i.e., messages that can be treated separately. In their rule system, [24] emphasize this point with a rule for checking the presence of a semantic link between the visual part and the vocal part of the message. To the contrary of a VU meter or a video where the temporal synchronization is sufficient for the user to put together the sound track and the visual track, the association of a vocal message to a visual feedback might be seen as two distinct messages instead of one. Then, in this case there is the need for the system to provide some indications to the user so that he puts together the visual feedback (e.g., emphasizing a particular visual object) to the related part of the vocal message (e.g., a referring expression such as "this object").

A way to do this is to indicate with a modality that another part of the message is conveyed using another modality. An additional visual feedback can thus make the presence of the vocal message obvious. Using natural language, vocal messages such as "on the currently displayed map, you can see..." or "flight 102 is the one that flashes" include a reference to the visual modality. In fact, such a reference is a kind of 'deixis' and has been well studied in linguistics and computational linguistics works [14]. To us, an IMMPS has to handle, with particular care, deictic cases, in order to well manage the interactions between natural language and visual perception. As an example, the generation of "flight 102 is the one that flashes" can be seen as the following suite of processes:

1. The dialogue manager produces a presentation request using a logical form that corresponds to something like "make-obvious-to-the-user (flight-102)";
2. The IMMPS chooses both a visual and vocal realization with the generation of a deixis, so that the user brings the two realizations together;
3. The IMMPS asks the natural language generation module to materialize the inter-modal deixis, i.e., the IMMPS indicates the nature of the display;

4. The natural language generation module produces the expression “the one that flashes”;
5. The IMMPS produces “Flight 102 is the one that flashes” and activates the visual flashing rendering.

Since the nature of the visual feedback is clearly explicit, this way of proceeding corresponds to an explicit inter-modal deixis, as opposed to the following implicit inter-modal deixis (using the same example but with another choice from the natural language generation module):

4. The generation module produces the expression “here is”;
5. The IMMPS produces “Here is the flight 102” and activates the visual flashing rendering.

Thus, deixis management is one important aspect of multimedia information presentation, and is integrated in our architecture of Figure 1.

4.1.4 Reinforcing the message by exploiting redundancy

When the level of urgency is high and in other cases of presentation, there is the need to reinforce the message in order to increase the probability of it being well perceived and assimilated by the user. This can be done by duplicating the information over two or all the communication channels, i.e., exploiting redundancy.

In fact, redundancy is the classical way to emphasize information when generating in a multimedia context. We want to soften this method, with the following arguments. First, there are of course a lot of arguments for the exploitation of redundancy: (a) if a communication channel does not work well, the other one makes up for it, (b) the more information is emitted, the more chances the addressee has to receive it, (c) the more information is presented again, the more chances the addressee has to become imbued with it. Second, these arguments can be opposed to human factors preoccupations that work against redundancy: (a) too many messages do not encourage the addressee to maintain his persistent attention, (b) too many messages increase the processing time and therefore the expected reaction delay. As an illustration, remember the famous example of an air crash due to a bad interpretation of redundant information: “– Why didn’t you answer the control tower who indicated to you that your landing gear was not out? – Because I had a klaxon that was sounding in my ears! – That’s incredible! That signal precisely indicated to you that your landing gear was not out!” As a statement, we propose the following basic but important rules to handle redundancy:

- Exploit redundancy only if the addressee should be able to make the link between the various emissions of the same information, i.e., if he can notice that it is redundancy;
- Do not exploit redundancy in the same communication channel (e.g., sound and voice like in the air crash example);
- When the message is so urgent or important that it cannot be ignored, be careful that redundancy does not introduce any perturbation.

4.1.5 Information repartition and multimodal fission

It is now well stated in the literature that managing input multimodality corresponds to the ‘multimodal fusion’ problem, and that managing output multimodality corresponds to the ‘multimodal fission’ problem [27]. But it is very rare to find work dealing with several levels of fusion or fission, with the aim to make precise what these processes are. We want to propose a unified vision of multimodal fusion and fission that is linked to our semantics and pragmatics related approach.

Multimodal fission consists of splitting the information into several parts considering the presentation aims, means and context. Now, information can be split at different levels. At the signal level, the information, considering its nature, is sent to the correct communication channel. This is typically the case for a video, the sound track being sent to the auditory channel and the visual track to the visual channel. This is also the case for a linguistic utterance accompanied by one or more deictic gestures, such as “I am putting that there” with two gestures, one for “that” and one for “there”. In this example, IMMPS must be aware of the duration of the speech synthesis in order to provide the gestures, e.g., visual feedbacks, at the right moments. Splitting and synchronizing at the signal level is then a kind of multimodal fission, and is strongly linked to the constraint-based repartition over the communication channels.

At a semantic level, the information content can be dissociated over several modalities in order to better manage its complexity and to simplify the resulting monomodal messages. One important example related to human factors consists of displaying the part of the information that requires an important amount of persistent attention, and of verbalizing the part whose only aim is to capture selective attention. Splitting at a semantic level appears as another kind of multimodal fission, which is linked to the preference-based repartition.

At a pragmatic level, the message illocutionary force can be dissociated over several modalities in order to simplify the illocutionary force of each resulting monomodal message. For instance, a message labeled with a ‘feedback inform’ presentation act can be split into two messages: a first one that verbalizes the ‘inform’ and a second one that requires the ‘feedback’ using a text box. To us, this is a third kind of multimodal fission, as important as the previous ones, although it has not been studied in the literature.

To show the relevance of such a classification into three levels of fission, it is interesting to bring together fission with fusion. In fact, multimodal fusion can also be done at three different levels. At the signal level, the coordination of the various signals into one composite multimodal signal is a first kind of fusion. At a semantic level, the fusion of the message contents is a second kind of multimodal fusion, which is strongly linked to the resolution of references to objects [14]. At a pragmatic level, the fusion of illocutionary forces corresponds to the fusion of events and is also a kind of multimodal fusion. To conclude:

- At the signal level there is **multimodal coordination** for input signal processing and **multimedia coordination** for output processing;
- At a semantic level there is **content fusion** for input message processing and **content fission** for output message processing;
- At a pragmatic level there is **event fusion** for input event processing and **presentation act fission** for output event processing.

4.2 Second step: information valorization

Once the output multimodal message has been repartitioned over the communication channels, there is the need to optimize and to valorize each piece of information within each communication channel. We can distinguish a main strategy that consists of taking into account a set of constraints and a set of preferences, and some additional strategies dealing in particular with human factors.

First, constraints have to be taken into account, with (a) the constraints that are inherent to the information, for instance the numbers of lines and columns when displaying a table, (b) the constraints that are linked to the terminal, e.g., the screen size with the same example, and (c) the constraints that are linked to the presentation environment, for instance a threshold for the ambient noise.

Second, preferences can be taken into account using a set of rules relying on: (a) the message content, we can imagine for instance display rules related to data structure, (b) the communicative act, for instance favoring a strong intensity for a 'demand' act, (c) the user's preferences, for instance displaying with a font size of 16 if it is a preference, (d) human factors, for instance exploiting the color red for an alert (because red is perceived faster than other colors).

Moreover, IMMPS should be able to optimize the information content within a modality. For instance, information can be spread out to the limits of the terminal, with rules like the following one: when displaying a picture, take all the available space. IMMPS should also have to emphasize a content, to exploit a salience. In this way, one important aim will be to adjust the communicative structure for putting one element into salience. When an avatar is used as a conversational animated agent, IMMPS may have to render emotions on contents, with the exploitation of the prosody and if necessary the multiple possibilities of the animated character. Lastly, to manage the user's attention, IMMPS should take into account the distinction between selective attention (that is captured by a transient verbalization or display) and persistent attention (that requires a persistent or permanent display).

5 CONCLUSION AND FUTURE WORK

In this paper we have proposed a set of theoretical and operational principles for the design of intelligent multimedia presentation systems. These principles and the related classifications integrate the preoccupations from work dealing with adaptation to the terminal, to the environment, and to the user. Our proposal is based on the Speech Act Theory, and in a general manner on pragmatic preoccupations. Human factors are taken into account, and the foundations for more human-oriented systems are drawn. The method we propose for information presentation relies on two main phases, the first one consisting of the repartition of information over the communication channels, and the second one consisting of the valorization of each piece of information within each communication channel.

The recommendations we provide for the design of intelligent presentation systems have to be confronted with the applicative and design contexts in order to be translated into rule systems. In this paper our aim was not to present such rule systems, but to determine the underlying main preoccupations and methods, that can be applied to every kind of human-machine interactive system and not to a particular task-oriented one. Since we have focused mainly on theoretical aspects, a future paper will focus on technical details and implementations, with the presentation of a particular applicative context (an interactive support for cooperative decision making in the domain of air traffic management, which was at the basis of some of our observations and experimentations) and the related rule systems.

REFERENCES

[1] E. André and T. Rist, 'Multimedia presentations: The support of passive and active viewing', in *Proceedings of the AAAI Spring Symposium on Intelligent Multi-Media and Multi-Modal Systems*, pp. 22–29, Stanford, (1994).

[2] Y. Arens and E. Hovy, 'How to describe what? towards a theory of modality utilization', in *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 487–494. Lawrence Erlbaum Associates, (1990).

[3] J.L. Beckham, G.D. Fabbrizio, and N. Klarlund, 'Towards SMIL as a foundation for multimodal, multimedia applications', in *Proceedings of EUROSPEECH 2001*, pp. 1363–1366, Aalborg, Denmark, (2001).

[4] N.O. Bernsen, 'A reference model for output information in intelligent multimedia presentation systems', in *Proceedings of the ECAI'96 Workshop on: Towards a Standard Reference Model for Intelligent Multimedia Presentation Systems*, Budapest, Hungary, (1996).

[5] M. Bordegoni, G. Faconti, S. Feiner, M.T. Maybury, T. Rist, S. Ruggieri, P. Trahanias, and M. Wilson, 'A standard reference model for intelligent multimedia presentation systems', *Computer Standards and Interfaces*, **18**, (1997).

[6] C. Cadoz, *Les réalités virtuelles*, Flammarion, Paris, 1994.

[7] W. Chou, D.A. Dahl, M. Johnston, R. Pieraccini, and D. Raggett. EMMA: Extensible multi-modal annotation markup language. Available at http://www.w3.org/TR/emma/, 2002.

[8] J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R.M. Young, 'Four easy pieces for assessing the usability of multimodal interaction: the CARE properties', in *Proceedings of INTERACT'95*, Lillehammer, Norway, (1995).

[9] J.J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, 1979.

[10] P. Grice, 'Logic and conversation', in *Speech Acts, Syntax and Semantics (Vol. 3)*, eds., P. Cole and J. Morgan, 41–58, Academic Press, New York, (1975).

[11] J. Itten, *The Art of Colour*, Reinhold Publishing Corp., New York, 1961.

[12] C. Karagiannidis, A. Koumpis, and C. Stephanidis, 'Adaptation in intelligent multimedia presentation systems as a decision making process', *Computer Standards and Interfaces*, **18**(6-7), (1997).

[13] W. Kohler, *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*, Liveright Publishing Corp., New York, 1947.

[14] F. Landragin, 'Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems', *Signal Processing*, **86**(12), 3578–3595, (2006).

[15] I. Mel'čuk, *Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences*, Benjamins, Amsterdam, 2001.

[16] G.A. Miller, 'The magical number seven, plus ou minor two: Some limits on our capacity for processing information', *Psychological Review*, **63**, 81–97, (1956).

[17] H. Prendinger, S. Descamps, and M. Ishizuka, 'MPML: A markup language for controlling the behavior of life-like characters', *Journal of Visual Languages and Computing*, **15**(2), 183–203, (2004).

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

[19] C Rousseau, Y. Bellik, and F. Vernier, 'Multimodal output specification/simulation platform', in *Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, pp. 84–91, Trento, Italy, (2005).

[20] J.R. Searle, *Speech Acts*, Cambridge University Press, Cambridge, 1969.

[21] J.R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, Cambridge, 1979.

[22] D. Sperber and D. Wilson, *Relevance. Communication and Cognition*, Blackwell, Oxford, 2nd edn., 1995.

[23] R. Stevenson, 'The role of salience in the production of referring expressions: A psycholinguistic perspective', in *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, eds., K. van Deemter and R. Kibble, 167–192, CSLI Publications, Stanford, (2002).

[24] A. Sutcliffe and P. Faraday, 'Designing presentation in multimedia interfaces', in *Proceedings of CHI'94, Conference on Human Factors in Computing Systems*, pp. 92–98, Boston, (1994).

[25] W3C. Multimodal interaction activity, multimodal interaction working group. Available at http://www.w3.org/2002/mmi/.

[26] W3C. Synchronized multimedia integration language (SMIL) specifications, SMIL 2.1 proposed recommendation. Available at http://www.w3.org/AudioVideo/.

[27] W. Wahlster, 'Smartkom: Fusion and fission of speech, gestures, and facial expressions', in *Proceedings of the First International Workshop on Man-Machine Symbiotic Systems*, pp. 213–225, Kyoto, Japan, (2002).

Web experience as an expansion: a perspective on covert sales from multimodal discourse analysis

Arianna Maiorani[1]

Abstract – In this paper the multimodal discourse analysis method is applied to the study of the Internet as a multimodal semiotic system. The paper is aimed at testing the validity of the functional framework used in linguistics as a perspective to study Internet multimodal communication in relation to covert-sales strategies. Two cases have been taken as examples: *The Matrix* website and the *Lord of the Rings Online* game, both inspired by successful movies and both involving a thriving online market.

1. INTRODUCTION: INTERNET AS A SEMIOTIC CHALLENGE

Internet is a most challenging as well as interesting field of research for the multimodal branch of systemic functional linguistics known as multimodal discourse analysis. Multimodal discourse analysis is based on the concept of culture as a compound of semiotic systems: language is one of them and its primary function is to communicate through the verbal mode. Systemic functional linguistics, through the model of Functional Grammar created by M.A.K. Halliday, [1] studies verbal discourse as a function-oriented form of semiosis and allows us to recognise the systemic connection between the context in which a message is created and the lexico-grammar[2] through which it is realised. Furthermore, since all act of communication is functional to a specific *context of situation*, within a specific *context of culture*, the systemic functional study of language allows us to understand how texts encode the society and culture they belong to. Thus, the systemic functional perspective allows the study of society through the language as social semiotic. In order to allow communication language construes human experience, enacts social relationships, and builds intelligible sequences of texts through which this construal and enacting is discursively organised. Halliday calls the basic functions of language respectively *experiential, interpersonal, and textual*[3]. The context of a message is formed by three basic components, *Field, Tenor* and *Mode*, which activate the basic meanings: *experiential, interpersonal,* and *textual*; these meanings are then realised in the specific lexico-grammar of the text. As a social semiotic theory of representation, the functional framework has been developed to create models through which other modes of communication can be investigated [2, 3, 4, 5, 6]. Multimodal discourse analysis studies discourse realised through different semiotic modes in a systemic functional perspective based on the three metafunctions (Fig. 1).

Internet is by nature a multimodal communicative dimension. It is an expansion of the outside world which is not subjected to the physical laws of time and space its same users have to obey. Can the functional framework be used as a tool to investigate the hyper-textual multimodal discourse? Can we use multimodal discourse analysis to investigate a systemic relationship between society, hyper-context and hypertext?

2. INTERNET COMMUNITIES AND COVERT SALES STRATEGIES: TWO CASES.

As an expansion of the world outside, the Internet world has its own communities of many different typologies[7], and has also favoured thriving community markets. Communities of game players or movie fans, in particular, can keep on watching trailers, discussing in forums, searching for information and, most of all, buying related products through dedicated websites. Websites allow a continuous updating of on line games. They also keep a movie market 'alive' well before its theatre release and long after it. The most important items of this market are clothing and gadgets and, very often, role play on line games which promise to 'expand' the movie plot. Massive Multi-player Online Role Play Games (MMORPG) are another very interesting multimodal Internet phenomenon. Players can create one or more game identities and join the other members of the community during the game sessions. The way in which the presence of a user is represented in covert-sales oriented hyper-contexts will be the focus of the following multimodal analysis, which will be performed both on *The Matrix* movie website and on *The Lord of the Rings Online* MMORPG (hereafter LOTRO) This choice has been determined by three factors: both internet products are inspired by very successful movie trilogies; both movie trilogies have developed a thriving market of by-products for fans; they both belong to the science fiction/fantasy genre of narrative events.

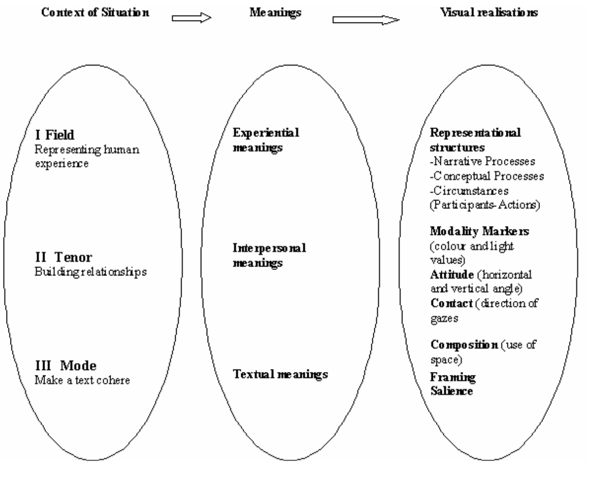


Figure 1. Functional relations between visual text and context. Based on Kress & van Leeuwen 1996; 2006.[2]

3.a *THE MATRIX* WEBSITE COVERT-SALES STRATEGY AND THE *COCKPIT PERSPECTIVE.*

The Matrix trilogy website is basically sales-oriented and advertises all items related to the Matrix world and philosophy. *The Matrix* was the first movie of the Internet era which was entirely based on the dichotomy virtual world/real world. The movies were released between 1999 (*The Matrix*) and 2003 (*The Matrix Reloaded* and *The Matrix Revolutions*). During the almost four years which separate the release of the first movie from its sequels, the website had the primary function of giving

[1] University of Bologna "Alma Mater Studiorum", Faculty of Foreign Languages and Literatures, Bologna, Italy. E-mail: arianna.maiorani@unibo.it

[2] Halliday sees the unity of lexis and grammar (*lexico-grammar*) in the realisation of all texts as the two poles of a single continuum (see Halliday 2004: 43 ff.)

[3] The term 'metafunction' is used to remind that 'function' is an intrinsic component of language: language is by nature 'functional' (see Halliday 2004: 31)

news and updates about the second and third movies, keeping the interest of the public alive. At the same time it also developed a thriving online market based on video/audio material, clothing, fan gadgets.

The website communicative strategy is based on the core of the movies plot [8]: the Matrix is a virtual world where the outside world is just replicated for those who enter it unconsciously, while it can be unlimitedly expanded for those who access it consciously. Online sales are construed as a consequence of experiencing the website as a 'door' to the Matrix world. The buyer's perspective as a traveller connected to the Matrix is construed though the use of cockpit-like environments through which the user can access different sections of the website and, most of all, through Quick Time .mov files, created by Apple to handle and reproduce simultaneous multimodal data. These files, located in the website Mainframe, are capable of containing data from different kinds of tracks (audio, visual, textual, etc.): each different track contains codecs or reference to its specific media stream, so that data of different nature can be kept separate but reproduced simultaneously. This allows the visualisation of digital environments that the site visitor can edit around his/her central point of view, which these environments implicitly reproduce. Figure 2 shows a schematic representation of how the .mov file, represented as a rotating cylinder, construes the central point of view of the web user:

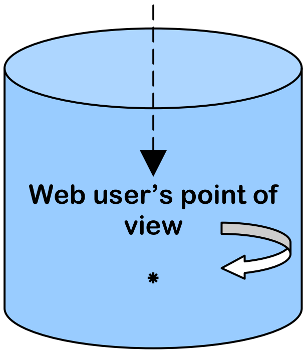
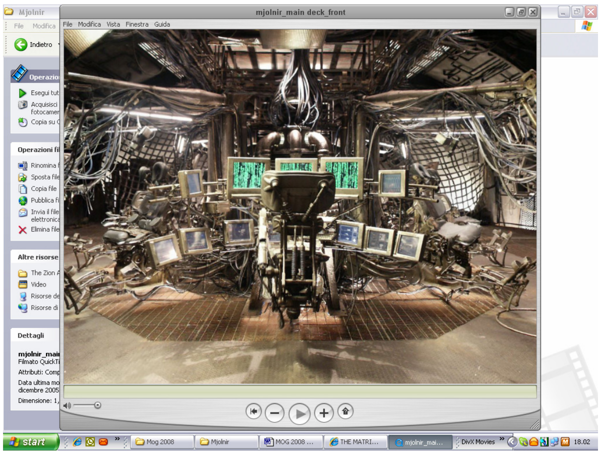


Figure 2. Schematic representation of a .mov file perspective.

Rather than displaying images or trailers to the site visitor on a classic bi-dimensional webpage, these files offer the possibility of virtually navigating in spaces developed around the implicit presence of a viewer, who, just because he/she is construed as being *inside* the Matrix, must not be reproduced by a multimodal (or, at least, visual) *alias*. A multimodal functional analysis performed on this kind of .mov generated environments reveals the intrinsic sales-oriented quality of this typology of files: in terms of experiential metafunction they allow the web user to visit a tridimensional environment by clicking on screen buttons (Fig. 3) provided on the online console frame. The user can also click on some highlighted spots (screens, levers, etc.) to scroll data pages or be directed to consecutive environments. All items that the user experiences during this visit are linked to items on sale on the website. In terms of interpersonal metafunction, they are structured to construe the web user as having the power to decide which environment to visit, which hyperpath to take; however, the user is always guided by predisposed mouse sensitive links that interact with him through beaming lights and beep sounds signaling sales oriented points of interest. In terms of textual metafunction, the web user is always construed as being the implicit centre of the .mov file multimodal output, the one around whom the whole Matrix world and its market rotate.

Thus, these files elicit interaction aimed at purchasing items on sale online as a consequence of having been *into* the Matrix world, rather than having watched it from outside. For this reason, I have defined the .mov files that have revealed this specific functionality as *Elicitors*. Furthermore, in order to experience these files, the web user has to purchase and download a specific software from the Apple website, which is a business partner of the Matrix market.



Fiure3. Example of *Elicitor* on the Matrix website.

Elicitors are linked to each other as a sort of 'net within the net'. They construe the web user as *interactively represented Participant*: that is, he/she is implicitly represented as 'point of view' when making the hyper-context move and activating links around his/her non-physical presence. He/she is interactively construed as being already *inside* the Matrix and constantly put in the position of 'piloting' his/her experience in the hyper-context of the website[4]. The functional analysis of these multimodal environments has therefore revealed that *Elicitors* are .mov files specifically used for covert sales multimodal on line strategies and that their functions are determined by sales-oriented hyper-contexts related to entertaining events and products. This specific multimodal strategy is based on what I will define as the *cockpit perspective*.

Representing the web user as if he/she were *inside* an Internet environment is a strategy used also for selling subscriptions to MMORPG. The following paragraph will show how, in this case, perspective changes according to the on-sale product.

3.b LOTRO: THE *DIRECTOR'S PERSPECTIVE*.

As a Massive Multi-player Online Role Play Game, *The Lord of the Rings Online*, can be said to be the by-product of the cinematographic transposition of a verbal text. It has been chosen as an example for the repeated transmediality of Tolkien's verbal text and therefore for its worldwide renown.

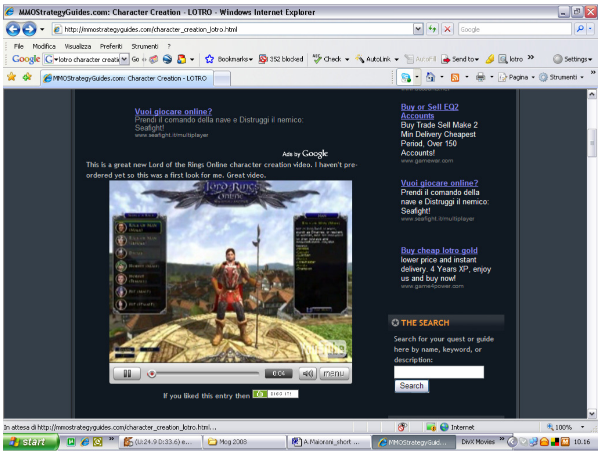
In order to start playing, the on line player has to create the character who will be his/her own representation in the multimodal hyper-context of the online game. This *alias* will have to be able to communicate with game generated characters and other players' representations[5] [9]. The creation of a character is achieved through three consecutive attributive phases which imply the choice of a race and gender, the choice of a class and of a name, a geographical background and specific physical features. In each phase, the player has to make choices within a framework displayed on a webpage whose central focus is the character under construction, surrounded by a suggestive landscape (Fig.4) .

Choices are presented through verbal texts within Celtic style templates and then visualised on the character during the creation process. The system of choices offered on the left side is paradigmatic: all elements are available at the same time and in the same hyper-context. The system of choices on the right side

[4] With respect to the definition of a web surfer interacting with a hypertext as an "ergodist", discussed by A.K.K Chiew in O'Halloran 2004: 132 ff., the path described by visitors of *Elicitors* is characterised by a specific sales-oriented disposition of *lexia* (complex multimodal scrolling pages) and links that aim at covertly guiding the user towards purchase.

[5] J.P.Gee (2004) associates what he calls the "inherently social" nature of video games to their usefulness as learning tools. As in this article, in terms of social interaction patterns, he doesn't make any difference between game generated interactive characters and players' aliases.

is syntagmatic and depends on the choices made within the left-side system, which determine a series of possible features combined with character/gender specific powers and tools.


MMOStrategyGuides.com: Character Creation - LOTRO - Windows Internet Explorer
http://mmostrategyguides.com/character_creation_lotro.html
Google
File Modifica Visualizza Preferiti Strumenti ?
Google lotro character creati Go
Bookmarks 352 blocked Check AutoLink AutoFill Send to lotro Settings
MMOStrategyGuides.com: Character Creation - LOTRO
Pagina Strumenti
Ads by Google
0:04
menu
DIGG IT!
THE SEARCH
Search
In attesa di http://mmostrategyguides.com/character_creation_lotro.html...
Internet
100%
start
(U:24.9 D:33.6) e...
Mog 2008
A.Maiorani_short ...
MMOStrategyGuid...
DivX Movies
10.16

Vuoi giocare online?
Prendi il comando della nave e Distruggi il nemico: Seafight!
www.seafight.it/multiplayer

This is a great new Lord of the Rings Online character creation video. I haven't pre-ordered yet so this was a first look for me. Great video.

If you liked this entry then

Buy or Sell EQ2 Accounts
Buy Trade Sell Make 2 Min Delivery Cheapest Period, Over 150 Accounts!
www.gamewar.com

Vuoi giocare online?
Prendi il comando della nave e Distruggi il nemico: Seafight!
www.seafight.it/multiplayer

Buy cheap lotro gold
lower price and instant delivery. 4 Years XP, enjoy us and buy now!
www.game4power.com

Search for your quest or guide here by name, keyword, or description:

Figure 4. Creation of a LOTRO character.

Choices will enable the character to perform specific actions, establish specific relationships and appear with a specific aspect: the character is therefore a multimodal text that realises the meanings activated by the three systemic metafunctions.

The context of situation in and by which this multimodal text is created is determined by the *Lord of the Rings* transmedial text experience of the game player. Thus the player is required to create a multimodal text which is a product and a process[6] that in order to be developed and enhanced will need more experience of the same hyper-context in which it will work: the player will therefore have to keep on learning about the game by buying subscriptions and enhancing computer hardware with up to date audio-visual processors, especially of the ATI Radeon series.. The initial phase of the game enacts in this way a covert-sales strategy based on the multimodal representation of the player in the game world.

After having completed the character and entered the game, several icons appear at the bottom of the screen during the whole game session that link the player to text windows displaying the character's resources. In the top right corner a compass is visualised, in the top left the character status is displayed. Furthermore, all characters speak mainly through written texts[7] appearing on the screen either as verbal pop-ups or in text windows.

The on line game page looks therefore like the display of a digital camera. Background music also follows the various audio/visual performances: it can also be disactivated as when using a camera.

Unlike what happens with the *Elicitors*, in the LOTRO environment it is the web user's point of view that rotates around the character. The player is construed as a movie director moving his/her character within tri-dimensional hyper-environments. The environments rotate around the character that represents multimodally the player in the game hyper-context. An example of this perspective is shown in Figure 5.

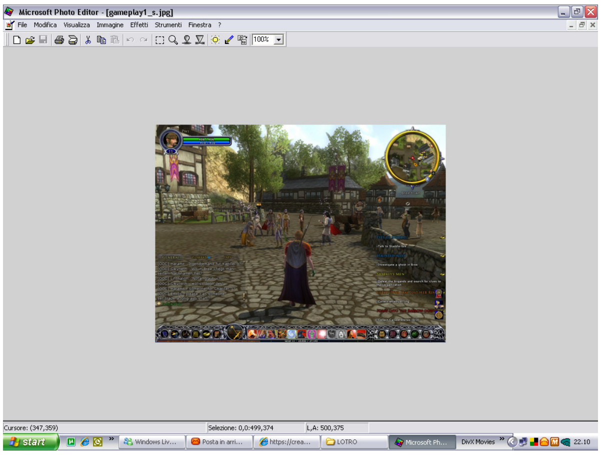


Figure 5. An example of *director's perspective* in a LOTRO screenshot.

4. CONCLUSIONS

The multimodal analysis performed on both *The Matrix* website and the LOTRO game has shown how the functional framework can be applied to an analysis of different multimodal texts realised in the Internet dimension. Analysis has focused on entertainment internet products where covert sales strategies have been enacted.

Two main perspectives have been highlighted through which the web user is construed in the multimodal hyper-context: the *cockpit perspective*, which implicitly incorporates the web user within the multimodal hyper-context, and the *director's perspective,* which construes multimodally the web user's presence as an *alias* in the hyper-context directed by a decision-making viewer from 'outside'.

It has also been shown that these two perspectives are linked to covert-sales strategies merchandising products related to a movie trilogy market and to a MMORPG which, in itself, is part of a movie-and-book trilogy market. Furthermore, it has been observed that the enactment of these strategies is linked to the use of specific files and processors that imply, for the web user and game player, the visit of partner companies websites and the purchase of their on line software products.

This analysis has been performed on finished products: it would be interesting to study if and how the multimodal systemic functional perspective would effectively influence the creation of different or enhanced multimodal output devices. Would a pre-existent knowledge of the functional theory of communication change or influence the display of internet environments or the structuring of an *alias* creation process? And if so, would it allow the production of a different kind of online identity? And what kind of internet communicative strategies would this serve? These are all questions that only future interdisciplinary research can try to answer.

REFERENCES

[1] M.A.K. Halliday, *An Introduction to FunctionalGrammar*, 3rd ed., Arnold, London, 2004 [1994].

[2] G. Kress and T. Van Leeuwen, *Reading Images. The grammar of visual design*, 2nd ed., Routledge,London and New York, 2006 [1996].

[3] G. Rose, *Visual Methodologies*, 2nd ed., SAGE, London, 2007 [2001].

[4] T. van Leeuwen, *Speech, Music, Sound*, Macmillan, London, 1999.

[5] E. Ventola/ C. Charles/ M. Kaltenbacher (eds.), *Perspectives on Multimodality*, John Benjamins, Amsterdam – Philadelphia, 2004.

[6] K.L. O'Halloran, *Multimodal Discourse Analysis: Systemic Functional Perspectives*, Continuum, London, 2004.

[7] J. Slevin, *The Internet and Society*, Polity Press, Cambridge,2000.

[8] A. Maiorani. 'Reloading' movies into commercial reality: A multimodal analysis of *The Matrix* trilogy's promotional posters. In: *Semiotica* 166- 1/4: 45-67 (2007).

[9] J.P. Gee, *What Video Games have to teach us about learning and literacy*, Palgrave Macmillan, New York, 2004.

[6] See the complex notion of text as instantiation and realization in Halliday 2004: 26 ff. A text is an instantiation of the language system functional to a context which, on its turn, it contributes to change.

[7] Interestingly, this unexpectedly extensive use of written text is a characteristic typical of mulyimodal advertising texts; as H. Stöckl (in Ventola et al. 2004: 21 ff.) underlines: "Print advertising is a textual genre whose reliance on language-image-combinations is almost obligatory." [5]. Indeed, some of the game generated characters seem to speak by slogans.

Multimodal content representation for speech and gesture production

Bergmann, Kirsten and **Kopp, Stefan** [1]

Abstract. This paper presents a computational perspective on the joint production process for speech and gesture. Based on empirical evidence indicating a mutual influence of speech and gesture in utterance production, we propose an interface between imagistic and propositional knowledge at the level of content representation. This is integrated into a generation architecture in which the planning of content and the planning of form across both modalities proceeds in an interactive manner.

1 INTRODUCTION

When giving spatial explanations, most humans inevitably move their hands and arms to gesture. These gestures have been subject of extensive empirical and theoretical research and current conjectures entail both, that gestures are used for communicative purposes [5, 35] and that they are connected to the cognitive processes involved in the speaker's current mental activity [22, 17]. This twofold role particularly pertains to representational gestures like iconics, which to some extent resemble the entity being referred to and thus (seem to) contribute designated meanings to the communicated message.

Albeit their prominence in natural spatial communication, work on computational models of speech and representational gesture are still sparse, with the majority of approaches targeted at automatic recognition and understanding of such behavior. In this paper we present work towards a computational model of the *production of speech and iconic gesture* which could drive multimodal behavior of embodied agents like the virtual human Max [26]. Computational approaches to producing multimodal behavior with artificial agents can be characterized as conceiving of the generation problem in terms of three consecutive tasks (cf. [42]): figuring out what to convey (content planning), figuring out how to encode it (micro-planning), and realizing the behaviors (surface realization). Almost all existing systems have either circumvented the modeling of large parts of this generation process, by focusing on non-representational gestures that can be selected and sequenced from a statistical model of a particular speaker [20], i.e. essentially neglecting the meaning a gesture conveys, or they have simplified matters by utilizing predefined lexicons of both words and gestures, i.e. pre-fabricating the meanings and forms of gestures [9, 14].

In previous work [27, 25], we have devised and applied a framework to analyze gestural images into semantic units (*image description features*), and to link these units to morphological features of hand gesture (handshape, trajectory, etc.). This feature-based framework allowed for implementing an integrated micro-planner for multimodal directions that derives the form of both natural language and gesture directly from communicative goals. The goals, along with the entire content of the communicative intention including all the image description features of the objects and events involved, were coded in one common propositional format. This parsimonious approach afforded an integrated micro-planning stage, in which the meanings of words and on-the-spot-created gestures could be unified and set against each other. Yet, there are also several shortcomings of this appraoch. Devising a sentential representation of a complex visuo-spatial content requires a large ontology that is difficult to set up and introduces an arbitrary degree of discretization and schematization, e.g. by prescribing qualitative symbols for different extents. Further, deciding upon a particular description entails operations like perspective-taking that are hard, if at all, to realize efficiently upon a symbolic spatial representation. Finally, although there are proponents of an amodal conceptual representation of space, especially in consideration of the schematization of spatial language (e.g. [15, 30, 49]), a large body of literature argues for the cognitive plausibility of an analog representation of space and imagery, which are widely assumed to underlie gesture in prevalent psycholinguistic theories. Indeed, the very opposition between two differing modes of thinking, one linguistic-propositional and the other imagistic, has been argued to fuel speech and gesture production [33].

In this paper we present work towards a computational model of the production of speech and iconic gesture, with a focus on landmark descriptions as found in direction-giving. The model is novel in several respects. First, it comprises two different, but interconnected, representations, a logic-based propositional one and an analog visuo-spatial one, and grounds the production of language and gesture in them. Second, it comprises an approach to micro-planning iconic gestures, which exceeds previous attempts on this problem in that it does not directly map visuo-spatial meaning onto gesture forms, which has so far found weak empirical support. Instead, it additionally incorporates the notion of more general *representation techniques* into this mapping. Third, it offers an account not only of micro-planning but also of content planning, in which the two representational formats interact in order to figure out portions of multimodal meaning that can be turned into coherent multimodal behavior. Finally, it rests upon the assumption that these two planning stages cannot be separated into successive stages of deriving coordinated speech and gesture, as in other systems, but that one must model a more interactive, bi-directional production process.

We start in Section 2 with a review of findings that evidence a mutual influence of speech and gestures. Section 3 introduces the conceptual design of our approach to model this production process by employing two different kinds of meaning representation, which are interconnected by multimodal concepts to constitute a level of multimodal meaning. This level informs the derivation of coordinated

[1] Artificial Intelligence Group, University of Bielefeld, Germany, email:{kbergman, skopp}@techfak.uni-bielefeld.de

speech and gestures. In Section 4 we conclude with a discussion how the simulation model accounts for the empirical data.

2 SPEECH AND GESTURE IN MULTIMODAL COMMUNICATION

Building computational models of multimodal human behavior requires a detailed conception of the mechanisms that underlie the production of speech and gesture. Several different hypotheses have been put forward to explain these mechanisms, all of which implicating differing psycholinguistic models of speech and gesture production. Particularly controversial among these models is the point of origin from where the relationship of speech and gesture arises. Building upon the assumption that gestures are generated "prelinguistically", Krauss et al. [29] assume gestures to derive from spatial imagery. While there is no influence of the linguistic production processes onto the gesture in this model, the readily planned gesture facilitates lexical retrieval through a kind of cross-modal priming. In de Ruiter's Sketch Model [10, 11] the common origin of speech and gesture production is located in one single central process which is responsible for the selection and distribution of information to be expressed. One submodule performs content planning for gesture ("Sketch Generation"), while another submodule plans the message to be verbalized ("Message Generation"). There is no further interaction between the planning processes after this integrated content processing in the conceptualizer. In a different account, Kita & Özyürek [22] assume an online-interaction between the message generation process for speech ("Message Generator") and the process generating the gestural content ("Action Generator"). According to this, a gesture is generated during the conceptual process which organizes spatio-motoric imagery into a suitable form for speaking. Another proposal, the Growth Point Theory [34], claims that gestures arise from growth points which are units combining imagery and categorial content. This combination is unstable and initiates cognitive events through which speech and gesture unfold. In this process, bidirectional interactions take place between language and imagery (for a discussion of the above models, see [11]). Finally, a novel model has been proposed by Hostetter & Alibali [17], the Gestures as Simulated Action framework. This framework emphasizes a perspective on how gestures may arise from processes of embodied cognition. It thus situates gestures in a larger context including mental imagery, embodied simulations, and language production. According to this view, whether a gesture is produced or not depends on the strength of activation of a mentally simulated action, the height of the speaker's current gesture threshold, and the simultaneous engagement of the motor system for speaking.

Contingent upon the differing origins of gestures, these models imply different mutual influences of speech and gesture during the multimodal production process. To computationally generate speech and gesture we must capture these mechanisms, and thus the question where to locate a gesture's origin during the process of utterance production is of major relevance. The following two sections review empirical evidence for the mutual influence of speech and gesture.

2.1 How speech influences gesture

Coverbal gesture has long been considered a by-product of language production [18]. Indeed, on a closer inspection, gestures are found to be influenced by the conceptual, syntactic, and lexical structure of concomitant speech. In a cross-linguistic study, Kita & Özyürek [22] could show that the packaging of content for gestures parallels linguistic information packaging. Speakers of Japanese, Turkish and English had to re-tell cartoon events for which their languages provide differing means of encoding. English speakers, for example, used the verb "swing" for a character's action, encoding an arc-shaped trajectory, while Turkish and Japanese speakers employed a trajectory-neutral, change-of-location predicate such as "move". Gestures follow this packaging in a way that Japanese and Turkish speakers were more likely to produce straight gestures, whereas most English speakers produced arced gestures. In another cartoon-event, the character rolled down a hill. Again, speakers of English typically described this by combining manner and path of the movement in a single clause (e.g. "he rolled down"), accompanied by a single gesture encoding both semantic features. In contrast, Turkish and Japanese speakers encode manner and path separately in two clauses (e.g. "he descended as he rolled") and also used two separate gestures for these two features.

Further evidence along the same line comes from a study comparing the gestures of native Turkish speakers, who are at different levels of proficiency in English [39]. In contrast to beginners, the advanced L2 speakers typically encoded manner and path information in one clause, just as native English speakers do, and the gestures they used followed this packaging of information. Another recent study showed that this effect also occurs when stimulus events are manipulated in order to make first language (English) speakers produce one-clause and two-clause descriptions of manner and path [23].

Recently, Gullberg analyzed how the semantics of placement verbs affect gestures [13]. In French, for example, the verb "mettre" ("put") as a general placement verb is typically accompanied by gestures encoding only the placing movement. By contrast, in Dutch there are different verbs for putting things somewhere, e.g."zetten" ("set") or "leggen" ("lay"). The gestures accompanying these verbs typically represent the types of the object to be placed in addition to the placing movement.

Impact on gesture has also been observed when problems in verbalization occur. Bavelas et al. [5] investigated how the information distribution between speech and gesture is influenced in a description task by the degree of verbal encodability of a stimulus. The more difficult it was to describe the stimulus verbally, the more non-redundant gestures were used. Bavelas et al. conclude that humans compensate for verbal encoding problems by using gestures, which convey complementary information.

2.2 How gesture influences speech

Perhaps more surprisingly, the use of gestures conversely has an impact on the linguistic utterance they accompany. Kita's *Information Packaging Hypothesis* states that "gesture helps the speaker organize information in a way suitable for linguistic expression" [21, p.180]. This hypothesis claims that gestural influence on speech takes place at an early stage of utterance production. Different from the default thinking for speaking, the production of gestures supposedly involves a form of spatio-motoric thinking [43]. Support for this hypothesis comes from Alibali et al. [3] who found that placing higher demands on the conceptualization of speech (explaining vs. describing an environmental change), with comparable lexical access, results in higher numbers of gestures that convey perceptual dimensions of objects, as well as gestures that convey information not conveyed by the accompanying speech. Likewise, in another spatial description study, in which a configuration of dots with or without connecting lines were to be described, Hostetter et al. [17], supporting this hypothesis, found that, in more demanding conditions, what was more likely to be used was representational gestures and words that can be relatively easily accessed. These findings are taken as suggestive of an active role of representational gestures to help breaking down a complicated image into parts that can be organized into speech.

The influence of gesturing on language is perhaps most clearly seen in studies that prohibit gestures. Alibali & Kita [2] found that preventing children from gesturing has an influence on the content of their utterances, such that they are more likely to give non-perceptual explanations, i.e., describe states and actions from the past and make less reference to the physical properties of present objects (e.g. size and form). Alibali et al. conclude that gesturing seems to explore and structure information about the current visuo-spatial surrounding and make it available for verbal explanations. Another recent prohibition study during descriptions of motor tasks [16] showed that speakers who were free to gesture produced more semantically rich verbs (e.g., "cross", "fold", "replace") than participants who were prohibited from gesturing. Those were more likely to use generic verbs such as "take", "put" or "get". Further evidence in line with the idea that gesture has an impact on speaking on a level corresponding to content planning, is provided by a (unconstrained) picture description study [35], which revealed that the decision to gesture influences decisions about what is explicitly mentioned in both, concurrent and forthcoming speech. Speakers who voluntarily produced iconic gestures representing spatial relations omitted more required spatial information from their verbal descriptions than speakers who did not gesture.

In addition to these indications for a role of gesture in content planning (in particular, information packaging), another line of evidence suggests that gesturing helps the process of speaking more generally, and the task of micro-planning specifically. Allen [4] reports that individuals who gesture tend to speak more rapidly than others who gesture less frequently. Rauscher et al. [41] found that prohibiting gestures in cartoon narrations made speech less fluent. Specifically in phrases that include spatial prepositions, speakers spoke more slowly. In addition, speakers produced a higher proportion of filled pauses (e.g. "um", "uh"). The other way around, when restrictions were imposed on speech, e.g. to avoid all words containing the letter "c" or to use obscure words, speakers were found to produce more substantive gestures. Rauscher et al. interpret their findings as evidence of gestures facilitating lexical retrieval, which receives further support from other studies. Morsella & Krauss [37] found an increased gesture rate when lexical access is difficult, i.e. when describing visual objects from memory or when describing objects difficult to remember and encode verbally. Further, Morrel-Samuels & Krauss [36] found that the asynchrony between the onset of a gesture and its lexical affiliate is greater for less familiar lexical items. However, despite the above evidence for a functional role of iconic gestures in lexical retrieval, Beattie and Coughlan [6] found that–in a tip-of-the-tongue situation–participants who were prevented from gesturing, were more successful in the recall of words than others who were free to gesture. Although this difference was not significant, the result clearly odds with the theory of gestures facilitating lexical access.

3 COMPUTATIONAL MODELLING

The above sections have reviewed empirical evidence for a number of phenomena of multimodal behavior. On the one hand, information packaging for iconic gestures parallels linguistic packaging. A computational model, therefore, has to consider in particular the ways in which the syntactical and grammatical structures of a language are reflected in its conceptual structure, and the influence of this on the content of coverbal iconic gestures.

On the other hand, findings indicate a significant impact of representational gesture on speaking. This concerns the content planning and conceptualization stage, as well as lexical access. We note, however, that the effect on conceptualization and information packaging seems to be primordial and may also account for facilitated lexicalization (see [21] for a discussion). For computational models this means that we need to have an account of how "gesture thinking" interacts with the propositional thinking that is assumed to be necessary for speech.

In our conception of a computational model for speech and iconic gesture generation, we adopt the *Interface Hypothesis Model* by Kita & Özyürek [22] (IH model, henceforth) as a starting point. The IH model seems best suited, for it parallels the conclusions we have drawn above and lays out a production architecture that provides the required level of interactivity, while retaining a modular structure that lends itself to realization in computer simulations of artificial characters.

The IH model provides a layout for modeling how multimodal behavior is entrenched in multimodal thinking. Unfortunately, just as all other psycholinguistic accounts, it does not offer a detailed enough account of the visuo-spatial meaning retrieved from imagery and the process of turning it into concrete gesture morphology–questions one needs to answer in order to arrive at a prescriptive model that can be operationalized. To narrow down the scope of this problem, we will focus in the following on the generation of object and scene descriptions as found in route directions, leaving e.g. the explanation of actions for future work.

3.1 Representation of content

The representation of domain knowledge which is underlying the production processes is of major relevance. In his dual coding theory, Paivio [40] distinguishes two functionally independent systems, verbal memory and image memory, with associative links within each memory and possible referential links across the two systems. Imagery code is assumed to primarily represent shape and spatial and spatio-temporal relationships (rather than purely visual properties such as color or brightness). Verbal code was originally taken to be some form of inner speech in a natural language, which gave early rise to criticism and alternative suggestions to construe the two codes as imagery and conceptual ("mentalese") rather than imagery and English (e.g. [19]). Along this line, elements of the imagery code represent what they are images of; elements of the verbal code represent what words mean.

In line with this distinction, existing production models agree on the fact that speech and gesture are derived from two kinds of representation (spatial and propositional). Thus, going beyond previous approaches in computational modelling, which are solely based on a propositional representation (e.g. [27]), we adopt the dual coding perspective for our content representation. In addition, we propose an interface between imagistic and propositional knowledge. Thus, the representational layer underlying the production of multimodal behavior consists of three parts, (1) the imagistic description, (2) propositional knowledge, and (3) pairings of imagistic knowledge and propositions (see Figure 1). These three structures are organized as a multimodal working memory on which all modules involved in the generation process operate concurrently.

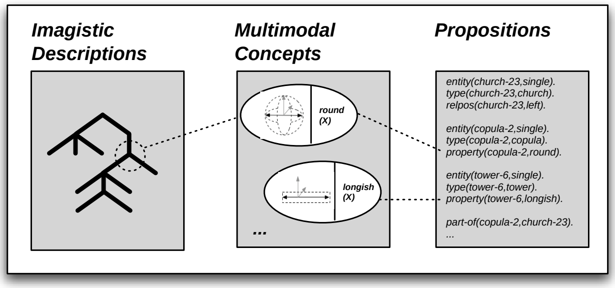


Figure 1. Three meaning representations for a chunk of multimodal utterance.

3.1.1 Imagistic Descriptions

We employ a hierarchical model of visuo-spatial imagery called IDT (Imagistic Description Tree)[46, 44]. The IDT model has been developed based on an empirical study to capture the imagistic content of shape-related gestures in a gesture interpretation system. Thus it is designed to capture all meaningful visuo-spatial features one predominantly finds in shape-depicting iconic gesture. The important aspects include (1) the hierarchical structure of shape decomposition (cf. [8]), with an abstract model of shape on each intermediate level, (2) the extents in different dimensions as approximation of shape, and (3) the possibility to leave certain dimensional information underspecified.

Each node in an IDT contains an Imagistic Description (IMD) which holds an object schema representing the shape of an object or object part, respectively [45]. Object schemata contain up to three axes representing spatial extents in terms of a numerical measure and a dimensional assignment value like ”max” or ”sub”, classifying an axis' extent relative to the other axes. Each axis can cover more than one dimension to account for rotation symmetries (becoming a so-called “integrated axis”). The boundary of an object can be defined by giving a profile vector that states symmetry, size, and edge properties for each object axis or pair of axes (cf. [32]). The size property reflects change of an axis extent as one moves along another axis; the edge property indicates whether an object‘s boundary consists of straight segments that form sharp corners, or of curvy, smooth edges. The links in the tree structure represent the spatial relations that hold between the parts and wholes, and are numerically defined by transformation matrices. It is thus possible to represent decomposition and spatial coherence.

One particular strength of the IDT model for our purpose is the possibility to represent underspecification and vagueness, both of which are immanent in gesture. Dimensional underspecification (e.g. when representing a 2D circle or simply a 1D breadth) is given when the axes of an object schema cover less than all three dimensions of space. Vagueness can hold with respect to the extent along a certain dimension (e.g. when representing something ”longish”) or the exact composition of a shape out of detailed subparts (e.g. when representing a church, without being able to recall all its single parts or their geometrical features).

3.1.2 Propositional representation

The second part of working memory is a propositional knowledge base which consists of logical formula based on a formal ontology of domain knowledge. A representation of knowledge to be drawn upon by the speech formulation processes needs to be designed to fit the needs and affordances of natural language. As discussed above, this requires a proper representation of spatial knowledge as well as conceptual background knowledge about the considered entities. As common in computational approaches to language semantics, we employ a propositional format for this, i.e. knowledge is encoded in terms of propositions over symbols that represent objects and relations, according to a given ontology. Since we focus on objects descriptions, the spatial knowledge to be captured pertains to objects, their geometrical properties, and the relations between them. The representation system thus consists of logical formulae based on a formal ontology, which encompasses entities (houses, streets, etc.) and their properties (proper name, color, quality, size, shape etc.). The entities are connected among each other by different types of relational links, such as taxonomic (is-a), partonomic (part-of), or spatial relations (on-top-of, left-of). Such a relation represents the core element of each (preverbal) message to be processed by the Speech Formulator, along with propositions stating further information about the subject entities of the particular relation.

3.1.3 Multimodal concepts

The third part of the content representation are a number of *multimodal concepts* which are bindings of IDTs with corresponding propositional formulations (see Figure 2). The imagistic parts of these multimodal concepts are characterized by underspecification, i.e. they contain more than one alternative interpretation and thus represent abstract concepts. It is this underspecification which draws the distinction between the imagistic part of the multimodal concepts and the imagistic description of concrete spatial objects. For example, the property of being ”longish” can be represented as an underspecified IDT in which one axis dominates the other one or two axes, as well as in terms of a logical formula (e.g. ”longish(X)”). Such an underspecified IDT can be matched with any other IDT by means of a formal graph unification algorithm as described in [44]. Currently, multimodal concepts for dimensional adjectives (longish, round, tall, etc.), stereotyped object shapes (box, circle, etc.), and basic spatial relations (right-of, side-by-side, above, etc.) are predefined as part of long-term memory. Matching of IDT-structures is realized by the unification procedure which also determines the necessary geometrical transformations for one IDT.

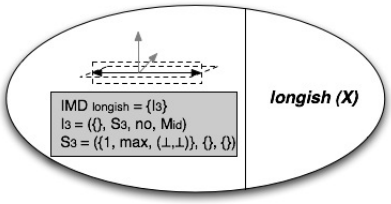


Figure 2. Multimodal concept corresponding to the property “longish”.

3.2 Generation model

3.2.1 Overall architecture

The three representations described above form the basis for simulating speech and gesture generation. Figure 3 shows an overview of the model and its underlying architecture [24]. It is inspired by, but extends and substantiates Kita & Özyürek's IH model in several ways. First, we conceive of four processing modules to be involved in content planning and micro-planning of speech and gestures: *Image Generator*, *Message Generator*, *Speech Formular*, and *Gesture Formulator*. That is, in contrast to the IH model, we embrace the idea advocated in other production models, e.g. by de Ruiter [10], of a functionally separable module that turns visuo-spatial imagery into gestural form (see [25] for a discussion). As a consequence, IH model's Action Generator is replaced by two components, one for activating visuo-spatial imagery and picking imagistic features from it (the Image Generator), and one for turning these features into gesture forms (the Gesture Formulator). In addition, two dedicated modules, Motor Control and Phonation, are in charge of surface realization of synchronized speech and gestures. Further components of the model consist of a discourse model as well as distinct long-term memories for imagery and propositional knowledge.

Second, we assume that all modules involved in the generation process operate concurrently on a central working memory, realized using a globally accessible blackboard. The overall production process evolves by each module observing entries in the working memory, taking local action if necessary, and modifying existing entries or posting new entries in result. As in the IH model, the production process is finished when all entries associated with a multimodal communicative act stabilize and form specifications of verbal and gestural acts have been formed in working memory. In this sense, the interaction between these four modules can realize content planning and micro-planning in a highly interleaved and interactive manner. This also enables bottom-up processes in both modalities.

Third, working memory is organized into units of multimodal thinking for speech and gesture (see Figure 1). Each unit incorporates all structures of meaning and form involved in one multimodal delivery, i.e., ultimately, one intonation phrase and one gesture phrase [33]. This includes representations for meaning to be conveyed as well as representations of the linguistic and gestural surface forms to be realized for this. As described in Section 3.1, three types of meaning representation are maintained, notably for visuo-spatial imagistic meaning (a single IMD or sub-trees of an IDT), propositional meaning, and multimodal concepts that act as interface between the former two.

It is normal that a number of working memory units are entertained at the same time on the blackboard. This is to account for the fact that the modules do not operate on the same level of utterance construction. For example, an IMD selected by the Image Generator may require the Gesture Formulate to employ a complex gesture consisting of two successive gesture strokes. In this case, the Gesture Formulator can, just as each of the four modules, introduce a new unit for the second stroke and its meaning, thereby preparing grounds for speech to follow accordingly.

3.2.2 Image and Message Generation

Image and Message Generation are content planning processes which perform information selection and perspective assignment. This first step in translating a thought into speech and gesture is similar for speech and gesture production [11, 12]. The production of one chunk of multimodal object description starts upon the arrival of a new communicative goal in the generator modules.

The Image Generator accesses the IDTs stored in long-term memory, which constitute the agent's experiential imagistic knowledge of the world, and activates the imagistic descriptions (IMDs) of all objects involved in the communicative goal. This activation generally leads to import into the respective working memory structure. Likewise, the Message Generator starts by selecting knowledge from propositional long-tem memory and asserts the selected facts into working memory.

For IMDs with a significantly high activation, the Image Generator performs spatial perspective taking. That is, it determines which spatial perspective to adopt towards the objects. Direction-givers usually adopt either a route (1st person) or survey (3rd person) perspective [31], with frequent switches between the two. For simplicity, we currently assume that our agent adopts the more prominent route perspective in describing an object. Still, since the IDT is defined in a world coordinate frame, the Image Generator has to figure out how the objects look from the particular point of view adopted and along that particular view direction. This operation is directly implemented as a transformation of object schemas.

The Image Generator tries to map the perspective IMD onto the imagery poles of multimodal concepts in long-term memory. If this succeeds, the corresponding multimodal concept is added to the working memory chunk. Likewise, activation of certain propositions spreads out to the multimodal concepts if they unify with their propositional pole. Of particular importance to this process of cross-modal activation is the fact that neither the original IMD nor the original propositions have to be identical with the respective pole of a multimodal concept to match. Instead, a *similarity value* between 0 and 1 is calculated by comparing the two IMDs, and a multimodal symbol is selected in dependence of a customizable threshold of similarity. Crucially, multimodal concepts are tested in inverse order of their underspecification. In other words, concepts with highly underspecified IMDs, like ”longish” or ”thing”, are tested last. That is, in case no more specific multimodal concept has matched before, it is still possible to activate meaning corresponding to a general linguistic expression, and to leave it to a complementary gesture to compensate for this (as observable in humans; see Sect. 2).

3.2.3 Speech Formulation

The Speech Formulator monitors the working memory and carries out sentence planning for each set of propositions posted by the Message Generator. As in previous work [27, 25], we employ the SPUD system [47], a grammar-based micro-planner using a Lexicalized Tree Adjoining Grammar (LTAG), to generate natural language sentences. Figure 4a shows an example of an utterance tree of a verbal description generated by our system. In contrast to previous work, however, we do not employ this micro-planner for constructing a tree for the whole multimodal utterance, employing a specially introduced node type in the LTAG formalism. This ‘sync’-node was used to combine a linguistic constituent with a gesture, which then had to be produced in synchrony. Here, we avoid extending the linguistic formalism to account for gesture integration on the level of micro-planning speech. Instead, we let this come about via the interactive content planning, in the first place, and leave it to the Speech Formulator to lay down the necessary temporal constraints for speech-gesture synchrony. To this end, SPUD's ability to provide back information on the semantics of each linguistic constituent is utilized in order to determine the lexical affiliate of the gesture that the Gesture

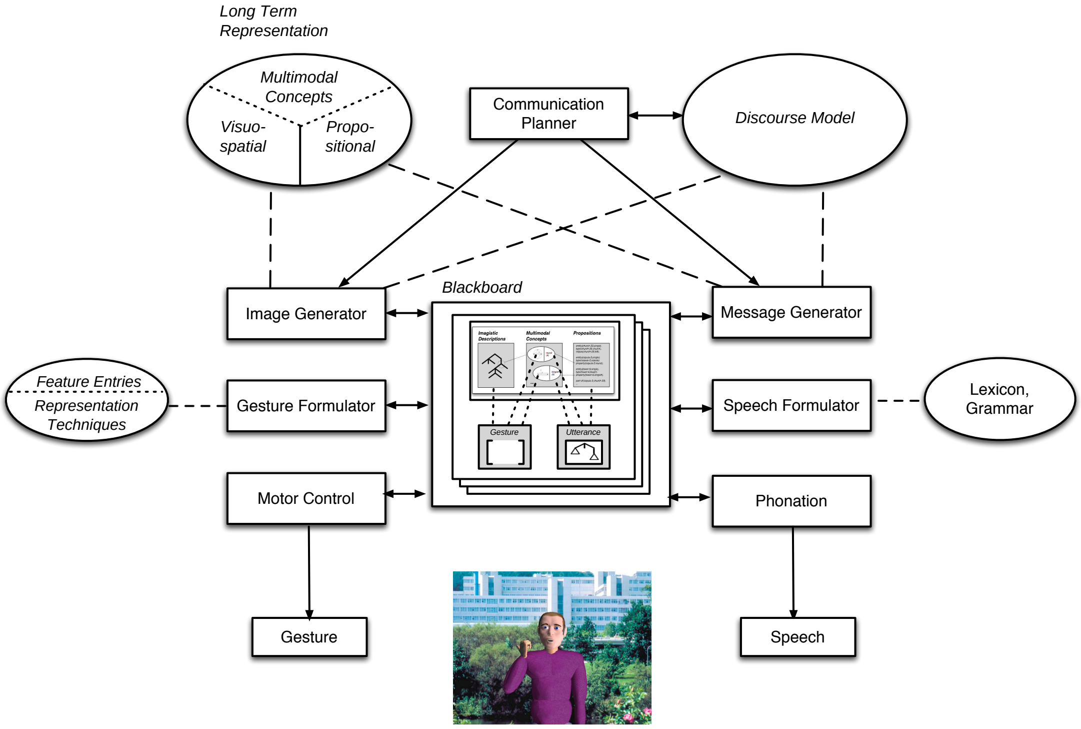


Figure 3. Overview of the proposed integrated model of speech and gesture production.

Formulator has proposed. This builds upon the multimodal concepts as interface between gesture and speech meaning.

Further, the Speech Formulator interacts with content planning by posting back the extent to which propositions are verbalizable. Upon noting this feedback, the Message Generator can retrieve or activate propositions or multimodal concepts. Similarly, the Gesture Formulator takes the visuo-spatial representation from working memory and reports back if these semantic structures can find a counterpart in gesture.

3.2.4 *Gesture Formulation*

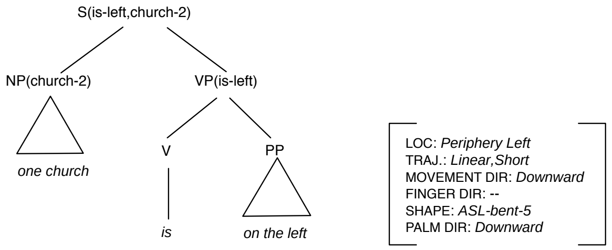
The Gesture Formulator notes when a perspective object schema is posted on the blackboard memory and then starts to look for apposite gestural renditions of the imagistic information entailed. This search is guided by the different representation techniques that people employ preferrably in these situations. Representation techniques capture the aspect that iconic gesturing does not seem to entirely follow the goal to maximize similarity with the referent model, but also brings into play conventions, experiences, and habituations, which people seem to have acquired and apply in their multimodal deliveries. Kendon [18] summarizes different attempts to classify theses techniques (e.g. [38], [48]) as a division between *modelling*, *enactment*, and *depiction*. In modelling a body part is used as a model for an object, while enactment means that the gesturing body part has features in common with the action that is being referred to. In depicting the hands ”create” an object in the air. According to the domain of reference and the purpose of the classification others differentiate in more detail.

The differentiation of these representation techniques is a promising extension of the process that transforms the mental image of an object into a gesture, beyond just relying on iconicity. Table 1 summarizes the set of representation techniques we currently employ for the domain of direction giving. A corpus analysis is underway to analyze the circumstances under which a particular representation technique is used. We expect, for example, visuo-spatial properties of the referent or the dialogue context to be correlated with the use of a certain representation technique.

Table 1. Categories of representation techniques

Indexing	Pointing to a position within the gesture space
Shaping	An object’s shape is contoured in the air
Drawing	Hands trace the shape of an object
Posturing	The hands form a static configuration standing for the object itself
Grasping	The hands grasp an object (often in combination with a placing movement)
Sizing	An object’s size is displayed, whereas the gesture typically refers to one axis’ extent of the object
Counting	The number of outstretched fingers is used to count sth.
Pantomime	Hands are used to imitate an action

Representation techniques are represented as templates that consist of two parts, one form part and one semantic, imagistic part. The semantic part is constituted by an underspecified IMD that represents the shape properties this technique is suitable to encode. In other words, it encodes the technique’s ”affordance” to depict shape-related aspects of 3-dimensional objects. The form part specifies the morphology of a gesture as a typed feature structure, which accommodates slots for handshape, hand position, movement, palm orientation, and extended finger direction (see Figure 4b). These slots conform the features that were successfully used in previous work to specify a gesture componentially with a behavior markup language (MURML [28]).



LOC: *Periphery Left*
TRAJ.: *Linear,Short*
MOVEMENT DIR: *Downward*
FINGER DIR: --
SHAPE: *ASL-bent-5*
PALM DIR: *Downward*

(a) LTAG tree generated by the Speech Formulator for the utterance “one church is on the left”.

(b) Gesture form feature matrix generated by the Gesture Formulator.

Figure 4. Generation results of speech and gesture formulation.

Some of the slots in a technique’s form feature structure are refined to specific values or a restricted value set. For example, the technique of placing requires a handshape that resembles a grasping, either a precision grasp with the index finger and the thumb or a power grasp with the whole hand, a short linear movement that is directed towards the target location, and a palm orientation that is perpendicular to this movement vector. Shaping, in contrast, is less restricted by conventions than by the IMD to be depicted. Its handshape can be either a one-hand or two-hand grasp or touch, its movement must trace the dominant (”max”) axes of the IMD either with one hand or symmetrically with two hands, and its palm orientation and extend finger direction are orientated parallel to the remaining axis or axes. All of these possible values are defined as restrictions over the value sets for the single slots of the feature structures.

The Gesture Formulator determines possible gestural renditions by selecting a technique and then turning it into a fully specified gesture. This starts with creating a disposition as to which technique to employ in order to convey the given object schema as part of the overall communicative goal. This disposition is realized by differently activating the possible techniques. Each representation technique is tested for how well it can carry the imagistic content at hand. To this end, two kinds of mappings are applied. First, mappings from single object schema components (i.e. axes) onto morphological gesture features that were drawn from a previous study on the detailed used of iconic gestures for describing simple geometrical objects [44]. Second, the numerical values defined in the object schema (relative to the taken perspective) are mapped onto gesture space coordinates. The results of both mappings are used to fill the gesture features structure, i.e. to adapt the representation technique to the particular context of usage.

3.2.5 Surface Realization

The four aforementioned modules interact until an adequate multimodal encoding of the intended message is achieved. The temporal coordination of verbal and gesture parts is achieved via a direct connection between Formulator and Gesture Generator. At last, Motor Control and Phonation are concerned with the concrete realization of speech and hand/arm movements for our virtual human Max employing the Articulated Communicator Engine (ACE, for short) for behavior realization [26]. ACE allows to create virtual animated agents, and to synthesize for them multimodal utterances including speech, gestures, or facial expressions from XML descriptions in MURML. Assuming that uttering runs incrementally in chunks (see [7] for empirical evidence), the ACE production model simulates the mutual adaptations that take place between running, continuous speech and gesture. Within each chunk, animations are formed such that the gesture spans the co-expressive word or sub-phrase; between successive chunks, the movement between two strokes as well as the silent pause between two intonation phrases are adapted to prepare the synchrony of the upcoming behaviors.

4 CONCLUSION

Based on what is currently known about the processes underlying the production of multimodal behavior, we have presented a computational model of speech and gesture generation. Our content representation consists of two different kinds of representation which are interconnected by multimodal concepts to constitute a level of multimodal meaning informing the derivation of coordinated speech and gestures.

How is this content representation able to account for the observed mutual influence of speech and gesture reported in Section 2? First, accounting for the influence of gesture onto speech, the ability to gesture results in the activation of the imagistic representation, and the resulting activation of a set of multimodal concepts. The propositions activated this way are then considered for the process of formulation. Note, that this way of modelling the influence of gesturing on language use also accounts for the fact that gesture frequency is significantly increased when expressing spatial information (for a review see [1]). When conveying information for which there is no imagistic representation, there is only activation within the propositional part of working memory.

Second, considering the influence of language on gesture, the Formulator provides feedback to the content planning level if and how a set of propositions is verbalizable. According to the set of activated multimodal concepts, the Action Generator is able to plan one or multiple gesture(s) in dependence on the linguistic part of the utterance. If, for example, two of the activated multimodal concepts are planned to be realized within speech, two gestures will be planned accordingly.

We are currently in the process of implementing out the model in full detail and further intricacies of the single processing stages are likely to arise. These technological challenges notwithstanding, we are confident from our extensive previous work as well as promising first results that the architecture and formalisms described allow us to simulate many of the phenomena of speech-gesture alignment that previous computational models have not been able to realize.

ACKNOWLEDGEMENTS

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 “Alignment in Communication”.

REFERENCES

[1] M.W. Alibali, ‘Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information’, *Spatial Cognition and Computation*, **5**, 307–331, (2005).

[2] M.W. Alibali and S. Kita, 'The role of gesture in speaking and thinking: Insights from piagetian conservation tasks', in *Communicating skills of intention*, ed., T. Sakamoto, Hitsuji Shobo, (in press).

[3] M.W. Alibali, S. Kita, and A.J. Young, 'Gesture and the process of speech production: We think, therefore we gesture', *Language and cognitive processes*, **15**, 593–613, (2000).

[4] G.L. Allen, 'Gestures accompanying verbal route directions: Do they point to a new avenue for examining spatial representations?', *Spatial Cognition and Computation*, **3**(4), 259–268, (2003).

[5] J. Bavelas, C. Kenwood, T. Johnson, and B. Philips, 'An experimental study of when and how speakers use gestures to communicate', *Gesture*, **2:1**, 1–17, (2002).

[6] G. Beattie and J. Coughlan, 'An experimenatl investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon', *British Journal of Psychlogy*, **90**, 35–56, (1999).

[7] K. Bergmann and S. Kopp, 'Verbal or visual: How information is distributed across speech and gesture in spatial dialog', in *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, eds., David Schlangen and Raquel Fernandez, pp. 90–97, (2006).

[8] I. Biederman, 'Recognition-by-components: A theory of human image understanding', *Psychological Review*, **94**(2), 115–147, (1987).

[9] J. Cassell, M. Stone, and H. Yan, 'Coordination and context-dependence in the generation of embodied conversation', in *Proceedings of the First International Conference on Natural Language Generation*, (2000).

[10] J.P. de Ruiter, 'The production of gesture and speech', in *Language and gesture*, ed., D. McNeill, Cambridge University Press, (2000).

[11] J.P. de Ruiter, 'Postcards from the mind - the relation between speech, imagistic gesture, and thought', *Gesture*, **7:1**, 21–38, (2007).

[12] J.P. de Ruiter, 'Some multimodal signals in humans', in *Proceedings of the Workshop on Multimodal Output Generation 2007*, eds., I. van der Sluis, M. Theune, E. Reiter, and E. Krahmer, (2007).

[13] M. Gullberg, 'Language-specific encoding of placement events in gestures', in *Event Representation in Language and Cognition*, eds., E. Pederson, R. Tomlin, and J. Bohnemeyer, (submitted).

[14] B. Hartmann, M. Mancini, and C. Pelachaud, 'Implementing expressive gesture synthesis for embodied conversational agents', in *Gesture in Human-Computer Interaction and Simulation*, eds., S. Gibet, N. Courty, and J.-F. Kamp, (2005).

[15] A. Herskovits, *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*, Cambridge University Press, 1986.

[16] A.B. Hostetter, M.W. Alibali, and S. Kita, 'Does sitting on your hands make you bite your tongue? The effects of gesture inhibition on speech during motor descriptions', in *Proceedings of the 29th annual meeting of the Cognitive Science Society*, eds., D. S. McNamara and J. G. Trafton, pp. 1097–1102. Erlbaum, (2007).

[17] A.B. Hostetter, M.W. Alibali, and S. Kita, 'I see it in my hands' eye: Representational gestures reflect conceptual demands', *Language and Cognitive Processes*, **22**, 313–336, (2007).

[18] A. Kendon, *Gesture - Visible Action as Utterance*, Cambridge University Press, 2004.

[19] D. Kieras, 'Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory', *Psychological Bulletin*, **85**, 532–554, (1978).

[20] M. Kipp, M. Neff, K.H. Kipp, and I. Albrecht, 'Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis', in *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, eds., C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, (2007).

[21] S. Kita, 'How representational gestures helps speaking', in *Language and gesture*, ed., D. McNeill, 162–185, Cambridge University Press, (2000).

[22] S. Kita and A. Özyürek, 'What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking', *Journal of Memory and Language*, **48**, 16–32, (2003).

[23] S. Kita, A. Özyürek, S. Allen, A. Brown, R. Furman, and T. Ishizuka, 'Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production', *Language and Cognitive Processes*, (2007).

[24] S. Kopp and K. Bergmann, 'Towards an architecture for aligned speech and gesture production', in *Proceedings of 7th Conference on Intelligent Virtual Agents*, eds., C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, LNAI 4722, pp. 389–390. Springer, (2007).

[25] S. Kopp, P. Tepper, K. Ferriman, K. Striegnitz, and J. Cassell, 'Trading spaces: How humans and humanoids use speech and gesture to give directions', in *Conversational Informatics*, ed., T. Nishida, chapter 8, 133–160, John Wiley, (2007).

[26] S. Kopp and I. Wachsmuth, 'Synthesizing multimodal utterances for conversational agents', *Computer Animation and Virtual Worlds*, **15**(1), 39–52, (2004).

[27] Stefan Kopp, Paul Tepper, and Justine Cassell, 'Towards integrated microplanning of language and iconic gesture for multimodal output', in *Proceedings of the International Conference on Multimodal Interfaces*, pp. 97–104, New York, NY, USA, (2004). ACM Press.

[28] A. Kranstedt, S. Kopp, and I. Wachsmuth, 'Murml: A multimodal utterance representation markup language for conversational agents', in *AAMAS'02 Workshop Embodied conversational agents- let's specify and evaluate them!*, (2002).

[29] R.M. Krauss, Y. Chen, and R.F. Gottesman, 'Lexical gestures and lexical access: A process model', in *Language and gesture*, ed., D. McNeill, Cambridge University Press, (2000).

[30] B. Landau and R. Jackendoff, '"What" and "where" in spatial language and spatial cognition', *Behavioral and Brain Sciences*, **16**(2), 217–265, (1993).

[31] S. Levinson, 'Frames of reference and molyneux's question: Cross-linguistic evidence', in *Space and Language*, eds., P. Bloom, M.A. Peterson, L. Nadel, and M.F. Garrett, 109–169, MIT Press, (1996).

[32] D. Marr and H. Nishihara, 'Representation and recognition of the spatial organization of three-dimensional shapes', in *Proceedings of the Royal Society of London B, 200*, pp. 269–294, (1978).

[33] D. McNeill, *Hand and Mind - What Gestures Reveal about Thought*, University of Chicago Press, Chicago, 1992.

[34] D. McNeill and S. Duncan, 'Growth points in thinking-for-speaking', in *Language and gesture*, Cambridge University Press, (2000).

[35] A. Melinger and W.J.M. Levelt, 'Gesture and the communicative intention of the speaker', *Gesture*, **4**, 119–141, (2004).

[36] P. Morrel-Samuels and R. Krauss, 'Word familiarity predicts temporal asynchrony of hand gestures and speech', *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 615–622, (1992).

[37] E. Morsella and R.M. Krauss, 'The role of gestures in spatial working memory and speech', *The American Journal of Psychology*, **117**, 411–424, (2004).

[38] C. Müller, *Redebegleitende Gesten: Kulturgeschichte - Theorie - Sprachvergleich*, Berlin Verlag, 1998.

[39] A. Özyürek, 'Speech-gesture synchrony in typologically different languages and second language acquisition', in *Proceedings of the Boston University Conference on Language Development*, pp. 500–509. Cascadilla Press, (2002).

[40] A. Paivio, *Mental Representations - A Dual Coding Approach*, Oxford University Press, 1986.

[41] F.H. Rauscher, R.M. Krauss, and Y. Chen, 'Gesture, speech, and lexical access: The role of lexical movements in speech production', *Psychological Science*, **7**, 226–231, (1996).

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

[43] D. I. Slobin, 'From "thought and language" to "thinking for speaking"', in *Rethinking linguistic relativity*, eds., J.J. Gumperz and S.C. Levinson, 70–96, Cambridge University Press, (1996).

[44] T. Sowa, *Understanding Coverbal Iconic Gestures in Shape Descriptions*, Akademische Verlagsgesellschaft Aka, Berlin, 2006.

[45] T. Sowa and S. Kopp, 'A cognitive model for the representation and processing of shape-related gestures', in *Proceedings of the European Cognitive Science Conference*, eds., F. Schmalhofer, R. Young, and G. Katz, p. 441, New Jersey, (2003). Lawrence Erlbaum Assoc.

[46] T. Sowa and I. Wachsmuth, 'A model for the representation and processing of shape in coverbal iconic gestures', in *Proceedings of KogWis05*, pp. 183–188, Basel, (2005). Schwabe Verlag.

[47] Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer, 'Microplanning with Communicative Intentions: The SPUD System', *Computational Intelligence*, **19**(4), 311–381, (2003).

[48] J. Streeck, *Gesture: The Manufacture of Understanding*, To appear.

[49] B. Tversky and P.U. Lee, 'How space structures language', in *Spatial Cognition: An interdisciplinary approach to representation and processing of spatial knowledge*, eds., C. Freksa, C. Habel, and K. F. Wender, pp. 157–176, Berlin, (1998). Springer.

ECA gesture strategies for robust SLDSs

Beatriz López[1] (student) and **Álvaro Hernández**[1] (student) and **David Pardo**[1] (student) and **Raúl Santos**[1] (student) and **María del Carmen Rodríguez**[2]

Abstract. This paper explores the use of embodied conversational agents (ECAs) to improve interaction with spoken language dialogue systems (SLDSs). For this purpose we have identified typical interaction problems with SLDSs and associated with each of them a particular ECA gesture or behaviour. User tests were carried out dividing the test users into two groups, each facing a different interaction metaphor (one with an ECA in the interface, and the other implemented only with voice). Our results suggest user frustration is lower when an ECA is present in the interface, and the dialogue flows more smoothly, partly due to the fact that users are better able to tell when they are expected to speak and whether the system has heard and understood. The users' overall perceptions regarding the system were also affected, and interaction seems to be more enjoyable with an ECA than without it.

1 INTRODUCTION

In this paper we examine certain complementarities of spoken dialogue and visual communication in Human-Machine Interaction (HMI). More specifically, we wish to identify effects of incorporating an animated agent onto a spoken language dialogue system (SLDS). Such dialoguing animated agents are commonly referred to in the literature as embodied conversational agents (ECAs) [1].

Our primary concern is to find benefits that may be gained by adding to a SLDS a visual channel of communication featuring an ECA. We are, of course, especially interested in improving aspects of human interaction with SLDSs that are particularly problematic. One major problem area is robustness. Dialogues often run into trouble for various reasons. For instance, when speech recognition errors occur it is usually difficult to recover from them. Error spirals are common [2], and even when the dialogue strategies designed specifically for error recovery are successful, interaction tends to become awkward, inefficient and "unnatural." Turn management is also tricky, and users are often not sure when they are supposed to speak.

What does an ECA bring into the picture? Most generally, a human-like figure adds a social element to the interaction. It may convey supra-linguistic information by performing gestures, including some designed as visual cues specifically to smoothen the flow of the dialogue making it seem more "natural" (for instance, by marking turn transitions), and others characterising expectations, mental processes (e.g., how well the system is understanding the user) and emotions (e.g., using emotional and empathic strategies to control user frustration when errors occur) - [3], [4], [5], [6], and [7].

According to some critics, however, no real benefits of interaction with ECAs have ever been proved. ECAs, they add, can be misleading and create false expectations regarding the system's interactional and functional capabilities. Furthermore, they can be confusing, distracting, and even increase user anxiety and reduce the users' sense of control ([8], [9]).

As part of the research we are currently undertaking in the context of COMPANIONS -a European Union project [10]-, we have performed user tests on a dialogue system with and without an ECA in an attempt to isolate the effects of the ECA on the interaction. Our comparative analysis is focussed especially on finding gesture sequences that complement dialogue strategies designed to improve dialogue flow and robustness, thus resulting in improved overall interaction.

The application scenario we have designed is a domotic videotelephony service where users call "home" using mobile phones (simulated on a computer screen) to check the state of various home appliances. This task isn't important in itself in the scope of our experiment (here we are not especially interested in designing a real remote domotic control service); we use it solely to motivate dialogue that may go through the main stages identified in the literature for automatic dialogue generation [11].

This notwithstanding, remote domotic control applications are certainly interesting in their own right. Today new videotelephony applications are being developed for mobile terminals, gradually moving towards the use of directed spoken dialogue to access a variety of information services (like voicemail or videomail). Incorporating ECAs onto this new visual channel affords challenges of its own. For instance, screen space is more limited, so what ECA size, appearance and gestures are best and whether it is appropriate to have an ECA on screen in the first place are all relevant questions for research.

The rest of the article is organized as follows: Section 2 presents the dialogue strategies we have implemented to increase robustness and the ECA behavioural schemes we have associated with them. Section 3 describes the types of ECA parameters considered in our evaluation. In Section 4 we explain how the empirical test was set up, and we show its structure. Section 5 shows the main results of the experiment, with discussion. Section 6 brings together the main findings and anticipates the next steps of research.

2 DIALOGUE AND GESTURE STRATEGIES

Among the more critical dialogue situations for which it is worth examining the positive effects an ECA could have are the following:

- Turn management: Here the body language and expressiveness of agents could be exploited to help regulate the flow of the dialogue [12]. Usability experimental analysis on how the facial feedback provided by avatars can

[1] ETSIT Universidad Politecnica de Madrid, Spain; email: beatriz@gaps.ssr.upm.es

[2] Telefónica I+D, Spain; email: mcrg@tid.es

make turn-taking smoother in the COMIC multimodal dialogue system has been presented in [13].

- Error recovery: The process of recognition-error recovery typically leads to a certain degree of user frustration (see [14]). In fact, once an error occurs it is common to enter an error spiral, because as the user becomes increasingly frustrated, her frustration leads to more recognition errors, making the situation worse [15]. ECAs may help to limit such feelings of frustration and by so doing make error recovery more effective [16].
- User confusion: A common problem in dialogue systems is that the user isn't sure what the system is doing and whether or not the dialogue process is working normally [17]. This sometimes leads the dialogue to error states that could be avoided. The expressive capacity of ECAs could be used to help the user keep track of what stage the dialogue is in (i.e., what the system is doing and expecting from the user).

We have designed a dialogue strategy to deal with various critical dialogue stages, react to different recognition confidence levels and manage error situations. Associated with the dialogue strategy is an ECA gesture scheme, with a set of gestures corresponding to each dialogue stage. Table 1 shows each dialogue stage, what prompts it, and the associated ECA behaviour. The gesture repertoire of our ECA is partially based on relevant gestures described in [1] and [12], and on recommendations in [18], [19], [20], [21], and [22], to which we have added a few suggestions of our own.

Aiming to define ECA behaviour during the interaction, we have tried to exploit the following supra-linguistic resources: conversational skills (such as beat gestures to emphasize information, nodding and "don't understand" gestures, waiting posture, etc.), shifts in camera shots and lighting intensity (in order to create "proxemic" effects that might be meaningful to the user), and the recreation of an empathic attitude in the ECA (smiling or offering an expression of apology) to try to keep user frustration low when interaction problems occur.

In the rest of this section we explain in a little more detail the dialogue-gesture scheme for each stage summarised in Table 1.

Initiation. Upon first encountering an ECA the user may "humanise" the system [23] and expect from it a lot more than it is actually capable of. Users may tend to speak with less restraint, making it more difficult for the system to understand them. The end result is likely to be somewhat disappointing and frustrating. Another possible effect we should consider is that contact with a dialoguing animated character may have the effect that the user's level of attention to the actual information that is being given is reduced ([24], [25]), especially in the case of new users (as our test users are). Thus, the goal at initiation is to present a human-like interface that is upon first contact less striking and less distracting, and one that clearly "lays down rules" of the interaction and sets the user on a track that is tightly focussed on the task at hand.

In order to try to foster a sense of ease in the user and help her focus we have designed a welcome gesture for our ECA based on the recommendations in Kendon [20], (see Table 1).

Termination. It is confusing if a dialogue concludes without the user being aware of it. It is important to end with a clear farewell message. We have complemented this with typical farewell gestures in human-human interaction [1].

MAIN DIALOGUE		
Dialogue stage	**Description (when it occurs)**	**ECA behaviour (movements, gestures and other cues)**
Initiation	At the beginning of the dialogue	Look straight at the camera, smile, wave hand. Zoom in for task explanation. Zoom out, lights dim.
Turn management	*Take Turn:* when the system starts to speak	Look straight at the camera, raise hand into gesture space. Camera zooms in. Light gets brighter.
	Give Turn: when the system prepares to listen to the user	Look straight at the camera, raise eyebrows. Camera zooms out. Lights dim.
Wait	When a timeout occurs	Slight leaning back, one arm crossed and the other touching the cheek. Shift of body weight.
Help	When the system gives some explanation to the user	Beat gesture with the hands. Change of posture.
Confirmation (low confidence)	When the system cannot understand something the user has said.	Slight leaning of the head to one side, stop smiling, mildly squint.
Confirmation (high confidence)	The system has recognised the user utterance with a high level of certainty	Nod gesture, smile, eyes fully open.
Acknowledgement of misunderstanding	After user informs the system that it has misunderstood what he or she has said. Speech: a) apology; b) repetition or rephrase request	Apology: Head aside, raise inner eyebrow central, head down, eyebrow of sadness (to show remorse). Request: Show expression of interest by opening eyes, and smiling slightly.
Error recovery with correction	When the user has corrected a recognition error and the system confirms the correction	Lean towards the camera, beat gesture.
Termination	Goal: to show that the dialogue is being closed. Speech: farewell message.	Looks straight at the camera, nod, smile, wave hand.

Table 1. Gesture repertoire for the main dialogue stages

Turn management. Turn management involves two basic actions: taking turn and giving turn. Dialogue fluency improves and fewer errors occur if alternate system and user turns flow in orderly succession with the user knowing when it is her turn to speak. It is important to point out that we have not allowed barge-in (i.e., the user cannot interrupt the system because the system doesn't listen to the user –the speech recogniser is inactive– while the system is speaking). This makes for a less flexible dialogue scheme than may be generally desirable, but we hope it offers at least two advantages: firstly, in certain problem situations such as error spirals [26] it may well be most advisable never to allow the user to interrupt while the system is trying to reach a stable, mutually understood dialogue state. Since these are the cases we are most interested in, it makes sense to work with barge-in-free dialogue. Secondly, if users try to speak when they're not "supposed to" (our users are not told they cannot interrupt the system) this usually leads to no-inputs (when the system isn't aware that the user has said something), no-matches (the system is unable to understand the incomplete utterance it "hears"), and perhaps recognition errors. Turn management then becomes more critical, and the consequences of confusion regarding who's turn it is more obvious. Thus the role an ECA may play in clarifing turn possession and turn transitions should be more apparent.

Our ECA strategy is as follows: When it's the ECA's turn to speak the camera zooms-in slightly and the light becomes brighter; while the ECA is approaching it raises a hand into the gesture space to "announce" that it is going to speak (see Figure 1). When it's the user's turn the camera zooms out, the lights dim and then the ECA raises its eyebrows to invite the user to speak. The idea is that, hopefully, the user will associate different gestures, camera shots and levels of light intensity with each of the turn modes.

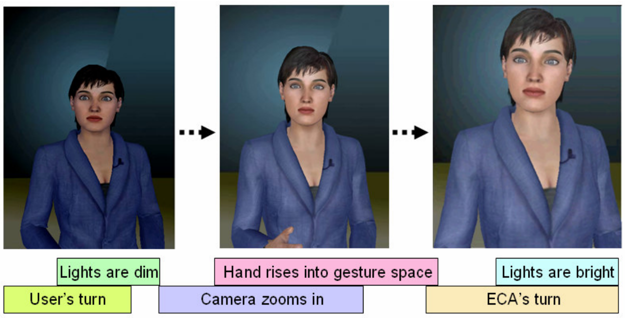


Figure 1. Visual sequence of turn transition from user to ECA.

Confirmation. Once the user utterance has been recognised, information confirmation strategies are commonly used in dialogue systems. Different strategies are taken depending on the level of confidence in the correctness of the user locution as captured by the speech recognition unit [22]. Our dialogue scheme and the associated gestural strategies are as follows:

- *High confidence in recognition:* The dialogue continues without confirmation request. The ECA nods her head [1], smiles and opens her eyes wide to show the user that everything is going well and the system understands her.
- *Intermediate confidence:* The result is regarded as uncertain and the system tries implicit confirmation (by including the uncertain piece of information in a question about something else). This allows the user to correct the system if an error did occur, and to feel everything is going well if what the system understood was correct. No specific ECA gesture was designed for this case. The idea is to keep the user speaking normally and without hyperarticulating (which would make recognition more difficult [15]).
- *Low confidence:* The dialogue becomes more guided with the system asking the user to repeat or rephrase. The ECA leans her head slightly to one side, stops smiling and mildly squints (a "what was that you said?" gesture; see Figure 2).

Acknowledgement of misunderstanding. A particularly delicate situation arises when the system misunderstands the user. If the user tries to correct the system or point out that it has misunderstood, the system will hopefully realise what has happened. It then tries to keep the user in a positive attitude and avoid her distrust while seeking to obtain the correct information. The dialogue scheme to pursue this consists in an apology followed by a kind request for a repetition or rephrase. The ECA gestures accordingly (see Table 1), stressing the system's "interest" in getting it right to further motivate the user and preserve her trust.

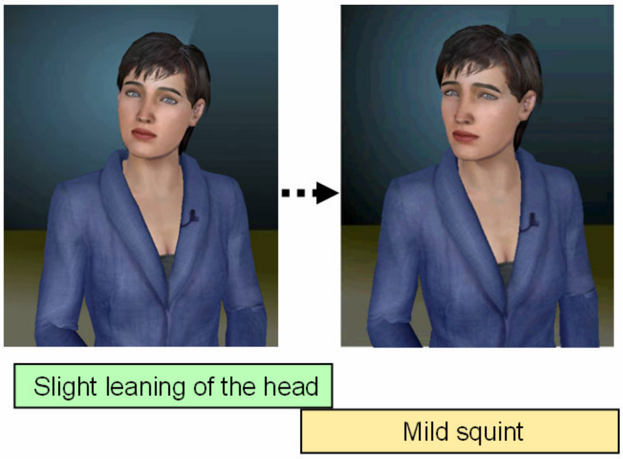


Figure 2. ECA gesture sequence expressing low confidence in the comprehension of the user's utterance.

Error recovery with correction. If the user says that recognition errors have taken place and gives the correct information at the same time, the ECA repeats the corrected information by leaning towards the camera and marking the words by means of beat gestures with both hands.

Help. A help message is given either when the user requests it or when the system has failed to hear the user say anything for longer than a reasonable waiting period. The ECA emphasizes the more important information in the help message with beat gestures performed with the hands. The idea is to see whether this captures the interest of the user, makes her more confident and the experience more pleasant, or if, on the contrary, it is distracting and makes help delivery less effective.

Wait. As we discussed before, it sometimes happens that the user doesn't realize it is her turn to speak. To help the user realise the system is waiting for her to say something the ECA performs a waiting gesture: leaning back slightly with her arms crossed and shifting the body weight from one leg to another.

3 EVALUATION PARAMETERS

As was mentioned in the introduction, our main goal is to evaluate how well ECAs can work in improving HMI performance parameters and user satisfaction. The approach we have taken is based on Möller et al.'s taxonomy of quality factors for dialogue systems [27] and the ITU P.851 recommendation [28] on evaluating dialogue systems, to which we have added questions as we have seen appropriate to evaluate user perceptions related to the ECA. We combine the system and interaction, performance and event data registered automatically, with user's responses to questionnaires (note that although recognition performance data is interesting, the goal of the experiment reported in this paper is not to evaluate how well the speech recogniser works).

In order to measure the influence of an ECA on user satisfaction we have compared a dialogue system that includes an ECA in the interface with one without ECA along a range of user-centred parameters. These parameters fall into three classes:

- *Typical dialogue system parameters* (automatically collected in the questionnaires) covering aspects of system performance, dialogue flow, information offered by the system, usefulness and overall evaluation, overall impression and perception of task success.
- *Impressions felt while using the system*: User emotions ("relaxed", "confident", "happy", "bored", "dejected", "angry" and "clumsy") and sensations ("pleasant", "fun", "interesting", "frustrating", "confusing" and whether they were surprised by anything). Participants were also invited to write comments about different aspects of the system.
- *Specific questions concerning the ECA* regarding both gesture design (clarity, naturalness, range of the gesture repertoire) and the perceived personality of the ECA ("expressive", "likable", "polite", and how comfortable it is to speak with the agent).

We have also added a time dimension to see whether we can determine how users' expectations evolve through use of the system (other studies, such as that in [29], have focused primarily on user expectations). We do this by repeating certain questions at different stages of the test.

4 EXPERIMENTAL SETUP

4.1 System implementation

The architecture of the test environment is based on web technology, with which we simulate a mobile phone interface. Figure 3 shows the two different interaction scenarios we have compared: one (on the right) corresponding to what we have called the ECA metaphor scenario, and the other (on the left) with a still image (representing "home") that we call VOICE metaphor scenario (SLDS without ECA). Different users interact with these two different scenarios providing contrastive experimental data that will allow us to evaluate the ECA metaphor vs. the VOICE metaphor. The system is implemented on a web page that contains two frames. In the left frame there is a column of labels that show the test user what stage of testing he or she is (not to be confused with the dialogue stage which is not indicated). The main interface is displayed in the right frame and shows a mobile phone running a videotelephony application. Tactile interaction is not active at any stage.

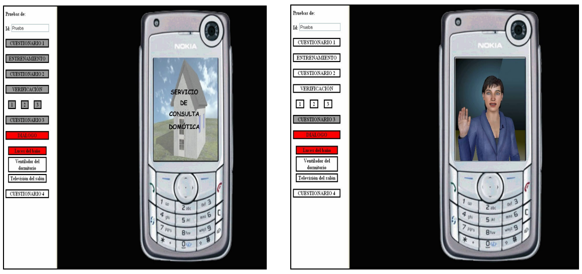


Figure 3. Interface displays for VOICE (left) and ECA (right) metaphors.

All the contents of the evaluation are hosted on an Apache Tomcat web server. Throughout the test, users face a series of evaluation questionnaires and experimental dialogue interactions. The questionnaires are implemented using HTML forms, and the information collected on them is transferred to JSP files and then stored in a database. Our test environment uses Nuance Communications' speech recognition technology [30]. The ECA character was created by Haptek [31]. The dialogues are implemented with Java Applet technology, and they are all packed and signed to guarantee fast download and access to the audio resources. Dialogue dynamics are programmed. Nuance's speech recognition engine provides a useful Java API that allows access to different grammars and adjusting a range of parameters depending on the characteristics of each application. Finally, interaction parameters (such as utterance durations, number of turns, number of recognition errors, etc.) are recorded automatically during the test interactions.

4.2 Description of the experiment

Testing was done in a small meeting room. Each user was sat at the head of a long table in front of a 15" screen. Two different views of the user interacting with the system where video-recorded to provide us with visual data to inspect and annotate the subject's behaviour. A frontal view was taken from the top edge of the user's screen, and a lateral view was recorded from a wide-angle position to the right of the user. Both views were taken with Logitech Quickcam Pro 4000 webcams. The users interacted with the system using a headset microphone, and the system prompts are played through two small speakers. Half of the users interact with a system only through spoken dialogue; the other half encountered an interface that includes an ECA. All user-system dialogue was in Spanish.

The evaluation was designed so that users could carry out the test with minimal intervention on the part of experimenter. The stages of the test were as follows:

1) Brief explanation: An experimenter briefly explains to each test participant what the general purpose (to "evaluate automatic dialogue systems") and methodology of the test are, as well as the tasks that lie ahead. We try to emphasize the importance of the answers given in the questionnaires.

2) Opening questionnaire to learn about the users' prior experience and expectations.

3) Training and verification phase (and associated questionnaires): Users are asked to enrol in a secure access application using voice recognition, and then to verify their identity. The interaction method is a rigid, directed dialogue, with an ECA for half of the users. We will not deal with this aspect of our test in this paper. We mention it here for two reasons: the first is for completeness and accuracy of our account of the testing procedure; the second is that some questions contained in the questionnaires at this stage are repeated later in the final questionnaire, which enables us to analyse how users' opinions and expectations evolve throughout the test.

4) Dialogue phase: Users are given three dialogue tasks. In each task users are asked to find out the state (on/off) of a household device ("the bathroom lights", "the fan in the bedroom", and "the living-room television set". The automatic speech recogniser and the dialogue system function freely (i.e., they are not programmed to give certain answers; it is a real working system). Half of the users interact with an ECA and the other half without.

5) Final questionnaire to get the user's overall impression of the system, its main elements and the more important aspects of using it. As mentioned before, some questions are the same as in previous questionnaires to provide information regarding the evolution of users' perceptions throughout the various stages of system use.

5 EXPERIMENTAL RESULTS

We carried out testing with 16 undergraduate and graduate students (7 female and 9 male), of ages ranging from 19 to 33, divided into two groups (8 users in each group), one to test the system with the ECA interaction metaphor and the other with the VOICE metaphor.

Our analysis is mainly based on the users' answers to the questionnaires and on the performance parameters registered during the course of the user-system interaction. As previously mentioned, we based our questionnaire design partly on the ITU-T P.851 recommendation [28], which identifies a variety of conceptual dimensions or categories that should be taken into account when writing questions to evaluate users' opinions on a comprehensive range of aspects related to quality of interaction with dialogue systems. We have added similar question categories that deal with ECA presence and gestures, and also a set of questions to inquire about the users' emotions while using the system.

We have reached the following results by a) comparing the performance and the answers to the questionnaires of the ECA metaphor group of users with those of the VOICE metaphor group; and b) analysing how performance and responses to certain questions evolve throughout the test. In addition, we have looked at users' comments, given at certain points in the questionnaires, and compared them to the findings in a) and b).

We carried out a series of two sample t-tests, setting the significance level at 5% (p=0.05). Questionnaire responses were collected on Likert-type 5-point response formats.

In the rest of this section we present the main findings obtained from these comparative analyses.

5.1 Sensitivity to errors and user frustration

We found some statistically significant differences between ECA and VOICE metaphors regarding certain factors related with robustness in difficult dialog situations (e.g., when the system acknowledges having misunderstood something, or when the system doesn't "get" what the user has said). Specifically:

Average user awareness of system recognition errors is lower for ECA users. In spite of the fact that the minor difference we found in the actual average numbers of recognition errors between both of the tested interaction metaphors was not statistically significant, there was a striking, statistically significant, difference in the answers to the question *"Did the system make many mistakes?"* (1- very many ... 5 - none): a mean value of 3.8 for the ECA metaphor vs. 2.6 for the VOICE metaphor (t(12)=3.16; p-value=0.004). User frustration while interacting with the system was also markedly lower for the ECA group, as indicated by the 1.4 (ECA) vs. 2.6 (VOICE) mean values (t(9)=-2.52; p-value=0.016) of the responses to the question: *"Was the experience [of using the system] frustrating?"* (1 – no, not at all ... 5 – yes, very much so).

The measured differences in the two previous parameters between the ECA and VOICE scenarios possibly reflect relevant advantages, at least in terms of how it affects user perception, of the use of ECAs with appropriately designed gestures, both to deal with problematic dialog stages such as error recovery situations and to provide users with visual cues of how well the system is understanding her (i.e., with what level of confidence; see Section 2). We could be seeing here a variant of the persona effect [32], a phenomenon widely reported in the literature according to which users tend to perceive a particular task as easier when they interact with an ECA in order to carry it out, without there being any real improvement in performance (success in task execution and efficiency) when compared with users doing the same without an ECA. In our case no significant difference was found between the two test groups regarding perception of ease of use. However, believing the system made fewer mistakes could be a related effect.

There may be more to it, though, and user frustration and perception of performance quality may be linked to actual improvements in dialogue flow and in the users' knowledge of what is going on (what the system is doing and expecting the user to do). We now turn to exploring these possibilities briefly.

5.2 Dialogue coordination and fluency with visual cues for turn-switching

Efficiency and fluency of interaction are important factors (identified in [28]) in which we have also found differences between the ECA and VOICE metaphors. Users' perception that *"Dialoguing with the system led quickly to solve the task proposed"* (1 - totally disagree ... 5 - totally agree) was on average greater in the ECA group (4.2) than in the VOICE group (3.2) (t(12)=3,16; p-value=0.004).

This is not simply, or not solely, a subjective impression induced by the presence of the ECA, which would make it an instance of the persona effect. In fact, a close examination of our experimental ECA-supported dialogues shows that users easily learn when they are supposed to speak to the system (i.e., when it is their turn). This helps prevent most of the typically observed failed barge-in attempts and time-outs, which we found occurred more often for our VOICE metaphor users. Some of these users said they had felt confused at certain stages of the dialogue (e.g., *"between tasks there were silences and I didn't know if I was supposed to say anything," "a couple of times I think I spoke too early and that's why the system didn't get what I said," "it would be better if some sort of visual sign told you when the system is ready to listen"*).

However, we found no statistically significant differences between the two groups of users as regards task duration and number of turns taken, which are, of course, two important efficiency indicators. This notwithstanding, all of the main performance indicators were slightly better for the ECA group than for the VOICE group: average dialogue duration (μ_{ECA}=38655ms (std=18688); μ_{Voice}=47657ms (std=34043)), total duration of user turns (μ_{ECA}=4267ms (std=1745); μ_{Voice}=6182ms (std=4615)), number of dialogue turns (μ_{ECA}=5.70 (std=2.17); μ_{Voice}=6.33 (std=4.43)), number of time-outs turns (μ_{ECA}=0.08 (std=0.40); μ_{Voice}=0.20 (std=0.72)), number of times a help message is given (when the system “realises” the user may be in trouble or confused) (μ_{ECA}=0.04 (std=0.20); μ_{Voice}=0.12 (std=0.44)) and number of no-matches (μ_{ECA}=0.25 (std=0.53); μ_{Voice}=0.41 (std=0.50)) were all lower for the ECA group.

Our sample sizes are rather small so we need to increase them to see if these observed differences become statistically significant. But for the time being it is reasonable to interpret our findings as possible evidence of a combination of a persona effect with the fact that ECA-metaphor users learn how to interact with the system more easily and feel more in control, and actually achieve a more coordinated dialogue (if not significantly more efficient in terms of time) than VOICE-metaphor users.

Thus, it seems our visual feedback channel featuring an ECA displaying contextual dialogue management cues may be providing supra-linguistic information that users are able to interpret correctly, which translates into an improved coordination, which in turn increases the users’ impression of the dialogue being fast, efficient and under control. This could also be related with our finding that user perception of system mistakes and user frustration were lower for the ECA group, as reported above.

But what are these visual cues that appear to be so useful? Our findings allow us to suggest that the visual information strategy for turn-switching that we have implemented –involving a combination of gestures and lighting and camera zoom effects– may be creating a “proxemic-code” that helps avoid the complicated, problem-laden interaction patterns reported in [13], where user-ECA interaction suffers from rather severe coordination problems. Furthermore, our proxemic strategy is as simple as the good old invitation to speak using beeps, with the advantage that in our tests we haven’t observed any sign of rejection as may arise with the use of artificial-sounding beeps. Users seem to accept meaningful proxemic shifts as a “natural” part of dialogue interaction.

5.3 User expectations and perception of dialogue capability

The users’ impression of how powerful the system’s dialogue capabilities are, combined with the users’ expectations regarding these capabilities, has an important impact on the users’ overall assessment of the system [28]. Our experimental results show that the ECA-metaphor group was impressed with the system dialogue capability, although somewhat less than the VOICE-metaphor group, the former grading with an average of 3.9, and the latter 4.5 (3.0 being the neutral score), on the question: *“Were you positively or negatively surprised by the system’s dialogue capability?”* (1 - very negatively surprised … 5 - very positively surprised) (t(13)=-2.12; p-value=0.027). This result is in agreement with the findings in other research efforts (see, e.g., [10]).

A plausible explanation has to do with the effect, discussed in Section 2, by which users that encounter an “embodied” interface tend to be overoptimistic with regard to the system’s capabilities, assuming these to be more on a par with those of human beings. But, since in fact we have the same dialogue engine behind both our ECA and VOICE-metaphor interfaces, users of the former tend to end up being less impressed with the system’s conversational skills –having expected more but getting the same– than users of the latter.

This, of course, notwithstanding the fact that the users in the ECA group don’t really “get the same,” if we consider that, on average, they experience a smoother dialogue, as we saw previously. The following qualitative impressions expressed by our test users may add a little perspective to the analysis:

“In the beginning my main feeling was one of mistrust because it was a new experience, but afterwards it was pleasant and it was very easy to become accustomed to it.”

“I thought that the interaction with the system would be less comfortable, but the system understood me very well.”

Here we see that initial expectations might not be so positive after all, and that the experience of interacting with the system did in fact exceed at least some of the users’ expectations. We clearly need to carry out further tests to shed light on the intricacies of user expectations and their evolution through system use.

5.4 Emotions

Apart from frustration, the only other feeling for which our data shows a statistically significant difference between the ECA and the VOICE group is happiness (users in both groups felt similarly relaxed, confident, bored, dejected, angry and clumsy, for instance). The ECA group averaged 4.0, against 3.1 for the VOICE group, in their replies to the question: *“While you were interacting with the system, did you feel happy?”* (1 - no, not at all … 5 - yes, very much so) (t(13)=1.99; p-value=0.034).

It is clear that the observed difference in emotional response between the two test groups, favouring as it may the use of an ECA, was only very slight. After all, the whole experimental procedure is short and fairly simple, and test users have very little at stake performing the test, so it seems unlikely that strong emotional responses might appear. However, in future experiments we plan to design longer, more complex tasks and, by increasing the sample size, we hope to be able to determine more precisely how our ECA affects user emotions, if at all, and how these might affect overall usability and user acceptance.

5.5 ECA expressiveness

We invited the test users to give us their views regarding the ECA’s gestures and expressiveness. These are a few revealing samples:

“I very much liked the expressiveness of the animations.”

“I found the agent and the agent’s gestures surprising.”

“The face gestures were very well designed, but the hand gestures could distract you.”

“I liked the ECAs very much. They’re very funny.”

These opinions are encouraging, especially as there are studies that point out that in order to improve the believability and naturalness of an ECA it is essential to give it a consistent personality and to make it expressive (see, e.g., [33]).

Furthermore, in our study we have observed that the users’ opinion of the ECA’s expressiveness increases with use after first contact (which occurs in the identity verification phase of the test): the average score for *“Is the agent expressive?”* (1 - no, absolutely not … 5 - yes, very much so) increased from 3.5 after first contact to 4.1 at the end of the test (t-value=-3.42; p-value=0.006). Similarly, users’ impression of ECA friendliness (another relevant factor connected to user expectations; see [34]) also increases slightly with use, from 4.1 to 4.5 (t-value=-2.05; p-value=0.040).

Expressiveness and friendliness may be "humanising" the ECA [35], but in a way that, rather than leading ultimately to disappointment, keeps users in a positive attitude and raises their interest in a natural-feeling interaction. This happens though the course of time (the little time our test lasts), which may be yet another piece of evidence that our ECA doesn't trigger unrealistic expectations upon first appearance, but gradually "wins users over."

Finally, we mention that in the present work we have not focused on specific gesture design (which gestures were preferred, which were perceived as being the clearest, and so on). However, prior to the present experiment we carried out a successful gesture validation test on the repertoire displayed by our ECA [36]. The comparative experiment discussed in this paper also serves as *implicit* overall gestural validation thanks to the interaction improvements we have observed. By analysing the video recordings of the user tests (which we will do shortly) we hope to obtain deeper insights on the effects of specific gestures –especially those we have designed with a view to improving dialogue robustness in difficult situations– and on how we might refine them.

6 CONCLUSIONS AND FUTURE WORK

Our line of research is intended to help make some progress in identifying the pros and cons of Embodied Conversational Agents (ECAs). In this article we have presented a research scheme in which we consider the main problem situations that typically arise in automatic dialogue generation. In order to improve the robustness and the ease-of-flow of the dialogue we have implemented a gesture repertoire to be displayed by an ECA at each stage of the dialogue. These gestures are designed to convey to the users meaningful supra-linguistic information regarding the state of the dialogue throughout the interaction, and to try to keep the user in a positive frame of mind. We have proposed evaluating how well these strategies work by setting up an experiment to compare two interaction scenarios or metaphors (ECA metaphor vs. VOICE metaphor).

We found that the ECA contributed to keeping user frustration low, especially when recognition errors occurred (which is the most delicate scenario). This result suggests that the error management strategies employed are working, particularly: a) implicit confirmation with no ECA reaction when confidence in recognition is intermediate; b) performing a "What was that you said?"-type gesture to show the user the system isn't sure it has understood but is making an effort to (when confidence in recognition is low); and c) acknowledging misunderstandings with an apology and an accompanying gesture sequence to reassure the user that the system knows what has happened and is trying to put things right.

Also worth mentioning is the observed improvement in dialogue fluency (especially in connection with turn changes) with the ECA interaction metaphor. The combined use of specific gestures and proxemic effects (playing with "camera" shot distance and light intensity) seems to be a promising alternative to the traditional 'beep' signal. In the absence of acoustic signals or visual cues, some users start speaking before the system is ready to listen. When visual cues are added, however, users display a greater tendency to wait until they see the animated figure is inviting them to speak. These strategies add naturalness and smoothness to the flow of the interaction.

On the negative side, the ECA's human-like appearance could potentially cause users to ultimately be somewhat disappointed with the system's dialogue ability, probably because of the false expectations such an appearance gives rise to, as has already been reported in the literature. Our results cannot confirm nor disprove this effect. However, we have seen indications that our ECA doesn't generate expectations in users that are too far off the mark. Indeed, users seem to appreciate the ECA more after they have interacted with it for a while. Nevertheless, this is an area we must examine more closely in future work.

The signs on which we have based our observations are only mild. We will continue testing with this experimental set-up, after which we will analyse all the gathered information, including the video recordings, to confirm (we hope) the effects reported in this paper and to refine our findings and discover more relationships between the interaction aspects we have considered (what we have presented here is a first batch of results that don't fully exploit the possibilities of the dialogue and gesture strategies we have developed).

One weakness in our study is the inadequacy of the experimental design for studying the evolution of system-user interaction and user impressions over long periods of time. It is most reasonable to assume that the ECA may have a noticeable novelty effect on inexperienced users, which affects our observations. Nevertheless, observations from another line of research we are undertaking on ECA interfaces for children with motor disabilities suggest that (at least in certain contexts) the influence of the novelty effect should not be overstated. We will report results in future work.

We are now annotating the videos of the interactions in such a way as to make it easier to accumulate information on a variety of test parameters and even to share it with other research groups. Finally, using these videos, we plan to design tests to study the reactions of users to the emotional behaviour of the ECA, as a first step to modelling different types of users (e.g., extroverted/introverted, patient/irritable, etc).

We hope our work may help to show ways in which ECA technology can make a positive contribution to natural dialogue interfaces.

ACKNOWLEDGEMENTS

This work was carried out with the support of the European Union IST FP6 program, project COMPANION, IST-34434 and by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

REFERENCES

[1] J. Cassell,, T. Bickmore, H. Vilhjálmsson, and H. Yan, More than just a pretty face: affordances of embodiment, *in Proceedings of the 5th international conference on Intelligent user interfaces*, pp. 52-59, ACM Press, 2000.

[2] S. J. Boyce, Spoken natural language dialogue systems: user interface issues for the future, *in Human Factors and Voice Interactive Systems,* D. Gardner-Bonneau Ed. Norwell, Massachusetts, Kluwer Academic Publishers: 37-62, 1999.

[3] D. McNeill, Hand and Mind: What Gestures Reveal about Thought. *The University of Chicago Press*, Chicago, 1992.

[4] P. Ekman, Facial Expression And Emotion, *American Psychologist*, 48(4), 384-392, 1993.

[5] M. Montepare, S.B. Goldstein, and A. Clausen, The iden-tification of emotions from gait information, *J. Nonverbal Behavior*, vol. 11, no. 1, pp. 33–42, Spring 1987.

[6] I. Poggi, C. Pelachaud and E.M. Caldognetto, Gestural Mind Markers in ECAs, *Gesture Workshop 2003*, pp 338-349, 2003.

[7] N. Leßmann, A. Kranstedt, and I. Wachsmuth, Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max, *Proceedings Workshop W12, AAMAS* 2004, New York, 57 - 65.

[8] R. Catrambone, Anthropomorphic agents as a user interface paradigm: Experimental findings and a framework for research, *in Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 166-171, Fairfax, VA, 2002.

[9] J. Xiao, Empirical Studies on Embodied Conversational Agents, *Ph.D. Dissertation*, Georgia Institute of Technology, Atlanta, GA, December 2006.

[10] COMPANIONS, European Commission Sixth Framework Programme Information Society Technologies Integrated Project IST-34434, http://www.companions-project.org/.

[11] M. McTear, Spoken Dialogue Technology: Towards the Conversational User Interface, *Springer*, 2004.

[12] T. Bickmore, J. Cassell, J. Van Kuppevelt, L. Dybkjaer, and N. Bernsen, (eds.), *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*, chapter Social Dialogue with Embodied Conversational Agents. Kluwer Academic, 2004.

[13] M. White, M. E. Foster, J. Oberlander, and A. Brown, Using facial feedback to enhance turn-taking in a multimodal dialogue system, *Proceedings of HCI International 2005*, Las Vegas, July 2005.

[14] S. Oviatt, and R. VanGent, Error resolution during multimodal humancomputer interaction, *Proc. International Conference on Spoken Language Processing*, 1 204-207, (1996).

[15] S. Oviatt, M. MacEachern, and G. Levow, Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication*, vol.24, 2, 1-23, (1998).

[16] K. Hone, Animated Agents to reduce user frustration, in *The 19th British HCI Group Annual Conference*, Edinburgh, UK, 2005.

[17] S. Oviatt, Interface techniques for minimizing disfluent input to spoken language systems, *in Proc. CHI'94*, pp. 205-210, Boston, ACM Press, 1994.

[18] J. Cassell, Y.I. Nakano, T.W. Bickmore, C.L. Sidner, and C. Rich, Non-verbal cues for discourse structure, *in Proceedings of the 39th Annual Meeting on Association For Computational Linguistics*, 2001

[19] N. Chovil, Discourse-Oriented Facial Displays in Conversation, *Research on Language and Social Interaction*, 25, 163-194, 1992.

[20] A. Kendon, Conducting interaction: patterns of behaviour in focused encounters, *Cambridge Univer-sity Press*, 1990.

[21] J. Cassell and K.R. Thorisson, The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, vol.13, pp.519-538, (1999).

[22] R. San-Segundo, J.M. Montero, J. Ferreiros, R. Córdoba, and J.M. Pardo, Designing Confirmation Mechanisms and Error Recover Techniques in a Railway Information System for Spanish, *SIGDIAL*, Denmark, 2001.

[23] S. Oviatt, and B. Adams, Designing and evaluating conversational interfaces with animated characters, *Embodied conversational agents*, MIT Press: 319-345, 2000.

[24] H. Schaumburg, Computers as tools or as social actors: the users' perspective on anthropomorphic agents, *International Journal of Cooperative Information Systems*, pp 217-234, 2001.

[25] R. Catrambone, J. Stasko, and J. Xiao, ECA as user interface paradigm, From brows to trust: evaluating embodied conversational agents, *Kluwer Academic Publishers*, Norwell, MA, 2004.

[26] I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg, Error correction detection and response generation in a spoken dialogue system, *Speech Communication* 45, pp 271-288, 2005.

[27] S. Möller, P. Smeele, H. Boland, and J. Krebber, Evaluating spoken dialogue systems according to de-facto standards: A case study. Computer Speech & Language 21 (2007) 26-53.

[28] ITU-T P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken. Dialogue Systems, *International Telecommunication Union (ITU)*, Geneva, 2003.

[29] K.Jokinen and T. Hurtig. User Expectations and Real Experience on a Multimodal Interactive System. In *INTERSPEECH-2006*, paper 1815-Tue2A3O.2.

[30] Nuance Communications' speech recognition technology, http://www.nuance.com.

[31] Haptek, http://www.haptek.com

[32] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone and R. S. Bhogal, *The persona effect: affective impact of animated pedagogical agents*, in Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press New York, NY, USA, pp. 359-366, 1997 .

[33] A.B. Loyall, and J. Bates, Personality-rich believable agents that use language. In Johnson, W.L., Hayes-Roth, B., eds.: Proceedings of the First International Conference on Autonomous Agents (Agents'97), Marina del Rey, CA, USA, ACM Press (1997) 106–113.

[34] N.C. Krämer, G. Bente, and J. Piesk, The ghost in the machine. The influence of Embodied Conversational Agents on user expectations and user behaviour in a TV/VCR application. IMC Workshop (2003) 121-128.

[35] B. Reeves and C. Nass. The media equation: How people treat computers, television and new media like real people and places. CSLI Publications, Stanford,CA, 1996.

[36] B. López, Á. Hernández, D. Díaz, R. Fernández, L. Hernández, and D. Torre, Design and validation of ECA gestures to improve dialogue system robustness, Workshop on Embodied Language Processing, in the 45th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 67-74, Prague, 2007.