EXPLORATIONS IN

MULTIMODAL INFORMATION PRESENTATION

Charlotte van Hooijdonk

Explorations in

Multimodal Information Presentation

Promotores:
Prof. Dr. A. Maes
Prof. Dr. E. Krahmer

Leden promotiecommissie:
Prof. Dr. J. Bateman
Dr. H. van Oostendorp
Prof. Dr. W. Spooren
Prof. Dr. M. Steehouder
Prof. Dr. M. Swerts
Dr. M. Theune

Contents

Acknowledgements (in Dutch)

Aan de totstandkoming van dit proefschrift hebben veel mensen een bijdrage geleverd die ik hier graag wil bedanken.

Allereerst, Fons Maes en Emiel Krahmer als mijn promotoren en Nicole Ummelen als mijn begeleidster. Met zijn vieren hebben we veel discussies gewijd aan de richting van dit proefschrift. Nicole was mijn eerste dagelijks begeleidster en heeft een belangrijke bijdrage geleverd aan het onderzoek dat beschreven is in Hoofdstuk 3. Na haar vertrek werd Emiel mijn dagelijks begeleider. Emiel is een vrolijke, enthousiasmerende en inspirerende onderzoeker. Ik wil hem in het bijzonder danken voor zijn inzet, geduld en optimisme. Fons is vriendelijke en inspirerende onderzoeker met wie ik, onder het genot van een appeltje, graag van gedachte wisselden over het lopende onderzoek. Ik wil hem in het bijzonder danken voor zijn steun en zijn geloof in mij.

Bij de sectie Communicatie & Cognitie vond ik een plek waar ik met veel plezier aan mijn proefschrift heb gewerkt. Ik dank dan ook mijn collega's voor hun steun en gezelligheid. In het bijzonder wil ik de volgende mensen danken:

- Carel van Wijk voor zijn statistische adviezen,
- Reinier Cozijn en Edwin Commandeur voor hun hulp bij het opzetten en uitvoeren van het oogbewegingsregistratie-experiment,
- Lennard van de Laar voor zijn technische ondersteuning tijdens de experimenten en zijn hulp bij het maken van de omslag van dit proefschrift.

Met Anja Arts en Pashiera Barkhuysen deelde ik samen een kamer. Ik wil Anja danken haar steun en goede raad. Met Pashiera, mijn paranimf, heb ik vier jaar lang lief en leed mogen delen. Bij Ingemarie Sam en Lauraine Sinay kon ik mijn verhaal kwijt tijdens een kop koffie of een lunchwandeling in het Warandebos.

Het onderzoek dat in dit proefschrift is beschreven, heb ik grotendeels uitgevoerd binnen het IMOGEN project. Ik wil daarom Wauter Bosma, Erwin Marsi en Mariët Theune danken voor de prettige samenwerking.

Tenslotte wil ik mijn achterban danken: mijn ouders, Jos en Marie-José, mijn zus Elise en mijn broer Olivier. Bedankt voor jullie steun en interesse en voor de lessen die jullie mij meegegeven hebben: "Volg je hart en benut je talenten ten volle." Tenslotte wil ik Martin danken. Hij is mijn steun en toeverlaat en ik ben erg blij dat

jij mijn paranimf wilde zijn. Bedankt dat je me hebt geholpen bij het verwezenlijken van mijn droom.

1 | General introduction
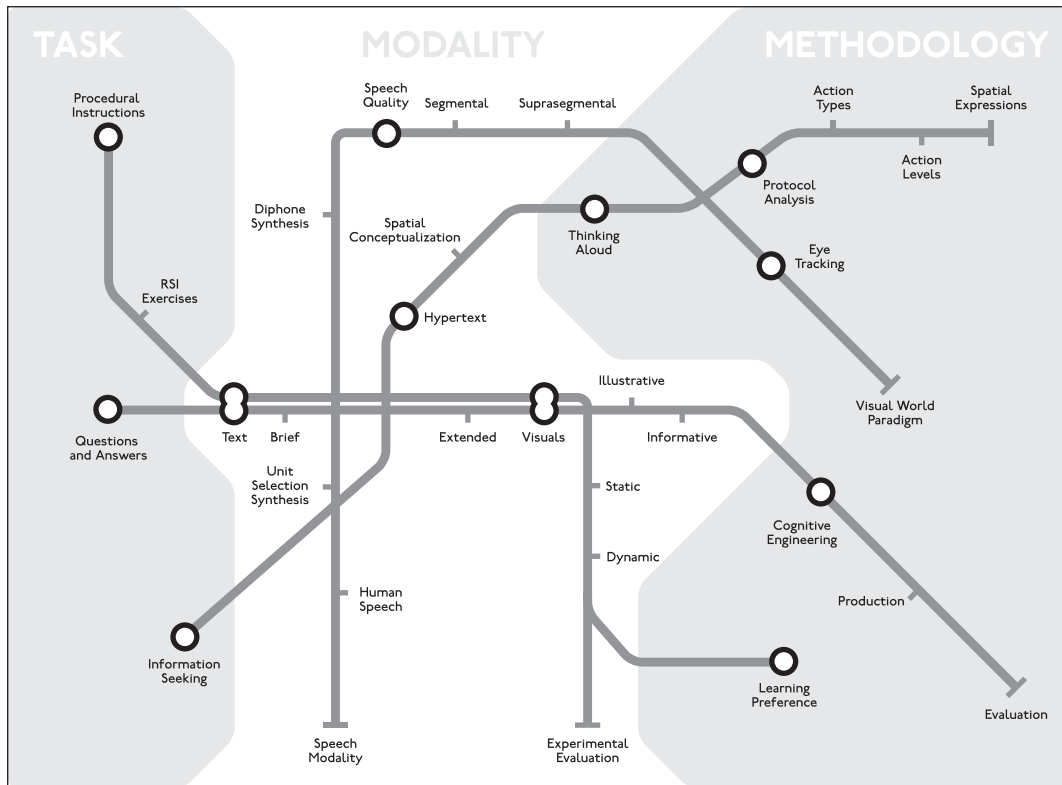
1.1 What is multimodal information presentation?

The cover of this thesis is inspired on the London Underground Map. This map has not only been a guide for travellers going from point A to point B, but it has also become a symbol for London itself (Roberts, 2005). The London Underground Map is a good example of a multimodal information presentation because it presents the information of London Underground by combing several presentation modes, i.e., text and visual representations of the tube lines. Moreover, the London Underground Map is an example of a good multimodal information presentation because the use of multimodal means matches the map's goal, i.e., guiding travellers in the right direction in a complex network of lines, stations, and zones.

A multimodal information presentation can be classified on the basis of three criteria or perspectives, i.e., the delivery medium, the presentation mode, and the sensory modality (Mayer, 2001). The first distinguishes presentations based on the devices used to deliver the information (e.g., paper, computer screen, loudspeaker). The second classifies presentations on the basis of the format of the message or the sign system used, like text or visuals. Finally, the third perspective starts from the human senses employed to process information, such as the auditory and visual senses. Note that these different views are highly related and often show different sides of the same coin (Maybury, 1993). For example, a particular medium may restrict the sensory modes involved (e.g., information on paper only serves the visual sensory mode), or a single medium may support several presentation modes (e.g., a piece of paper supports both text and visuals). Also, a single mode, like language, may be processed through different human senses (e.g., spoken text is processed aurally, while written text is processed visually). Although different distinctions can be made between 'mode' and 'modality', the ones formulated by Mayer (2001) enable us to define the modes and modalities discussed in this thesis. According to Mayer's tripartition, Chapters 2, 3 and 4 focus on different modes (e.g., text, graphics, and film clips) presented on a computer screen, while Chapter 5 focuses on the modality of the auditory sense.

The term 'multimodality' can also be defined from the perspective of written and spoken language research (Maes, 2005). In written language, multimodality refers

to the combination of verbal and nonverbal elements presenting information in documents. Examples of nonverbal elements are the visual vocabulary to organize text in lines, on a page, or in a document (see for an overview Kostelnick & Roberts, 1998), but also static (e.g., photos) and dynamic (e.g., animations) visuals. In spoken language, multimodality refers to the different modes with which spoken messages are communicated, such as intonation, speech quality, and facial expressions (Knapp, 1978). In this thesis, both research perspectives on multimodality are discussed: Chapters 2, 3, and 4 start from written language research, whereas Chapter 5 starts from spoken language research.

In this thesis, we speak of a multimodal information presentation if a chunk of information is presented through several presentation modes, like a combination of written or spoken text and visuals. There are reasons to believe that presenting information using multiple modalities is more effective than presenting information using a single modality (e.g., Mayer, 2001; Oviatt, 1999). Recent developments in computer technology have led to new possibilities of presenting information and to a renewed interest in the effects of different presentation modes. Naturally, this raises questions, like "Which presentation modes are most suitable in which situation?" and "How should different presentation modes be combined?" A research project which addresses these questions is the IMOGEN (Interactive Multimodal Output GENeration) project. This project is embedded in the IMIX (Interactive Multimodal Information eXtraction) research programme in the field of Dutch speech and language technology and is sponsored by the Netherlands Organisation for Scientific Research (NWO). Within the IMIX research programme a multimodal medical question answering (QA) system is being developed. A QA system is an automatic system that can answer a user's question posed in natural language (e.g., "What does RSI stand for?") with an answer formulated in natural language (e.g., "Repetitive Strain Injury"). Nowadays, QA systems are not only expected to give answers to these simple questions, but also to more complex questions, like "How should I organize my workspace in order to prevent RSI?" or "What is a good exercise to prevent RSI in my hands?" The answers to these questions might be more informative and effective if they contained multiple modalities, like text and a picture (Theune et al., 2007). In the IMOGEN project different aspects of multimodal information presentation are studied in order to improve the output quality of QA systems.

1.2 Research questions addressed in this thesis

Presenting information in a multimodal way is not trivial. It implicates a complicated mixture of characteristics of communicative tasks and goals, user characteristics and preferences, characteristics of sensory modalities, and qualities of presentation modes. One of the first questions that arises when presenting information in a multimodal way is which presentation mode(s) should be used. For example, suppose someone wants information on how to organize his / her workspace to prevent Repetitive Strain Injury. How should this information be presented to the user? A possibility would be to present the information through text (see Figure 1.1). However, the presentation would probably be more informative if it contained a visual as it would clarify the relations between the objects (e.g., chair, desk, and computer screen) within an ergonomic workspace in one glance (see Figure 1.1). Another possibility would be a multimodal information presentation in which a text and a visual are combined (see Figure 1.1). Note that the relation between the text and the visual should be considered when presenting them together (e.g., Carney & Levin, 2002; Twyman, 1987). For example, the visual can have a low or high informative value, e.g., the visual represents the information mentioned in the text or the visual explains the information mentioned in the text as in Figure 1.1. According to research by Glenberg & Robertson (1999), informative visuals allow readers to 'index' information presented in text to the information presented in a visual, hence helping readers to make relevant "affordances" (Gibson, 1972). The term affordances refers to the actions that an individual can potentially perform in their environment. Thus, in this example, when a multimodal information presentation is well-designed, users will be able to derive the proper actions in organizing an ergonomic workspace. Chapter 2 discusses these basic issues around multimodal information presentation through the following research questions:

- **When and how do people present information in a multimodal way?**
- **How do people evaluate unimodal and multimodal information presentations?**

Well-designed multimodal information presentations not only facilitate comprehension, they can also help users find the appropriate information quickly. This is especially important in large multimodal information presentations, like web sites. Users often experience problems when searching for information in web sites, like disorientation and cognitive overload (e.g., Ahuja & Webster, 2001; Conklin, 1987; Elm & Woods, 1985). Therefore, several multimodal navigational aids (e.g., sitemaps, bread crumbs) have been developed aimed at helping users to create a representation of the structure or content of the web site or to clarify the users' position within the web site (Maes, Van Geel & Cozijn, 2006). However, studies on the effectiveness of these navigational aids show equivocal results (e.g., Dias & Sousa, 1997; Hofman & Van Oostendorp, 1999). In order to help users finding the information they need, we first have to investigate how they conceptualize web sites. There are several indications that the spatial character of web sites plays an important role in users' conceptualisation (e.g., Boechler, 2001; Maglio & Matlock, 2003). Therefore, Chapter 3 sets out to explore how users conceptualize their actions when navigating a web site through the following research question:

- **How do users conceptualize their actions when navigating in multimodal information environments?**

Another question that arises in multimodal information presentation is which presentation mode is most effective for a particular learning task (e.g., learning how to organize an ergonomic workspace). For instance, it might be that a text is most effective in expressing abstract matters, whereas a static visual (e.g., photo or graphic) might be most effective in representing perceptual information. A dynamic visual (e.g., film clip or animation) is argued to be best in representing temporal aspects (Park & Hopkins, 1993). Moreover, much of the empirical research on the effectiveness of different presentation modes has focused on declarative tasks, where a learner acquires knowledge about a certain topic (e.g., meteorological changes as in Lowe, 2004) It is unclear to what extent findings for learning declarative tasks carry over to learning procedural tasks, where a learner acquires a certain skill (e.g., bandaging a hand as in Michas & Berry, 2000). Chapter 4 focuses on the effectiveness of different presentation modes for a specific learning task, i.e.,

Figure 1.1

Possible answer presentations to the question "How to organize my workspace to prevent RSI?"

procedural instructions. The characteristics of the presentation modes (i.e., text, photo, and film clip) as well as learners' preferences are taken into account through the following research questions:

- **Which presentation modes are most effective for learning and executing procedural instructions?**
- **Which presentation modes do people prefer when learning procedural instructions?**

Textual instructions on organizing an ergonomic workspace, could be presented visually but also auditorily. In fact, the modality principle states that when a multimodal information presentation consists of text and visuals, the text should be presented as spoken text rather than as visual text (Mayer & Moreno, 1998; Moreno & Mayer, 1999). But when following the modality principle and using spoken text instead of written text, the question arises which kind of voice should be used. Mayer, Sobko, and Mautone (2003) investigated the effectiveness of a human voice and a machine-synthesized voice that accompanied an animation that explained how lightning storms develop. They found that people learned better with a human voice than with a machine-synthesized voice. However, developments in speech technology have led to a frequent use of synthetic speech in computer applications, like computer-aided instructions and consumer products (e.g., navigational aids and mobile telephones) (Paris, Thomas, Gilson & Kincaid, 2000). There are two reasons why synthetic speech is harder to comprehend than human speech. First, synthetic speech is less intelligible than human speech as the acoustic signals of synthesized speech are impoverished (e.g., Luce, Feustel, & Pisoni, 1983; Nusbaum & Pisoni, 1985). Second, synthetic speech sounds unnatural compared to human speech due to the limited modeling of prosodic cues, like intonation, stress, and durational patterns (Nusbaum, Francis & Henly, 1995). Currently, there are two common ways to create speech synthesis. The first is diphone synthesis which is based on concatenating prerecorded phoneme transitions (i.e., diphones), followed by signal processing to obtain the required pitch and duration. The second is unit selection synthesis which is also based on concatenation and is realized by segmenting prerecorded human speech in units of variable sizes (e.g., sentences, words, and diphones). In

sum, evaluating multimodal information presentations not only implies evaluating different presentation modes, but also the quality differences within the same modality. Chapter 5 focuses on the quality differences between synthetic speech and human speech using the following research question:

- **How do quality differences within the speech modality influence its incremental processing and how can we assess these quality differences?**

1.3 Research approach

In the previous section, we mentioned that several factors should be considered when presenting information in a multimodal way. In this section, we will argue that knowledge on multimodal information presentation can be obtained using different research methodologies. In this thesis, each chapter discusses a different research methodology used to evaluate multimodal information presentations.

In the research field of speech and language technology there is a growing interest in multimodal human computer interaction. Past research in human-computer interaction has shown that the use of multiple output modalities makes systems more robust and efficient to use (Oviatt, 1999). Also, in the area of computational linguistics, research has been done on multimodal documents analysis and generation (e.g., Bateman, Kamps, Reichenberger & Kleinz, 2001). In multimodal systems guidelines are needed to combine the different modalities in such a way that each bit of information is presented in the most appropriate manner. A way to generate optimal multimodal presentations is investigating when and how human users present information in a multimodal way. Chapter 2 starts from multimodal human computer interaction and describes two experiments using the cognitive engineering approach (Tversky et al., 2006). In this approach, human users are asked to produce information presentations, which are then rated by other users (e.g., Agrawala & Stolte, 2001; Heiser, Phan, Agrawala, Tversky & Hanrahan, 2004).

Web site usability research investigates which factors influence the effectiveness and the efficiency in navigating a web site. For example, McDonald & Stevenson (1999) investigated the effects of different navigational aids (spatial map vs.

conceptual map) using performance measures, like the number of opened pages and the number of pages recalled. However, the relation between these performance measures and how users mentally conceptualise a web site is unclear. For instance, suppose users are presented with the spatial map and open many web pages. Does this mean that they have a clear overview of the web site or that they are disoriented? Users' representation of a web site can be investigated using other methods, like protocol analysis. In this research method, participants are asked to carry out a task, while verbalizing their thoughts. These verbalizations are written down in a verbal report and analyzed in a way that depends on the research question (Ericsson & Simon, 1993). Chapter 3 discusses an exploratory study in which protocol analysis is used to get a fine-grained view of how users conceptualize their actions when navigating a web site.

In the field of cognitive and instructional psychology research has been done on the influence of different presentation modes on the users' understanding, recall, and processing efficiency of the presented material (e.g., Mayer, 2005; Tversky, Morrison & Bétrancourt, 2002). Several studies compared the effectiveness of different presentation modes, however with mixed results (e.g., Bétrancourt & Tversky, 2000; Lewalter, 2003; Tversky et al., 2002). Various reasons have been mentioned for these findings: the lack of equivalence of information in the different presentation modes (Tversky et al., 2002), differences in learning tasks (Hegarty, 2004), or in learning performance measures (Brünken, Plass & Leutner, 2003). Apart from the objective effectiveness of different presentation modes, users' 'subjective satisfaction' (Nielsen, 1993) should also be taken into account, as an attractive and motivating presentation format could also influence its effectiveness. Chapter 4 describes two experiments. In the first experiment, the effectiveness of three presentation modes (i.e., text, photo, and film clip) was evaluated using several objective measures, like learning times and recall. In the second experiment, we investigated whether users subjectively preferred one of these three presentation modes.

Research in speech synthesis has evaluated the intelligibility and naturalness of synthetic speech with offline research methods. For example, in the Modified Rhyme Test (House, Williams, Hecker & Kryter, 1965) listeners are presented with spoken words and are instructed to select the word they heard from a set of alternatives that differ only in one phoneme. Another example is the Mean Opinion Score (Schmidt-

Nielsen, 1995) in which listeners have to rate the quality of spoken sentences on scales (i.e., excellent - bad). Yet, these research methods do not consider that speech is transient: spoken instructions are "gone" once they have been uttered. Online research methods, like eye tracking, give a direct insight in how speech is processed incrementally. Chapter 5 describes an eye tracking experiment in which the visual world paradigm (e.g., Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995) is used to evaluate the processing of synthetic speech and human speech.

1.4 Thesis Overview

Figure 1.2 gives a 'multimodal' overview of this thesis. Chapter 2 offers a general introduction into multimodal information presentation and presents two studies. The first study, a production experiment, was conducted to investigate when and how users present medical information in a multimodal way. The second study, an evaluation experiment, was done to investigate how users evaluate the informativity and attractiveness of unimodal and multimodal information presentations. The later chapters are more detailed case studies looking into multimodal information presentation from different perspectives.

Chapter 3 focuses on the research question how users conceptualize their actions when navigating a web site. Thinking aloud protocols were analyzed to distinguish users' actions involved in web sites navigation and the type of expressions used to verbalize these actions.

Chapter 4 also presents two studies. The first study describes an experiment investigating a specific kind of procedural instructions, i.e., RSI exercises, taking presentation mode (text vs. photo vs. film clip) and difficulty degree of the exercises (easy vs. difficult) as independent variables. The second study describes an experiment concentrating on which presentation people prefer when learning RSI exercises.

Chapter 5 takes a closer look at the speech modality. An eye tracking experiment was conducted to study the incremental processing of two forms of speech synthesis (i.e., diphone synthesis and unit selection synthesis) and human speech taking segmental and suprasegmental speech quality into account

Finally, Chapter 6 presents a review of the results found as well as a general discussion of the most interesting findings of this thesis.
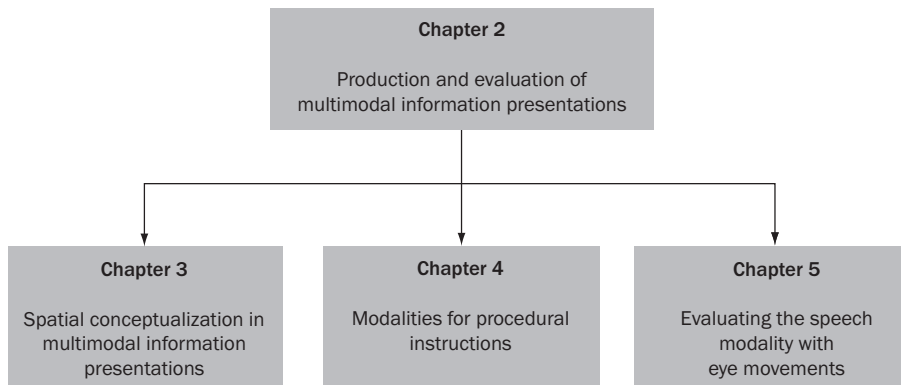
Figure 1.2

Overview of this thesis.

2 | Production and evaluation of multimodal information presentations

TASK MODALITY METHODOLOGY

Procedural Instructions

Speech Quality Segmental Suprasegmental

Action Types Spatial Expressions

Action Levels

Diphone Synthesis

Spatial Conceptualization

Thinking Aloud

Protocol Analysis

RSI Exercises

Eye Tracking

Hypertext

Illustrative

Questions and Answers

Text Brief

Extended Visuals Informative

Visual World Paradigm

Unit Selection Synthesis

Static

Cognitive Engineering

Dynamic

Production

Human Speech

Information Seeking

Learning Preference

Evaluation

Speech Modality

Experimental Evaluation

A journal paper based on this chapter is submitted for publication. Earlier versions of this chapter appeared as Van Hooijdonk, C.M.J., De Vos, J., Krahmer, E.J., Maes, A., Theune, M., & Bosma, W. (2007). On the role of visuals in multimodal answers to medical questions. *Proceedings of the International Professional Communication Conference (IPCC)*, Seattle, USA: IEEE and as Van Hooijdonk, C.M.J., Krahmer, E. J., Maes, A., Theune, M., & Bosma, W. (2007). Towards automatic generation of multimodal answers to medical questions: a cognitive engineering approach. *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*, Aberdeen, Scotland, pp. 93-104.

2.1 Introduction

This chapter offers a first exploration into multimodal information presentation from the perspective of human-computer interaction. More specifically, we take the perspective of multimodal presentation of answers in question answering (QA). Early research in the field of QA concentrated on answering factoid questions, i.e., questions that have one word or phrase as their answer, such as "Amsterdam" in response to the question "What is the capital of the Netherlands?" The output modality to these questions will typically be text. However, there is currently a growing interest in moving beyond factoid questions and purely textual answers, and then output generation becomes an important issue. Questions that arise are: how to determine for a given question, what the best combination of modalities for the answer is? And related to this: what is the proper length of a non-factoid answer? In this chapter, we address these basic issues around multimodal information presentation in the context of medical question answering.

In the medical domain several question types occur, such as definition questions and procedural questions, which require different types of answers. For example the answer to the definition question "What does RSI stand for?" would probably be a brief textual answer, like "RSI stands for Repetitive Strain Injury". However, a text only answer may not be the best choice for every type of information. In some cases other modalities (e.g., pictures, film clips, etc.) or modality combinations (e.g., text and a picture) may be more suitable (Theune et al., 2007). For example, the answer to the procedural question "How to organize a workspace in order to prevent RSI?" would probably be more informative if it contained a picture. Moreover, the length of the answer could also play an important role in the answer presentation. For example, the answer to the question "What does RSI stand for?" could be an extended one: "RSI stands for Repetitive Strain Injury. This disorder involves damage to muscles, tendons and nerves caused by overuse or misuse, and affects the hands, wrists, elbows, arms, shoulders, back, or neck". This answer provides the user with relevant background information about the topic of the question. In addition, including informative text in the answer may allow the user to assess the answer's accuracy in order to verify whether it is correct or not (Bosma, 2005). This raises the question how to determine for a given question, what the best combination of

modalities for the answer is. And related to this: what is the proper length of an answer?

Much research has been done in the field of cognitive and eductional psychology on the influence of (combinations of) different modalities on the users' understanding, recall and processing efficiency of the presented material (e.g., Carney & Levin, 2002; Mayer, 2005; Tversky et al., 2002). This research has resulted in several guidelines on how to present (multimodal) information to the user, such as the multimedia principle (i.e., instructions should be presented using both text and pictures, rather than text only) and the spatial contiguity principle (i.e., when presenting a combination of text and pictures, the text should be close to or embedded within the pictures) (Mayer, 2005). However, these guidelines are based on specific types of information used in specific domains, in particular descriptions of cause and effect chains which explain how systems work (e.g., Mayer, 1989; Mayer & Gallini, 1990; Mayer & Moreno, 2002) and procedural information describing how to acquire a certain skill (e.g., Marcus, Cooper & Sweller, 1996; Michas & Berry, 2000; Schwan & Riempp, 2004). Yet, these guidelines do not tell us which modalities are most suited for which information types, as each learning domain has its own characteristics (Van Hooijdonk & Krahmer, in press).

Several researchers have tried to make an overview of the characteristics of modalities, information types, and the matches between them. For example, Bernsen (1994) focused on the features of modalities in his Modality Theory, i.e., "Given any particular set of information which needs to be exchanged between user and system during task performance in context, identify the input/output modalities, which, from the user's point of view, constitute an optimal solution to the representation and exchange of that information" (Bernsen, 1994, p. 348). He proposed a taxonomy to define generic unimodalities consisting of various features. Other researchers proposed taxonomies of information types such as dynamic, static, conceptual, concrete, spatial, and temporal in order to select the appropriate modalities (e.g., Heller, Martin, Haneef & Guevka-Kriliu, 2001; Sutcliffe, 1997).

Other research has been concerned with the so-called "media allocation problem": "How does a producer of a presentation determine which information to allocate to which medium, and how does a perceiver recognize the function of each part as displayed in the presentation and integrate them into a coherent whole?" (Arens,

Hovy & Vossers, 1993, p. 280). According to Arens et al. (1993) the characteristics of the media used are not the only features that play a role in media allocation. The characteristics of the information to be conveyed, the goals and characteristics of the producer, and the characteristics of the perceiver and the communicative situation are also important. In order to create a multimodal information presentation, modalities should be integrated dynamically based on a general communication theory (e.g., Arens et al., 1993; André, 2000; Maybury & Lee, 2000; Oviatt et al., 2003).

In short, attempts have been made to generate optimal multimodal information presentations resulting in several guidelines, frameworks, and taxonomies. However, what is needed in addition is gaining knowledge on when and how people produce multimodal information presentations and how other people evaluate such presentations. To achieve this goal, we carried out two experiments following the cognitive engineering approach as used by Heiser et al. (2004). In this approach, people are asked to produce information presentations (e.g., route maps, assembly instruction, etc.), which are then rated by other people. Based on the results, design principles are identified and used to improve these information presentations.

This chapter describes two experiments carried out in order to investigate the role of visuals in multimodal answer presentations for a medical question answering system. First, a production experiment is described that focuses on which modalities users choose to answer medical questions. Participants were instructed to create a brief and an extended answer to different medical question types (i.e., definition questions, like: "Where is progesterone produced?" vs. procedural questions, like "How is a SPECT scan made?"). Next, an evaluation experiment is described that concentrates on how users evaluate different types of answer presentations. Participants were instructed to carefully study answer presentations that were either unimodal (i.e., consisting of text only) or multimodal (i.e., consisting of text and a picture), and that were based on the answer presentations collected in the production experiment. After the participants had studied these answer presentations, they had to assess them on their informativity and attractiveness. Subsequently, the participants received a post-test to determine how much of the information presented in the answer presentations they could recall.

2.2 Experiment I: Production

2.2.1 Research method

Participants

One hundred and eleven students of Tilburg University participated for course credits (65 female and 46 male, between 19 and 33 years old). All participants were native speakers of Dutch.

Stimuli

The participants were given one of four sets of eight general medical questions (see appendix A) for which the answers could be found on the Internet. The participants had to give two types of answers per question i.e., a brief answer and an extended answer. Besides, different (combinations of) modalities could be used to answer the questions. The participants had to assess for themselves which (combinations of) modalities were best for a given question, and they were specifically asked to present the answers as they would prefer to find them in a QA system. To make sure they could carry out this task, they were instructed about the working of QA systems in advance. Questions and answers had to be presented in a fixed format in PowerPoint™ with areas for the question ("vraag") and the answer ("antwoord"). This programme was chosen because it has the possibility to insert pictures, film clips, and sound fragments in an answer presentation. All participants were familiar with PowerPoint™ and most of them used it on a monthly basis (51,4%).

Of the eight questions in each set, four were randomly chosen from one hundred medical questions formulated to test the IMIX QA system (e.g., "How many X chromosomes does a female body cell have?"). Of the remaining four questions, two were definition questions and two were procedural questions. Orthogonal to this, two questions referred explicitly or implicitly to body parts and two did not. These four question types were given to the participants in a random order. Examples of the questions were:

- Definition question referring to body parts: "Where is progesterone produced?" or "Where are red blood cells produced?"
- Definition question not referring to body parts: "What are the side effects of ibuprofen?" or "What are thrombolytic drugs?"

- Procedural question referring to body parts: "How to apply a sling to the left arm?" or "What should be done when having a nosebleed?"
- Procedural question not referring to body parts: "What happens when a myelogram is taken?" or "How is a SPECT scan made?"

Coding system

Each answer was coded on the following variables: the presence of photos, graphics, animations, and the function of these visual media related to the text of the answer. The coding criteria for these variables are discussed below. To determine the reliability of the coding system, Cohen's κ (Krippendorff, 1980) was calculated.

- Photos: We distinguished whether the answer contained no photo, one photo or several photos.
- Graphics: We defined graphics as non-photographic, static depictions of concepts (e.g., diagrams, charts, and line drawings). We distinguished answers with no graphics, one graphic, or several graphics.
- Animations: We defined animations as dynamic visuals possibly with sound (e.g., film clips and animated pictures). We distinguished answers without animations, with one animation, or several animations.
- Function of visual media: We distinguished three functions of visuals in relation to text, loosely based on Carney & Levin (2002)[1]:
  1. *Decorational function:* a visual has a decorational function if removing it from the answer presentation does not alter the informativity of the answer in any way. Figure 2.1 shows an example of answer presentations with a decorational visual. The example shows an answer to the question: "What are the side effects of a vaccination for diphtheria, whooping cough, tetanus, and polio?" The answer consists of a combination of text and a graphic. The text describes the side effects of the vaccination, while the graphic only shows a syringe. The graphic does not add any information to the answer. The example on the right shows an answer to the question: "How many X chromosomes does a female body cell have?" The answer consists of a combination of text and a graphic. In text the answer is given (i.e., a female body cell has two X chromosomes). The answer would not be less informative if the graphic was absent.

2. *Representational function*: a visual has a representational function if removing it from the answer presentation does not alter the informativity of the answer, but its presence clarifies the text. Figure 2.2 shows two examples of answer presentations with a representational visual. The example on the left shows an answer to the question: "What types of colitis can be distinguished?" The answer consists of a combination of text and a graphic. The text describes the four types of colitis and their occurrence in the intestines. This information is visualized in the graphics. The example on the right shows an answer to the question: "How to apply a sling to the left arm?" The answer consists of three photos illustrating the procedure, which is described in more detail in the text on the right.

3. *Informative function*: a visual has an informative function if removing it from the answer presentation decreases the informativity of the answer. If an answer only consists of a visual, it automatically has an informative function. Figure 2.3 shows two examples of answer presentations with informative visuals. The example on the left shows the answer to the question: "How to apply a sling to the left arm?" The answer consists of four graphics illustrating the procedure. The example on the right shows an answer to the question: "How can I strengthen my abdominal muscles?" The text describes some general information about abdominal exercises (i.e., an exercise program should be well balanced and train all abdominal muscles). The photos represent four exercises that can be done to strengthen the abdominal muscles.

Coding procedure

In total 1776 answers were collected (111 participants × 8 questions × 2 answers). However, one participant gave 15 answers resulting in one missing value. Thus, the coded corpus consisted of 1775 answers. The coding scheme was given to six analysts. The annotation was done in two steps. First, each analyst independently coded a part of the corpus to determine the adequacy of the coding scheme. Differences between the analysts were discussed, which resulted in some adjustments of the coding system. Subsequently, every analyst independently coded the same set of 112 answers. Second, every analyst independently coded a part of the total corpus (i.e., approximately 300 answers).

To compute agreement we used Cohen's κ measure. Following standard practice, Cohen's κ scores between .81 and 1.00 signify an almost perfect agreement, between .61 and .80 signify a substantial agreement, between .41 and .60 is a moderate agreement, and between .21 and .40 is a fair agreement (Rietveld & Van Hout, 1993). It turned out that the analysts almost perfectly agreed in judging the occurrence of photos (κ = .81), graphics (κ = .83), and animations (κ = .92). Moreover, an almost perfect agreement was reached in assigning the function of the visual media (κ = .83).

Figure 2.1

Examples of answer presentations with decorational visuals

Figure 2.2

Examples of answer presentations with representational visuals

Figure 2.3

Examples of answer presentations with informative visuals

2.2.2 Results

Descriptive statistics

Table 2.1 shows the percentages of visual media (overall), photos, graphics, and animations in the complete corpus of coded answer presentations. Inspection of Table 2.1 reveals that almost one in four answers contained one or more visual media, of which graphics were most frequent and animations were least frequent. The presence of photos was between these two. In some answers several visual media occurred (i.e., photos, graphics, and animations). These instances were counted as one occurrence of visual media. Thus, the sum of the percentages of photos, graphics, and animations in the corpus exceeded the percentage of the variable visual media.

Table 2.1

Percentages of answer presentations containing text only (no visual media) and visual media (overall) divided into photos, graphics and animations in the complete corpus of coded answers (n = 1775).

No visual media	75.1
Visual media	24.9
Photos	8.6
Graphics	14.9
Animations	3.8

Table 2.2 shows the percentages of photos, graphics, and animations related to their function. Note that in some answers several visuals occurred (i.e., photos, graphics, and animations). These instances were counted as one occurrence of visual media. Thus, the sum of the percentages of photos, graphics, and animations in the corpus exceeded the percentage of the overall occurrence of visual media. Table 2.2 reveals that the distribution of photos related to their function differed significantly from chance (χ^2 (2) = 41.30, p< .001). Most photos had a representational function. Also, there was an association between graphics and their function (χ^2 (2) = 38.09, p< .001). Most graphics had a representational function. Finally, there was a relation between animations and the function of visual media (χ^2 (2) = 67.52, p< .001). Most animations had an informative function.

Table 2.2

Percentages of photos, graphics, and animations related to their function.

	Function of visual media			
	Decorational	Representational	Informative	Totals
Photos (n = 152)	20.4	57.9	21.7	100.0
Graphics (n = 265)	15.8	45.3	38.9	100.0
Animations (n = 67)	7.5	11.9	80.6	100.0

Within the corpus of collected answer presentations different types of photos and graphics occurred. It turned out that some photos and graphics contained text and some did not. Therefore, a sub-analysis was done to investigate whether the distribution of the functions of visual media differed between photos with and without text and between graphics with and without text. Table 2.3 shows the results. It turned out that photos without text occurred significantly more often than photos with text (χ^2 (1) = 60.63, p< .001). The reverse was found for graphics: graphics with text occurred significantly more often than graphics without text (χ^2 (1) = 38.49, p< .001).

There was a dependence between the function of visual media and photos with and without text (χ^2 (2) = 5.97, p = .05). Most photos without text were associated with a representational function or an informative function (χ^2 (2) = 37.37, p< .001).

Table 2.3

Percentages of types of photos and types of graphics related to their function.

	Function of visual media			
	Decorational	Representational	Informative	Totals
Photos without text (n = 124)	16.9	58.9	24.2	100.0
Photos with text (n = 28)	35.7	53.6	10.7	100.0
Graphics without text (n = 82)	30.5	40.2	29.3	100.0
Graphics with text (n = 183)	9.3	47.5	43.2	100.0

However, most photos with text were associated with a representational function or a decorational function (χ^2 (2) = 7.79 p< .025). Also, the distribution of the functions of visual media differed significantly between the graphics with and without text (χ^2 (2) = 19.54, p< .001). There was no association between graphics without text and their function (χ^2 (2) = 7.78, p = .41). Graphics without text were evenly associated with the three functions of visual media. However, there was an association between graphics with text and their function (χ^2 (2) = 48.13, p< .001). Most graphics with text had a representational or an informative function.

Answer length

The brief and the extended answers were related to different answer presentations. Table 2.4 shows the percentages and χ^2 statistics of the presence of visual media (overall), photos, graphics, and animations within the brief and extended answers. The results showed that visual media occurred significantly more often within the extended answers. Note that in some answers several visuals occurred (i.e., photos, graphics, and animations). These instances were counted as one occurrence of visual media. Thus, the sum of the percentages of photos, graphics, and animations in the corpus exceeded the percentage of the overall occurrence of visual media.

Table 2.4

Percentages and χ^2 statistics of the presence of visual media (overall) divided into photos, graphics, and animations related to the brief and the extended answers (Scores are percentages of answers; n = 1775).

	Length of the answer		
	Brief (n = 888)	Extended (n = 887)	χ^2 statistics
Visual media	11.4	38.4	χ^2 (1) = 173.89, p< .001
Photos	4.6	12.5	χ^2 (1) = 35.34, p< .001
Graphics	6.3	23.6	χ^2 (1) = 104.04, p< .001
Animations	.9	6.7	χ^2 (1) = 40.40, p< .001

Table 2.5 shows the percentages and χ^2 statistics of the functions of visual media related to brief and extended answers. The results showed that the overall distribution of the functions of visual media across the answer types differed significantly (χ^2 (2) = 34.31, p< .001). Decorational visuals occurred more often in brief answers, whereas representational visuals occurred more often in extended answers. Finally, informative visuals occurred more often in brief answers.

Table 2.5

Percentages of the function of visual media related to brief and extended answers (n = 444)

	Length of the answer		
	Brief (n = 102)	Extended (n = 342)	χ^2 statistics
Decorational function	26.5	12.9	χ^2 (1) = 4.07, p< .05
Representational function	20.6	52.9	χ^2 (1) = 126.73, p< .001
Informative function	52.9	34.2	χ^2 (1) = 23.21, p< .001

Type of question

We were interested whether different types of questions were related to different answer presentations. Therefore we analyzed a subset of the medical questions (i.e., the definition and procedural questions with and without reference to body parts). Table 2.6 shows the percentages and χ^2 statistics of the presence of visual media (overall), photos, graphics, and animations within the definition and procedural questions and within questions with and without reference to body parts. The distribution of all variables differed significantly across the question types. In general, visual media were more frequent within procedural questions with reference to

body parts. Looking at specific types of visual media, we see that graphics occurred more often in answers to definition questions with reference to body parts, but that photos and animations occurred more often in answers to procedural questions with reference to body parts.

Table 2.6

Percentages and χ^2 statistics of the presence of visual media (overall) divided into photos, graphics, and animations related to the four question types.

	Definition questions (n = 443)		Procedural questions (n = 444)		χ^2 statistics
	Body parts (n = 222)	¬Body parts (n = 221)	Body parts (n = 222)	¬Body parts (n = 222)	
Visual Media	31.1	10.0	47.7	33.3	χ^2 (3) = 53.09, p< .001
Photos	4.1	5.4	22.1	19.8	χ^2 (3) = 46.07, p< .001
Graphics	28.8	5.0	15.3	12.6	χ^2 (3) = 42.77, p< .001
Animations	.5	.9	14.9	5.4	χ^2 (3) = 55.17, p< .001

Table 2.7 shows the percentages and χ^2 statistics of the functions of visual media within definition and procedural questions and within questions with and without reference to body parts. The results show that the distribution of the functions of visual media differed significantly within the question types (χ^2 (6) = 91.84, p< .001). Decorational visuals occurred more often in definition questions without reference to body parts. Representational visuals occurred more often in definition questions with reference to body parts. Finally, informative visuals occurred more often in procedural questions with reference to body parts.

Table 2.7

Percentages and χ^2 statistics of the functions of visual media related to the four question types (n = 272).

	Definition questions (n = 91)		Procedural questions (n = 181)		χ^2 statistics
	Body parts (n = 69)	¬Body parts (n = 22)	Body parts (n = 106)	¬Body parts (n = 75)	
Decorational function	5.8	63.6	3.8	8.0	χ^2 (3) = 9.71, p< .025
Representational function	63.8	22.7	39.6	52.0	χ^2 (3) = 31.42, p< .001
Informative function	30.4	13.6	56.6	40.0	χ^2 (3) = 59.68, p< .001

2.2.3 Conclusion

The results of the production experiment showed that users do make use of multiple media in their answer presentations and that the design of these presentations is affected by the answer length and question type. However, what is not clear is how users evaluate different types of answer presentations (i.e., unimodal vs. multimodal). In the next section, an evaluation experiment is discussed in which users were instructed to assess answer presentations on their informativity and attractiveness.

2.3 Experiment II: Evaluation

2.3.1 Research method

Participants
Participants were 108 native speakers of Dutch (66 female and 42 male, between 18 and 64 years old). None had participated in the production experiment.

Design
The experiment had a 4 (answer presentation) × 2 (question type) mixed factorial design, with answer presentation (brief answer with an illustrative visual, extended answer with an illustrative visual, brief answer with an informative visual, and

an extended answer with an informative visual) as between participants variable and question type as within participants variable. The dependent variables were the participants' assessment of the informativity and the attractiveness of the text and visual combinations and the number of correct answers in the post-test. The participants were randomly assigned to an experimental condition.

Stimuli

For the evaluation experiment, 16 medical questions were selected from the set of 32 medical questions of the production experiment. We selected questions for which the production corpus contained two relevant types of visuals: informative visuals and decorational or representational visuals. For the purpose of this experiment, decorational and representational visuals were combined into illustrative visuals. An illustrative visual did not add any more information to the textual answer, whereas an informative visual did add information to the textual answer.

The selected set of medical questions consisted of eight definition questions and eight procedural questions. In both question types, half of the questions referred to body parts and half did not. Examples of the questions used in the evaluation experiment were:

- Definition questions: "Where is testosterone produced?" or "What does ADHD stand for?"
- Procedural questions: "How to apply a sling to the left arm?" or "How to organize a workspace in order to prevent RSI?"

The 16 medical questions were presented in four different answer presentation formats: a brief textual answer with an illustrative visual, an extended textual answer with an illustrative visual, a brief textual answer with an informative visual, and an extended textual answer with an informative visual. For the sake of comparison, two unimodal answer presentation formats were added: a brief textual answer and an extended textual answer.

For every question a brief and an extended textual answer was formulated. The brief and the extended textual answers were based on the answers found in the corpus of answer presentations collected in the production experiment. Small adjustments were made to these answers in order to make them more comparable. The brief answer always gave a direct answer to the question, while the extended answer also

provided some relevant background information about the topic of the question. The average length of the brief answer was almost 26 words and the average length of the extended answers was almost 66 words. The same brief and extended answers were also used in the text with an illustrative visual condition and in the text with an informative visual condition.

In the two text with an illustrative visual conditions, the brief and the extended textual answers were presented together with an illustrative visual. An illustrative visual had been given a decorational or a representational function in the production experiment (see section 2.2.1). Figure 2.4 shows an example of a brief textual answer and an extended textual answer with an illustrative photo. Both examples show the answer to the question: "How to organize a workspace in order to prevent RSI?" The answer presentation on the left contains a brief textual answer describing three tips for organizing a workspace in order to prevent RSI. The answer presentation on the right contains an extended textual answer describing an ergonomic workspace. Both answer presentations contain a photo illustrating a workspace. This photo represents an element (i.e., a desk) mentioned in the textual answers. However, the answers would not be less informative if the photo was not present.

Figure 2.4

Examples of a brief textual answer (left) and an extended textual answer (right) with an illustrative visual

Figure 2.5

Examples of a brief textual answer (left) and an extended textual answer (right) with an informative visual

In the two text with an informative visual conditions, we presented the brief and extended textual answers together with an informative visual. A visual was informative if it had been given an informative function in the production experiment. Figure 2.5 illustrates a brief textual answer and an extended textual answer with an informative graphic to the question: "How to organize a workspace in order to prevent RSI". Both answer presentations include a graphic depicting in detail an ergonomic workspace. Both answer presentations would contain less information if the graphic was not present.

We made sure that the type of question did not affect the answer length for brief textual answers (F [1,14] = 3.59, p = .08), nor for extended textual answers (F< 1). The illustrative and informative visuals were taken from the corpus of answer presentations collected in the production experiment. In a few cases, a visual was used from the Internet, when the corpus did not contain a suitable visual. Moreover, in a few cases the text within the visuals was enlarged to make it more readable.

The experiment was conducted using WWSTIM (Veenker, 2005), a CGI-based script that automatically presents stimuli to the participants and transfers all data to a database. This enabled us to run the experiment via the Internet. The answer presentations of procedural and definition questions were presented in one random order.

Procedure

The participants received an e-mail inviting them to take part in the experiment. 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. Figure 2.6 illustrates the procedure of the evaluation experiment.

When the participants accessed the experiment, they first received instructions about the procedure of the experiment. In these instructions, the participants were told that they would receive the answer presentations of 16 medical questions. They had to study these answer presentations carefully, after which they had to assess them on their informativity and on 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 the participants had filled out their personal data, they practiced the procedure of the actual experiment in a practice session: they were presented with the medical question "Where are red blood cells produced?" together with an answer presentation. The participants studied the answer presentation until they thought that they could assess its informativity and attractiveness. Subsequently, the participants were shown the medical question, the answer presentation, and a questionnaire. In the unimodal (i.e., text only) conditions, this questionnaire consisted of three questions addressing the formulation of the answer presentation, the informativity of the answer presentation, and the attractiveness of the answer presentation. In the four texts with a visual conditions, the participants filled out the above-mentioned questions and two other questions addressing the informativity and the attractiveness of the text and visual combination. The participants could indicate their assessment on a seven-point Likert scale, implemented as radio buttons. After completing the practice session, the participants started with the actual experiment, proceeding in the same way as during the practice session.

After completing the assessment of the answer presentations to the 16 medical questions, the participants received a post-test: they had to answer the same 16 medical questions by means 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. An example is "Where is testosterone produced?"
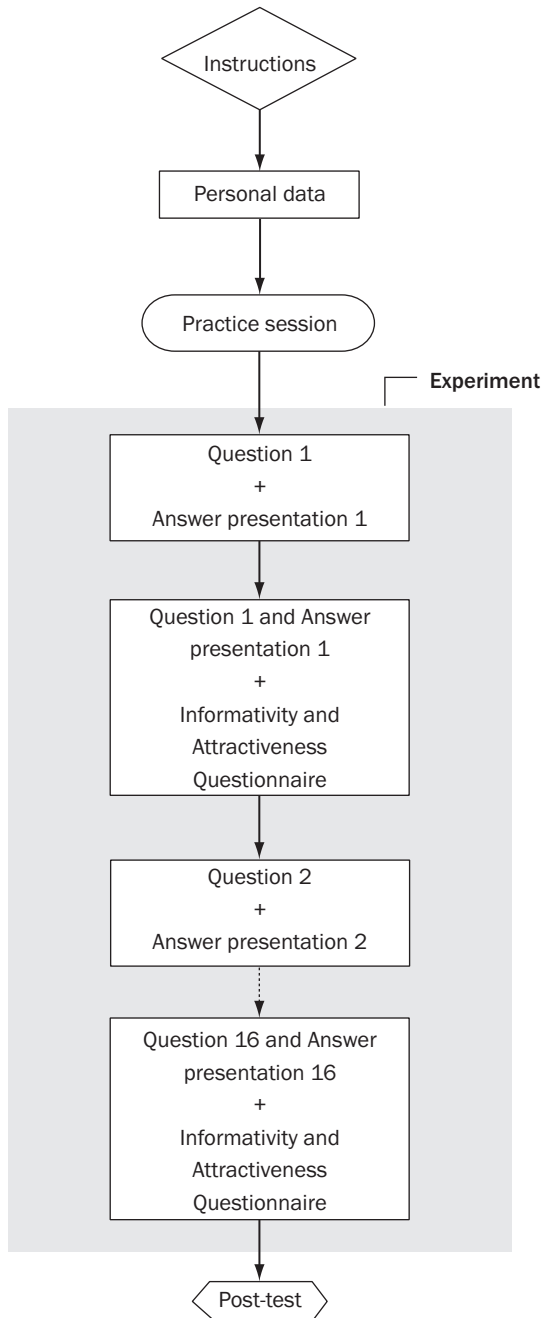
Figure 2.6

Procedure of the evaluation experiment

a. Testosterone is a sex hormone that is produced by males and females in the adrenal glands. Besides, males produce testosterone in the testes. (correct answer)
b. Testosterone is a sex hormone that is only produced by males. Testosterone is produced in the testes and in the adrenal glands. (incorrect answer)
c. Testosterone is a sex hormone produced by males and females. Testosterone is produced in the pancreas and in the hypothalamus. (incorrect answer)
d. Testosterone is a sex hormone produced by males and females. Testosterone is produced in the adrenal glands. (incorrect answer)

The order in which the medical questions were presented in the post-test was the same as in the actual experiment. Note that the information mentioned in the extended textual answers, and illustrated in the informative visuals was not necessary to answer the question in the post-test correctly.

Data processing
The following data were collected: the informativity and the attractiveness of the text and visual combination of the answer presentations, and the number of correctly answered questions of the post-test. Tests for significance were performed using a 4 (answer presentation) × 2 (question type) repeated measures analysis of variance (ANOVA), with a significance threshold of .05. For post hoc tests, the Bonferroni method was used. The participants were randomly assigned to an experimental condition. Note that inconclusive results were found for answer presentations to questions with and without reference to body parts. Therefore, we do not report on this any further.

2.3.2   Results

Informativity of the text and visual combinations
Table 2.8 shows the mean results of the assessment on the informativity of the text and visual combinations. A main effect was found of answer presentation format on the perceived informativity of the text and visual combinations, $F [3,68] = 9.32$, $p < .001$, $\eta^2_p = .29$. Brief answers with an informative visual were evaluated as most informative, while brief answers with an illustrative visual were evaluated as least informative. Post-hoc tests showed that brief answers with an illustrative visual did

not differ significantly from extended answers with an illustrative visual (p = 1.00). However, brief answers with an illustrative visual differed significantly from both brief (p< .001) and extended (p< .005) answers with an informative visual. Also, extended answers with an illustrative visual differed significantly from brief (p< .025) and extended (p< .025) answers with an informative visual. No significant differences were found between brief and extended answers with an informative visual (p = 1.00).

Table 2.8

Mean results of the assessment on the informativity and the attractiveness of the four text and visual combinations (Scores range from 1 = "very negative" to 7 = "very positive"; standard deviations in parenthesis).

Factor	Question type	Text with an illustrative visual				Text with an informative visual			
		Brief		Extended		Brief		Extended	
Informativity of the text and visual combination	Definition	3.83	(1.13)	4.01	(1.30)	4.91	(.81)	4.97	(1.20)
	Procedural	3.70	(1.26)	4.27	(1.18)	5.53	(.70)	5.40	(.84)
	Totals	3.76	(1.16)	4.14	(1.19)	5.22	(.69)	5.18	(1.00)
Attractiveness of the text and visual combination	Definition	3.93	(.87)	3.76	(1.14)	4.43	(.88)	4.69	(1.01)
	Procedural	4.18	(1.12)	4.18	(1.10)	4.95	(.84)	5.08	(.76)
	Totals	4.06	(.96)	3.97	(1.07)	4.69	(.75)	4.89	(.79)

Moreover, a main effect was found of question type on the perceived informativity of the text and visual combinations, F [1,68] = 15.13, p< .001, η^2_p = .18. The answer presentations of procedural questions were evaluated as more informative than the answer presentations of definition questions.

Finally, an interaction was found between answer presentation format and question type, F [3,68] = 4.27, p< .01, η^2_p = .16. This interaction can be explained as follows: for both brief (F [1,17] = 17.12, p< .005, η^2_p = .50) and extended (F [1,17] = 7.31, p< .025, η^2_p = .30) answers with an *informative visual* significant differences were found between the two question types in the perceived informativity of the text and visual combination. Procedural answer presentations were more informative than definition answers presentations.

Attractiveness of the text and visual combinations

A main effect of answer presentation format was found on the perceived attractiveness of the text and visual combinations, F [3,68] = 4.64, p< .01, η^2_p = .17. Extended answers with an informative visual were evaluated as most attractive, while extended answers with an illustrative visual were evaluated as least attractive (see Table 2.8). Post-hoc tests revealed that only extended answers with an informative visual differed significantly from brief (p< .05) and extended (p< .025) answers with an illustrative visual.

Also, a main effect of question type was found on the perceived attractiveness of the text and visual combinations, F [1,68] = 20.59, p< .001, η^2_p = .23. The answer presentations of procedural questions were evaluated as more attractive than those of definition questions. Finally, no interaction was found between answer presentation format and question type (F<1).

Table 2.9

Mean difference scores of correctly answered questions in the post-test per question type and answer presentation format (Standard deviations in parenthesis).

	Text with an illustrative visual				Text with an informative visual			
	Brief		Extended		Brief		Extended	
Definition	.00	(2.14)	.06	(2.01)	.78	(1.52)	.22	(1.90)
Procedural	-.06	(1.21)	-.17	(2.23)	-.33	(.97)	.11	(2.22)
Totals	-.06	(2.78)	-.11	(3.64)	.44	(1.89)	.33	(3.63)

Number of correct answers in the post-test

Table 2.9 shows the mean difference scores of correctly answered questions in the post-test for the brief and the extended answers with an illustrative and an informative visual. The mean difference scores represent the number of correctly answered questions within answer presentations with an illustrative or informative visual minus the number of correctly answered questions within the purely textual answer presentations. The mean difference scores were used to quantify the added value of the visuals in the answer presentations.

First, consider the total mean difference scores between the four answer presentation formats. Table 2.9 reveals that the participants who received answer presentations with an illustrative visual answered fewer questions correctly than the participants who received purely textual answer presentations. However, the participants who received answer presentations with an informative visual answered more questions correctly than the participants who received purely textual answer presentations. Nonetheless, the total mean difference scores did not differ significantly between the four answer presentation formats (F<1) presumably because the differences are relatively small and the standard deviations are relatively high.

Table 2.9 also shows that in the case of definition questions, participants who received answer presentations with an illustrative visual did not differ from participants who received purely textual answer presentations in the number of correctly answered questions. However, participants who received answer presentations with an informative visual answered more definition questions correctly than those who received purely textual answer presentations. The mean difference scores for procedural questions showed that participants who received answer presentations with an illustrative visual answered fewer questions correctly than the participants who received purely textual answer presentations. This was also the case for participants who received brief textual answers with an informative visual. However, participants who received extended textual answers with an informative visual answered more procedural questions correctly than those who received extended textual answer presentations. However, no effect of answer presentation format was found (F< 1).

2.3.3 Conclusion

The results of the evaluation experiment showed that answer presentations with an informative visual were evaluated as more informative than answer presentations with an illustrative visual, especially for brief answers. Moreover, it was found that answer presentations of procedural questions with an informative visual were evaluated as more informative than those of definition questions. It also turned out that informative visuals were judged more attractive than illustrative visuals.

The results for the post-test suggested that learning from answer presentation with an informative visual leads to a better learning performance than learning from purely textual answer presentations. However, no significant differences were found between the multimodal and unimodal answer presentations in the mean difference scores of the number of correctly answered questions in the post-test.

2.4 Discussion

This chapter describes two experiments carried out in order to investigate the role of visuals that can be used for multimodal answer presentation in a medical question answering system.

In a production experiment, we investigated when and how people produce multimodal information presentations. The types of visual media that occurred in the corpus of collected answer presentations were diverse, i.e., there were photos with and without text, graphics with and without text, and animations. Moreover, significant differences were found in the distribution of these visual media related to their function. Photos not containing text often had a representational function: they visually represented the information mentioned in the text. For example, the question "What complications can occur when suffering from the measles?" was frequently illustrated with a child suffering from the measles. A relatively large proportion of decorational photos did contain text, but in these cases, the text was not used to inform (what one may expect text to do in visuals). Photos that contained text often had a representational function too. For example, the question "How many X chromosomes does a female body cell have?" was often illustrated with a photo of a woman's chromosome pattern in which text indicated the particular sex chromosomes. Graphics without text often had a representational function. For example, the question "How to apply a sling to the left arm?" was illustrated with four graphics illustrating the procedure. Graphics with text often had a representational but also an informative function. For example, the question "What happens at a tympanometry test" was frequently illustrated with a textual diagram illustrating the procedure. These types of graphics schematize the procedure by indicating the key elements. Thus, while graphics without text visually represent the information

mentioned in text, graphics with text represent information in such a way that they contain more information than mentioned in the text. Finally, animations often had an informative function because they present the information dynamically as opposed to photos and graphics.

The type of answer (brief vs. extended) was associated with different answer presentations. Visual media were more frequent in the extended answers. Also, the distribution of the functions of visual media was associated with different answer types. Within brief answers, most frequent were visual media with an high informative value whereas visual media with a low informative value were more frequent within extended answers. A likely explanation for this result could be that when the answer does not contain much text, it is likely that a visual easily contains additional information with regard to the text. When the answer contains much text, it is likely that a visual adds less information to it (i.e., it visually represents the information already present in text).

The type of question was also associated with different answer presentations. Visuals with a low informative value were most frequent in the answers of definition questions, whereas visuals with a high informative value were most frequent in the answers of procedural questions. A possible explanation for this result could be that the textual answers to definition questions (e.g., "How many molars does a human have?") often explained an element of the question, like 'molars', which was represented with a visual. Visuals in the answers of procedural questions were often used to explain the steps within the procedure and therefore added information to the textual answer.

Next, we investigated how people evaluate different types of answer presentations. The results of the evaluation experiment showed that answer presentations with an informative visual were indeed evaluated as more informative than those with an illustrative visual. Moreover, the type of question influenced participants' assessment of the informativity of text and visual combinations. *Procedural* answer presentations with informative visuals were evaluated as more informative than *definition* answer presentations with informative visuals. An explanation for this result could be that medical procedures -as they occurred in this experiment- lend themselves better to be visualized than definitions, because they have a dynamic and spatial character, whereas definitions more often concern abstract concepts that are

less easily visualized. For example, it is easier to find an informative visual for the procedural question "What happens at a tympanometry test?" than to find a visual for the definition question "What does ADHD stand for?"

Another interesting result is that while brief answers with an informative visual were evaluated as most informative, extended answers with an informative visual were evaluated as most attractive. The information load of the textual answers might explain these results. Brief and extended textual answers differ in their information density, i.e., brief answers contain less information than extended ones. Therefore, an informative visual adds more information to brief answers than to extended answers. The perceived informativity of the answer could therefore be influenced by the added value of the visual in the answer presentation, making brief answers more informative than extended answers. Arguably, an informative visual primarily enhances the attractiveness of extended answers as less information is added to the textual answer presentations.

The results of the post-test seemed to indicate that learning from answer presentations with an informative visual improved the learning results. However, no significant effect of answer presentation format was found, presumably because of the individual variation among participants' scores. A possible explanation for this result could be that there was a ceiling effect: on average the participants answered 13 of the 16 questions correctly.

In both experiments, a consistent result was found: participants preferred visuals having a high informative value to visuals having a low informative value. Moreover, we found that adding a visual to a textual answer is not enough when designing multimodal information presentations. The content of the information presentation (i.e., the type of question) also plays an important role. In both experiments, participants preferred visuals with a high informative value in procedural answer presentations and visuals with a low informative value in definition answer presentations.

The experiments described in this chapter raise various more detailed questions. For example, it would be interesting to investigate whether individual differences, like prior knowledge or learning preferences (i.e., verbal vs. visual) affect participants' assessment on the informativity and attractiveness of different unimodal and multimodal answer presentations. Also, the results of the production

experiment showed that the participants included dynamic visuals (i.e., film clip and animations) in their answer presentations. Therefore, it would be interesting to investigate whether static and dynamic visuals are evaluated differently (and under which circumstances) on their informativity and attractiveness. Moreover, in both experiments offline research methods were used to investigate the role of visuals in multimodal information presentation. The production and evaluation experiment have provided insights on how and when people produce information in a multimodal way. However, what is unclear is how multimodal information presentation is actually processed. Eye tracking could be a useful method to investigate how people process and integrate information from different modes and whether different types of multimodal information presentation are processed and integrated differently.

In this Chapter, we investigated when and how people present information in a multimodal way, and how other people evaluate such information. In the following chapters, we present three detailed case-studies looking into multimodal information presentation from different perspectives. Chapter 3 starts from finding answers in web sites and how users conceptualize their actions when navigating such multimodal information environments. Chapter 4 focuses on the effectiveness of three presentation modes (i.e., text, picture, or film clip) for learning and executing procedural instructions. Finally, Chapter 5 takes a closer look at the speech modality and discusses an eye tracking experiment studying the incremental processing of diphone synthesis, unit selection synthesis, and human speech.

Footnotes

1. Several taxonomies have been proposed to investigate the relations between text and visuals (see Marsh & White (2003) for an overview). Our own classification of functions of visual media corresponded highly with the one formulated by Carney & Levin (2002).

Appendix A: Medical questions presented in Experiment I and II

	Dutch	English
1.*	Waar wordt somatostatine geproduceerd?	Where is somatostatin produced?
2.*	Hoe kan ik mijn buikspieren versterken?	How can I strengthen my abdominal muscles?
3.	Wat is PTSS?	What is PTSS?
4.*	Hoeveel kiezen heeft een mens?	How many molars does a human have?
5.	Wat wordt er gedaan bij een myelografie?	What happens when a myelogram is taken?
6.	Hoeveel X-chromosomen bevat een lichaamscel van een vrouw?	How many X chromosomes does a female body cell have?
7.*	Wat zijn thrombolytica?	What are thrombolytic drugs?
8.	Hoe lang is de incubatietijd van dementia paralytica?	How long is the incubation period of dementia paralytica?
9.*	Waar wordt progesteron geproduceerd?	Where is progesterone produced?
10.	Welke factoren kunnen leiden tot een holvoet?	Which factors could lead to a cavus deformity?
11.*	Wat kun je doen als je een bloedneus hebt?	What should be done when having a nosebleed?
12.	Wat is leukopenie?	What is leukopenia?
13.*	Wat gebeurt er bij een tympanometrie?	What happens at a tympanometry test?
14.*	Hoe moet ik mijn werkplek inrichten om RSI te voorkomen?	How to organize a workspace in order to prevent RSI?
15.	Wat zijn de bijwerkingen van ibuprofen?	What are the side effects of ibuprofen?
16.	Wat doet insuline met de bloedsuikerspiegel?	What is the effect of insulin on the blood sugar?
17.*	Waar vindt de productie van testosteron plaats?	Where is testosterone produced?
18.*	Wat is een goede oefening om RSI in je handen te voorkomen?	What is a good exercise to prevent RSI in your hands?
19.	Wat helpt tegen jetlag?	What may help when having jetlag?
20.*	Wat gebeurt er bij een arthroscopie?	How is arthroscopy performed?
21.*	Welke vormen van colitis worden onderscheiden?	What types of colitis can be distinguished?

22.	Wat zijn de bijwerkingen van een DKTP-prik?	What are the side effects of a vaccination for diphtheria, whooping cough, tetanus, and polio?
23.	Hoeveel mensen lijden aan hoge bloeddruk?	How many people suffer from high blood pressure?
24.	Welke complicaties kunnen optreden bij mazelen?	What are the complications of measles?
25.	Waar worden rode bloedcellen aangemaakt?	Where are red blood cells produced?
26.*	Hoe leg je een mitella aan bij de linkerarm?	How to apply a sling to the left arm?
27.	Hoe wordt de ziekte van Von Willebrand bestreden?	How is Von Willebrand disease treated?
28.*	Hoe wordt een SPECT-scan gemaakt?	How is a SPECT scan made?
29.	Wat is een allergie?	What are allergies?
30.*	Waarvoor worden NSAID's gebruikt?	For what conditions are NSAIDs used?
31.	Waar kan ik een griepprik halen?	Where can I get a flu vaccination?
32.*	Waar staat ADHD voor?	What does ADHD stand for?

The medical questions presented in Experiment II are marked with a *

3 | Spatial conceptualization in multimodal information presentations
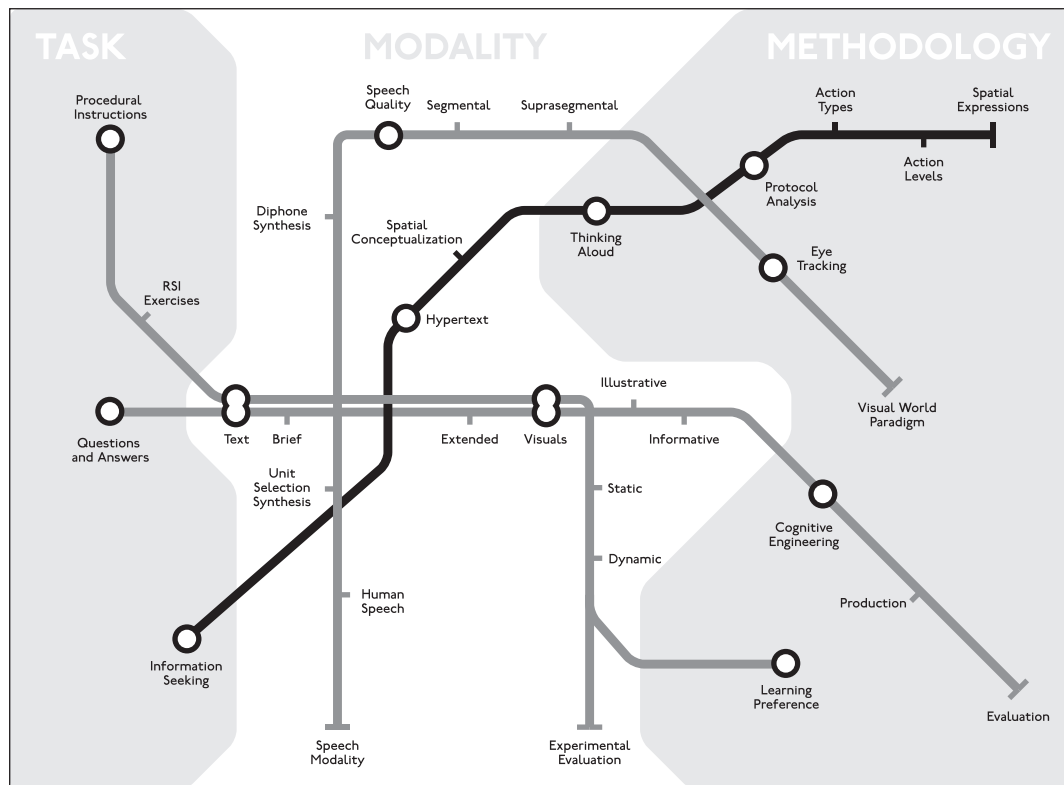
3.1 Introduction

In Chapter 2, we have seen that people make certain choices about positioning verbal and pictorial information in a multimodal information presentation, e.g., pictures were often placed below the text (see Figure 2.1 and Figure 2.3 in Chapter 2). Apparently, spatially positioning different types of information is important in multimodal information presentation. However, little research has been done on the notion of "space" in multimodal information presentations. Space can be considered as the minimal form of multimodality (Kostelnick & Roberts, 1998). For example, words are separated from each other by spaces. Also, in other forms of multimodality spatial information plays an important role, like in a hypertext.

A hypertext consists of a network of interlinked web pages in which users have to search for the information they need. However, hypertext users often cannot find the information they need. Therefore, several navigation aids (e.g., sitemaps and bread crumbs) have been developed which often present the hypertext's information structure spatially. This spatial character of navigation aids suggests that the concept of space is important for users who try to conceptualize a hypertext. Another indication that the notion of space is important when navigating in a hypertext are the large number of spatial metaphors used to talk about it, like "hyperspace" and "jumping from page to page". In order to help hypertext users to find the information they need, we first have to investigate when and how users represent a hypertext and whether or not they represent it spatially. This chapter sets out to explore this research question.

3.1.1 Effective navigation in hypertext: navigation maps

An important goal in usability research is to investigate which factors influence effectiveness and efficiency[1] in navigating hypertext. Studies focus for instance on individual differences in hypertext use (e.g., Campagnoni & Ehrlich, 1989; Vincente & Williges, 1988) and differences in users' tasks (e.g., Marchionini, 1989; Marchionini & Shneiderman, 1988). More recently, effects of the design of navigation aids were studied (e.g., Chen & Rada, 1996; Danielson, 2002; Dias & Sousa, 1997; Gupta & Gramopadhye, 1995; Maes et al., 2006; McDonald & Stevenson, 1999).

Hypertext users encounter at least two types of problems when they try to find information. The first problem is *disorientation* or *lostness*, embodied in the following three questions: Where am I? Where do I have to go next?, and How do I get there? (e.g., Edwards & Hardman, 1989; Elm & Woods, 1985; Otter & Johnson, 2000). The second is *cognitive overhead*, which refers to the amount of cognitive resources necessary to successfully complete a task in hypertext (Conklin, 1987). A large number of navigation aids have been developed to prevent users from getting lost and to reduce the cognitive overhead: hierarchical navigation bars (bread crumbs), paging buttons, alphabetical content lists, history tools, (expandable) menus, back buttons, spatial maps, etc. These tools typically present information structure in a schematic or spatial way. For example, a contents list presents the topics in a plain or indented list, based on theme or alphabet; bread crumbs show the depth of the information in a left-to-right order on the screen (Lida, Hull & Pilcher, 2003), and site maps offer many designs to present information structure spatially: top-bottom or left-right trees, spider structures, etc.

This spatial character of navigation aids suggests that the concept of space is important for users who try to conceptualize hypertext structure and tasks. Hypertext research indeed suggests that spatial site maps are beneficial to users. For example, McDonald & Stevenson (1999) concluded that users navigated more efficiently with a spatial map than with a contents list, although they did not find any differences in terms of effectiveness. Dee-Lucas & Larkin (1995) did find a difference in effectiveness: participants recalled more nodes in the spatial map condition than in the alphabetical index condition. These differences were smaller when participants were asked to read the content of the hypertext in order to summarize it, which suggests that spatial maps facilitate *finding* information rather than *studying* it.

Although the results of these and other experiments suggest that spatial navigation maps works better than other types of navigation instruments, like a contents list, there are reasons to question such a conclusion (Maes et al., 2006). First, the navigation maps as they are used in usability experiments differ substantially in the way they support the content, structure, task, and information access. They represent the hypertext's content and structure either globally or partially, they contain either 'labels only' or labels plus additional information, they either show the navigation history, or not. These differences themselves may result in substantial differences

in effect. Second, all maps that were studied contain semantic labels. The presence of these labels does not allow us to determine the exclusive contribution of the spatial element to the usability of the maps, as the labels (or other content oriented design variables) are likely to also affect usability. Third, even if a spatial map proves beneficial, it is not yet clear why, when and how it helps hypertext users. Does spatial design support the comprehension of the information structure? Does it enable users to set, monitor, and reach their goals more efficiently? Or does it mainly support the spatial-perceptual processes involved in hypertext use, like locating information, mentally replaying the navigating path, or transforming information into a spatial arrangement? These questions require more information about whether and how users conceptualize information environments and computer tasks spatially.

3.1.2 The role of space in conceptualizing hypertext and hypertext tasks

Space is one of the most powerful tools for humans to conceptualize abstract thought (e.g., Gentner & Boroditsky, 2001; Gibbs, 2005; Lakoff & Johnson, 1980; Tversky 2001, 2003). We 'translate' time into space (e.g., "We are entering a new age"), connect good things with *up* (e.g., "She is top!"), bad things with *down* (e.g., "I am feeling down"), important things with *near* (e.g., "Things you carry in your heart") and less important things with *far* (e.g., "A far-from-my bed show"), etc. It is therefore not surprising that hypertext also created a large number of spatial metaphors, like "lostness", "hyperspace", or "navigation". The pervasive conceptual force of space does not explain, however, when and how users represent information and tasks spatially.

The premise of the spatial metaphor is that navigation in hypertext is conceptualized and understood on the basis of navigation in a physical environment. But what does this mean exactly? Does it enable us to conclude that the distance of two web site pages is exactly 3,44 meters? Or does the spatial metaphor merely facilitate talking about hypertext and tasks in an intelligible way? Or is it something in between? In a lucid analysis, Boechler (2001) makes clear that space in hypertext can never be conceived of in purely literal or Euclidian terms. Navigating from one page to another is not literally going "deeper" in the web site, going to the homepage is only metaphorically "going up" and the distance between pages cannot

be expressed in metrical terms. This is not peculiar, as humans often conceive space in non-literal terms. But in her survey, Boechler makes clear that we have hardly any evidence on the working of spatial notions and metaphorical extensions in the minds of computer users.

There have been several attempts to apply spatial notions to hypertext use. For example, Shum (1990) applies the spatial notions *distance* and *direction* to hypertext, two elements which are known to be crucial in the study of how users conceive physical space, as it is clear from geography and psychology (e.g., Downs & Stea, 1973; Golledge, 1999; Taylor & Tversky, 1992a, 1992b; Tversky, 2003). The *whereness* of an object basically consists of a distance and a direction (Downs & Stea, 1977). According to Shum, each hypertext node has a certain *distance*, which can be quantified in absolute and relative terms, such as the number of nodes users have to visit, system response time, ease of returning to the previous node, or number of link traversals. *Direction* is defined as going forward and backward in the hypertext document. Although Shum tried to conceptually apply these definitions to hypertext, he did not investigate whether or not users really make use of these spatial concepts to mentally represent a hypertext environment. Similarly, Kim (1999) demonstrated the advantages of the familiar spatial metaphor of a shopping mall in accessing and using hypertext. Other researchers however contest the validity of the spatial metaphor. For example, Farris, Elgin and Jones (2002) concluded that the user's representation of a hypertext is non-spatial. They conducted an experiment in which participants had to explore a web site. The information on the web site was held constant, but the number of levels within the information structure varied. After exploring the web site, participants were asked to draw the web site's information structure. The analysis of these drawings indicated that the participants did not draw the spatial information structure of the web sites, but they drew conceptual relations between the information items instead. Therefore, Farris et al. (2002) concluded that the users' representation of a hypertext is non-spatial.

In sum, these studies seem to contradict each other at first glance. But this contradiction should be interpreted with great care, as the conclusions do not always seem to be reliable. Farris et al., for instance, offered their participants chunks with clear semantic relationships in a web site without any global spatial navigation aid, such as a site map, which makes it likely that participants are more guided by their

prefixed semantic knowledge than the somewhat ad hoc and unsupported division of the chunks in different information levels.

3.1.3 The investigation of spatial conceptualization in hypertext

Most usability studies draw conclusions about users' mental representations on the basis of performance results: the number of clicks, the recall of links or the quality of a drawing is assumed to reflect the adequacy of the representation. However, the relation between these dependent variables and the mental conceptualization of users is weak and always requires some type of subjective interpretation. Users' representations can also be investigated by other methods, like protocol analysis. In this research method, participants are asked to carry out a task, while verbalizing their thoughts. These verbalizations are written down in a verbal report and analyzed in a way that depends on the research question (Ericsson & Simon, 1993). Protocol analysis has been used in several research areas, like cognitive psychology (e.g., Newell & Simon 1972), reading comprehension (e.g., Presley & Afflerbach, 1995), and usability testing (e.g., Nielsen, 1993). Moreover, protocol analysis is a well-known tool for finding metaphorical language in interaction research (Kuhn, 1996; Maglio & Matlock, 2003). For example, Maglio and Matlock (2003) asked experienced and inexperienced hypertext users to verbalize what they do and think during their hypertext task. First, they asked users to execute free search tasks on the web. Afterwards, the participants were asked to tell what they just did. The transcripts of these interviews were coded to mark seven types of web actions. The results indicated that both novices and experts talked about their experiences in terms of physical motion and actions.

In this study, we elaborate on this elicitation method in an attempt to get a more fine-grained view of how users conceptualize their task and use spatial conceptualizations. Unlike Maglio and Matlock, we are not only interested in whether spatial metaphorical expressions are used, but also in the proportion of these expressions in their verbal production and in the relationship between spatial conceptualizations and the type of cognitive action of the user: do spatial conceptualizations mainly show up in verbalizing low level actions (such as clicking or typing), or also in planning and monitoring the task? Unlike Maglio and Matlock,

we ask users to verbalize their actions online, while they are executing their task. We realize that the resulting thinking aloud protocols do not directly tap cognitive processes. Furthermore, other drawbacks of this instrument may also apply here. For example, a user's conceptualisation of a hypertext may well be non-verbal, which would require a mental translation into a verbal form and thus additional cognitive processing. Yet, we consider protocol analysis to be a valuable tool for explorative work in this field, provided that the data are interpreted critically and carefully. Before describing the set up and the results of this exploration, we will discuss different ways of categorizing actions of users who are navigating in a hypertext.

3.1.4 Categorizing users' actions in hypertext

A generally accepted overall model of hypertext use is not readily available (Chen & Rada, 1996). How theoreticians model hypertext use depends on the type of computer task (e.g., solving open or closed information problems), and the perspective (e.g., a learning or usability perspective). Yet, several researchers have attempted to classify users' actions while navigating through hypertext. In this section, we will discuss some ways of classifying hypertext actions. These classifications should enable us to reliably determine action levels in thinking aloud protocols.
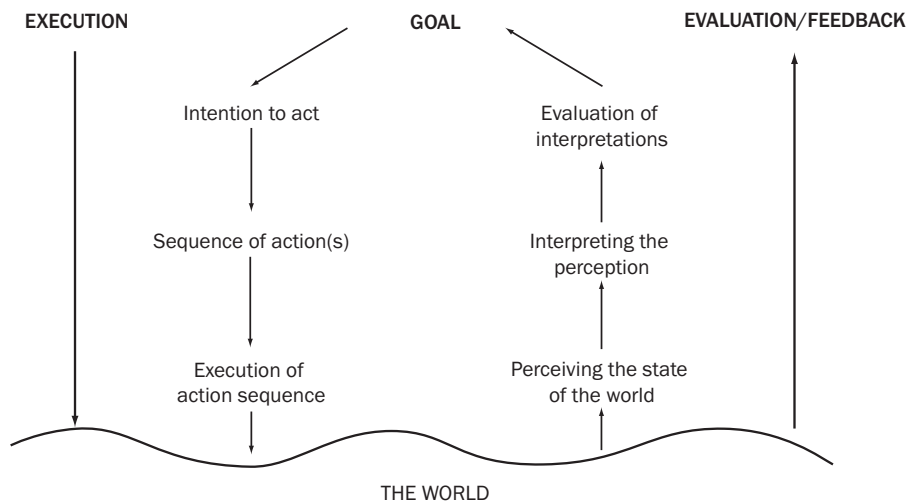
Figure 3.1

The Action Cycle (Norman, 1998, p. 47)

Using hypertext can be seen as an interaction between two actors: the user and the hypertext system. The user initiates an action, the computer responds, the user evaluates the computer's response, etc. That way, all users' actions can be categorized in either *executions* or *evaluations*. This distinction corresponds to a distinction in Norman's Action Cycle (1998) of human-computer interaction, see Figure 3.1. Users execute actions and evaluate the result by comparing the computer's reaction with their goal.
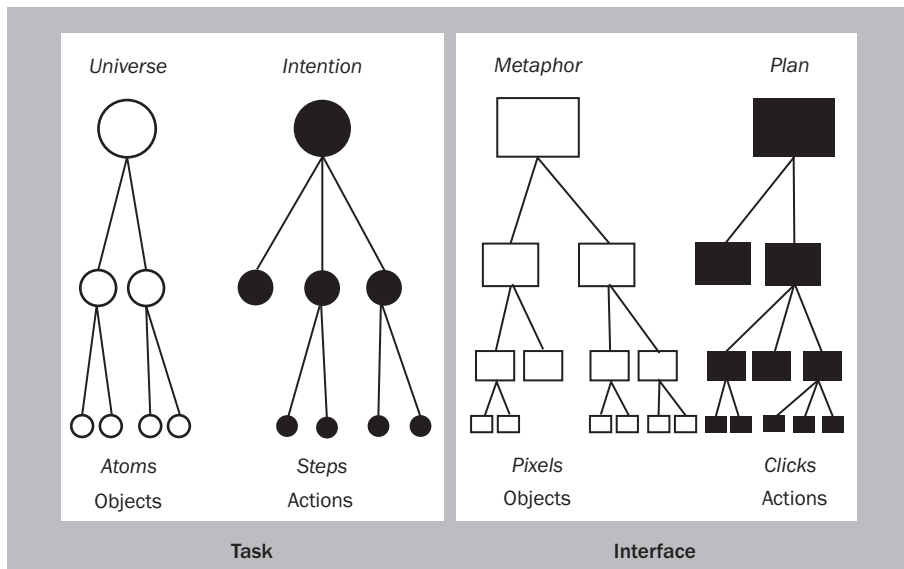
Figure 3.2

The OAI model illustrating the designer mapping the task (the real world universe of objects and intentions) to the interface (metaphors and plans). (Shneiderman, 1998, p.206)

More fine-grained models of hypertext use are based on the idea that users have to execute these cognitive actions on different levels, ranging from physical (e.g., pushing buttons, waiting) to conceptual (e.g., anticipating on new information behind a link, comparing computer response to their real world tasks). These different levels may be compared to three levels involved in language processing: readers are assumed to build a surface, a propositional and a mental representation when reading a text (e.g., Fletcher & Chrysler, 1990; Johnson-Laird, 1983; Kintsch &

Van Dijk, 1978). Hypertext users can be said to be mentally engaged in surface (i.e., executing physical, motional, perceptual actions), propositional (e.g., understanding the content and structure of hypertext) and mental/situational (e.g., planning and monitoring) actions. This analogy has also been used in other HCI models. For example, the Objects/Actions Interface (OAI) Model of Shneiderman (1998) follows a hierarchical decomposition of objects and actions in the task and interface domains, see Figure 3.2.

The task includes the world of real-world objects with which users work to accomplish their actions that they can apply on those objects. Both task objects and task actions can be decomposed into smaller units. For example, a high-level task object can be a letter which can be decomposed into paragraphs which in their turn consist of characters. Task actions start from high-level intentions which can be decomposed into individual steps. For example, the intention of writing a letter can be decomposed into knowing the addressee, knowing where to find the address of the addressee, and finally writing down the address. Also, the interface includes hierarchies of objects and actions. For example, some interface objects deal with storage. Users learn that a computer stores information and this information can be stored in directories. In turn, a directory consists of a set of files which in their turn consist of a set of characters. Interface actions can also be decomposed from high to low level actions. For example, an action on the highest level could be the plan to create a text file. This plan can be decomposed into lower action levels, such as creating a file, inserting text, and saving that file. But, also the action of saving a text file can be decomposed into lower action levels, like choosing the name of the file.

Finally, Dillon (2004) has developed the TIME framework of hypertext use consisting of four interactive levels, i.e., Task, Information model, Manipulations skills, and visual Ergonomics, see Figure 3.3. Dillon's Task level implies the users' goal in the real world. The Information level refers to the user's representation of the hypertext structure and content. Manipulation and ergonomics level refer to motional and perceptual activities.
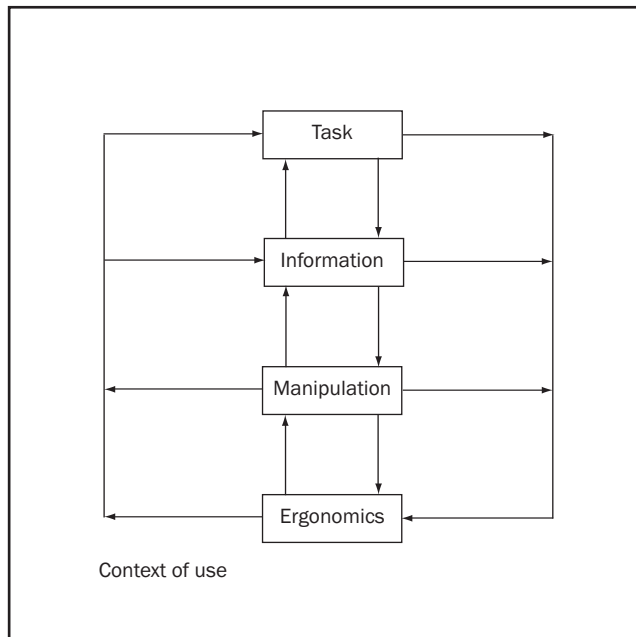
Figure 3.3

The TIME framework (Dillon 2004, p.140)

In our explorative study, we chose to depart from Norman's Action Cycle as this model represents users' actions into executions or evaluations. Moreover, Norman's Action Cycle reflects different levels of execution and evaluation. Executions start with higher-level actions (i.e., intending actions) and result in the low level actions (i.e., executing sequences of actions). Evaluations on the other hand start with low level actions (i.e., perceiving the state of the world) and result in higher-level actions (i.e., evaluating interpretations). In our analysis, we will distinguish two types of actions, i.e., executions and evaluations consisting of three levels of actions, i.e., first level, second level, and third level, see Figure 3.4.

Executions can be described in three action levels. For example, a user's goal could be writing a letter. In order to achieve this goal, the user has to formulate several intentions to act which corresponds to our third action level. An example of an execution on the third action level could be the intention to use the computer to

write a letter. Subsequently, the user has to translate this intention into a sequence of actions, which corresponds to our second action level. An example of an execution on the second action level could be turning on the computer and starting a computer program. Finally, the user executes the sequence of actions by pushing the ON button of the computer and by clicking on the WORD icon. These actions correspond to our first action level. Also, evaluations can be described in three action levels. For example, the user perceives a new window on the computer screen, after clicking on the WORD icon. This perception corresponds to our first action level. Next, the user interprets this perception, which corresponds to our second action level. For example, the user interprets the appearance of the new window as the start up of WORD. Finally, this interpretation has to be evaluated, e.g., the users evaluates that WORD can be used to write letters. This evaluation corresponds to our third action level. The actions types and action levels we distinguish do not intend to make direct claims about the representations involved in using hypertext, but they should serve as a useful analytical tool for analyzing actions involved in hypertext use as they are verbalized.
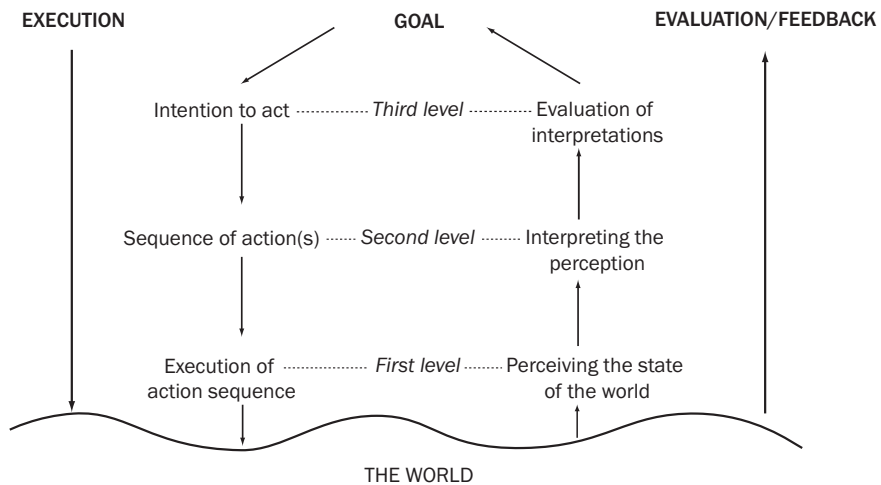
Figure 3.4

The Action Cycle consisting of two action types (execution and evaluation) and three action levels (first, second, and third).

3.2 Research method

We conducted an explorative thinking aloud study to investigate which actions types (i.e., executions vs. evaluations) and which actions levels (i.e., first level vs. second level vs. third level) are expressed in spatial terms.

3.2.1 Materials

We collected ten thinking aloud protocols in two different usability studies. One study was set up to investigate the usability of a web site about the European Commission, the other study to investigate the usability of a medical web site. The web sites in the two studies were conventional web sites with many textual hyperlinks and several standard search facilities (a sitemap and a search function), see Figure 3.5. In both studies, users were asked to perform simple search tasks in a hypertext (i.e., looking up the answers to factual questions), and to think aloud while executing these tasks. Participants' actions and verbalizations were recorded with Camtasia[2] camcorder software.

The Medical web site study
Seven participants (four women and three men, between 27 and 57 years of age) took part in this study. Together they produced five protocols of three different types:

- One individual thinking aloud protocol, in which a novice user was asked to perform simple search tasks in a medical web site while thinking aloud
- One individual thinking aloud protocol, in which an expert user was asked to perform simple search tasks in a medical website while thinking aloud.
- One co-discovery protocol (Dumas & Redish, 1993), in which two novice users worked together in performing simple search tasks in a medical website while thinking aloud.
- One co-discovery protocol (Dumas & Redish, 1993), in which two expert users worked together in performing simple search tasks in a medical website while thinking aloud.
- One instructing protocol, in which an expert user was asked to instruct an assumed (non-present) novice to find the answer on simple search tasks in a medical web site.

In all cases, the experimenter kept silent. She only reminded participants to keep thinking aloud by saying "keep talking" after a period of silence (Ericsson & Simon, 1993; Krahmer & Ummelen, 2004).
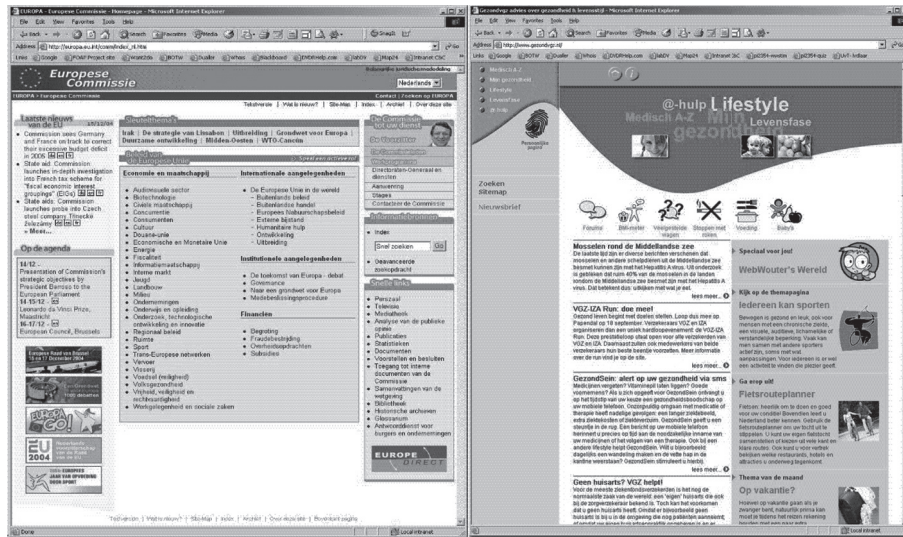
Figure 3.5

Screenshots from the homepages of the European Commission web site (left) and the medical web site (right)

As a first step, the participants were asked to solve a digital version of the tower of Hanoi puzzle while thinking aloud. This common practice task was used to familiarize the participants with the thinking aloud method. Then they were presented with a Dutch medical web site[3]. The participants were asked to answer fact-finding questions such as "What is the meaning of the word melatonin?" and "Which vaccinations do you need when travelling to Swaziland in Africa?" While executing these tasks they were asked to verbalize their thoughts. The experimenter kept silent. She only reminded participants to keep thinking aloud by saying "keep talking" after a period of silence (Ericsson & Simon, 1993; Krahmer & Ummelen, 2004). The participants were allowed to take as much time as they needed to complete a task. If they could not find the answer to a question they were allowed to move on to the next task. After finishing a particular task, participants were instructed to go to the homepage of the web site.

The European Commission study

Five participants (four women and one man, between 20 and 25 years of age) took part in this study. Each participant produced a thinking aloud protocol. First, they received the same practice task as the participants in the medical web site study (i.e., a digital version of the tower of Hanoi puzzle). Next, they were presented with the homepage of the Dutch web site of the European Commission[4]. The participants had to find the answers of six fact-finding tasks, like "Who is the current Dutch commissioner in the European Commission?" The procedure was the same as in the medical web site study, but only individual thinking aloud protocols were collected here.

3.2.2   Coding system

Each utterance was coded on the following variables: spatial verbalization, action type (i.e., executions and evaluations), and action level (i.e., first level, second level, and third level). In the following subsections, we will describe our criteria for coding the protocols.

Spatial or non-spatial verbalizations

We distinguished three types of spatial expressions:

- Verbalizations describing the user's next action as moving to or arriving at another place, by using expressions, like *gaan naar*, go to; *komen bij*, arriving at; *zoeken bij*, search at; *kijken bij*[5], look at, e.g.:

    "I am going back to the homepage"

- Verbalizations describing the user's location as being at a particular place, by using expressions, like *zijn in/bij/terug*, be in/at/back; *zitten in/bij*, sit in/at, e.g.:

    "I am in the main menu"

- Verbalizations describing information as being somewhere in a physical location, e.g.:

    "There is more information behind this hyperlink"

Verbalizations of action types: executions and evaluations

We distinguished two action types: execution and evaluation. **Executions** are defined as:

- Verbalizations of an action the user performs, e.g.:

  "I am clicking on GO"
- Verbalizations of the user's intention to act, e.g.:

  "I will go back to this item"

Linguistic characteristics that indicate this type of utterances are verbs reflecting actions, such as click, go, scroll, read, type, or move.

Evaluations are defined as:

- Verbalizations of a user's perception of elements in the environment, e.g.:

  "A pop-up appears"
- Verbalizations of an evaluation of a user's action, e.g.:

  "I cannot click on this item"
- Verbalizations of a user's evaluation of his task, e.g.:

  "I think I have found the answer"
- Verbalizations of a user's speculation on where information could be found, e.g.:

  "Maybe at the hyperlink called organization"

Verbalizations of action levels: first, second, and third level

We defined three action levels. Utterances at the **first level** are users' verbalizations of technical actions and perceptions of elements in the hypertext environment. We distinguished the following four types of first level verbalizations:

- Verbalizations of a user's perception of hypertext elements on the computer screen, e.g.:

  "I see three hyperlinks"
- Verbalizations of a user's coordination of actions with mouse and keyboard, e.g.:

  "I am double clicking on this object"
- Verbalizations of user's assumption on or question about technical aspects of the hypertext, e.g.:

  "Is this element clickable?"
- Verbalizations of user's technical actions, e.g.:

  "I type in the word movement"

Utterances at the **second level** concern a user's understanding of the meaning of the hypertext's content. We distinguished two types of verbalizations at this level:

- Verbalizations concerning the comprehension of the content on the screen, e.g.:

    "This is an interview about the books he likes"

- Verbalizations concerning inferences made during reading and interpreting, e.g.:

    "This seems to be about the nations who are united in the European Union"

Utterances at the **third level** concern users' reflections on their real-world goals. We distinguished four types of verbalizations on this level:

- Verbalizations reflecting questions about or relations with the search task, e.g.:

    "What is the name of the book I am looking for?"

- Verbalizations reflecting and evaluating screen results in terms of the search task goal, e.g.:

    "I think I have found the answer"

- Verbalizations reviewing the searching process, e.g.:

    "Maybe if I search on a new version of Publication Magazine, I will find the answer"

- Verbalizations of users' strategies concerning the search task, e.g.:

    "I am going to search on seats"

Not coded

Utterances that were not related to the task or took the form of fillers were not coded, e.g.:

    "Ehh"; "Well"; "Wait".

These utterances were left out of the analyses. Of a total of 694 utterances, 116 items were not related to the task (17%).

3.2.3   Coding procedure

Randomly chosen parts of the ten protocols were coded with the program MAXQDA[6]. One analyst coded the two Medical Web site fact-finding tasks mentioned above and eight randomly chosen tasks of the five verbal protocols of the European Commission study. The total corpus consisted of 694 coded segments.

To determine the reliability of the analysis, a second analyst independently coded parts of the corpus on the basis of the same coding scheme that was defined first (see section 3.2.2). Differences between the two analysts were discussed, which resulted in some adjustments of the coding system. This procedure took place two times. The second analyst coded 128 utterances during the final analysis. Following standard practice, we qualify Cohen's κ as adequate if its value was higher than .70 (Van Wijk, 2000). The results indicated that both analysts highly corresponded in judging the utterances as executions or evaluations (Cohen's κ = .80; n = 128). The two analysts also highly corresponded in judging the utterances as spatial or non-spatial (Cohen's κ = .87; n = 128). Finally, both analysts corresponded in judging the utterances as first level, second level, or third level (Cohen's κ = .78; n = 100[7]).

3.3   Results

3.3.1   Overall results

Table 3.1 shows frequencies of action types, action levels, and spatial verbalizations in the complete set of coded utterances. The table shows that overall, evaluations occurred more frequently than executions: χ^2 (1) = 39.91, p< .001, that the first action level occurred more frequently than the second level and third level levels: χ^2 (3) = 369,34, p< .001, and that non-spatial utterances occurred more frequently than spatial utterances: χ^2 (1) = 249.83, p< .001.

Not all utterances could be coded unambiguously in one of the three action levels. In a number of cases a segment could be interpreted as first level or second level. Utterances such as: "I am reading the headings", can be interpreted as second level because it refers to the second level comprehending of the information on the

screen. It can also be interpreted as first level, if it is expressing the technical, low-level activity of reading from the screen. Given the relatively large number of these cases, we included them as a separate category (first level/second level).

Table 3.1

Percentages of the occurrence of action types, action levels, and spatial verbalizations in 578 coded utterances from 10 verbal protocols (n = 578)

Action type	Execution	36.9
	Evaluation	63.1
Action level	First level	56.9
	First level/second level	11.9
	Second level	4.8
	Third level	26.3
Spatial verbalization	Spatial	17.1
	Non-spatial	82.9

3.3.2 Spatial verbalizations related to action type and action level

Is there a reliable relation between spatial expressions and action level or action type? Before discussing this main question, we will go into some effects of user and task characteristics that were intentionally or unintentionally varied in this study. A multinomial logistic regression analysis was performed with participant, experience of the participant (novice vs. expert) and the participant's role (thinking aloud, co-discovery, and instructor) as independent variables and spatial verbalizations as dependent variable. The analysis showed significant effects for the user characteristics: speaker, χ^2 (11) = 24.59, p< .05, experience of the user, χ^2 (1) = 4.23, p<. 05, and participant's role, χ^2 (2) = 7.08, p< .05. Experts tend to use more spatial expressions than novices (19% versus 11%), and co-discovery participants tend to use fewer spatial expressions (7%) than both thinking aloud participants (19%) and instructors (16%). Furthermore, individual users appear to differ somewhat in their tendency to use spatial expressions. The average percentage of spatial expressions per speaker is 17%, and some individual speakers use fewer or very occasionally more spatial expressions. Moreover, we checked whether the web sites (European Commission

web site vs. medical web site) had an effect on the amount of spatial verbalizations, and this turned out not to be the case, $\chi^2 (1) = 3.51$, p = .61. The following subsections does not go further into the reasons for these individual differences, but departs from the set of utterances that was collected and coded and tries to relate the occurrence of spatial expressions to action types and action levels.

Action type

Table 3.2 shows the frequencies of spatial verbalizations within executions or evaluations. Spatial verbalizations were most frequent when users were verbalizing executions, $\chi^2 (1) = 34.10$, p< .001.

Table 3.2

Spatial verbalizations related to executions and evaluations (Scores are percentages of utterances; n = 578)

	Executions (n = 213)	Evaluations (n = 365)
Spatial verbalizations (n = 99)	29.1	10.1
Non-spatial verbalizations (n = 479)	70.9	89.9

Action level

Table 3.3 shows the frequencies of spatial verbalizations at different action levels. Spatial verbalizations were found more frequently on the first level action level than on the other levels, $\chi^2 (3) = 25.98$, p< .001. Table 3.3 entails an intermediate level as well (first level/second level), containing the cases where the analysts disagreed between a first level and second level classification (see section 3.3.1).

Table 3.3

Percentages of spatial verbalizations related to the action level (n = 578)

	First level (n = 329)	First / second level (n = 69)	Second level (n = 28)	Third level (n = 152)
Spatial verbalizations (n = 99)	22.2	0.0	0.0	17.1
Non-spatial verbalizations (n = 479)	77.8	100.0	100.0	82.9

3.3.3 Spatial verbalizations related to other performance data

In additional analyses we looked for relations between spatial verbalizations and other performance data, such as the type of search task and the correctness of the task outcomes.

In order to see whether or not the number of spatial verbalizations depends on the type of search task, we divided the search tasks in both web sites in subtasks and tested whether the proportion of spatial verbalizations was related to the specific subtask. In both web sites the spatial verbalizations differed depending on the subtask: European Commission web site, χ^2 (5) = 13.42, p< .025, medical web site, χ^2 (2) = 19.60, p< .001. In the European Commission web site, most spatial verbalizations occurred during a subtask that required participants to search for the chairman's name of the EU. In this instance, subjects typically had trouble finding the answer and had to search a large part of the site. In the medical web site, most spatial verbalizations occurred in the task: "Go back to the homepage" in which participants were instructed to return to the homepage of the web site. This was a content-free task and related to the web site's information structure. Finally, the amount of spatial verbalizations was not related to completing the task in a successful way, χ^2 (1) = 1.43, p = .23.

3.4 Discussion

The exploration executed in this study served different purposes, which all merit to be discussed shortly. The main purpose was to investigate whether and how hypertext users spatially conceptualize cognitive actions they are involved in. Second, we wanted to discover whether thinking aloud protocols are a suitable method to shed light on the types of cognitive actions at work while using hypertext. Furthermore, we wanted to know whether the action types (i.e., executions vs. evaluations) and action levels (i.e., first level vs. second level vs. third level) we distinguished can be regarded as a suitable mould for the classification of cognitive actions of hypertext users. Finally, we wanted to find legitimization for the widespread design decision to represent digital information spaces as spatial constructs, i.e. sitemaps, instead of verbal summaries.

The exploration suggests a clear-cut result: users predominantly use spatial expressions to conceptualize executions and first level actions. Arguably, first level actions are straightforward and therefore more directly related to perceptual space than higher order actions. Also, executions are more goal-oriented than evaluations, and therefore more suited to be conceptualized by the well-known GOAL AS DIRECTION metaphor (i.e., concrete or abstract motion toward a goal, like "picking up the telephone" or "working to get a promotion", Maglio & Matlock, 2003). Still, it is strange that there are so little spatial conceptualizations on the second level and third level, although space is perfectly suited for conceptualizing information structures or plans and goals of users. It should be noted that not only the number of spatial second level and third level expressions is low, but also the overall proportion of second level and third level (as opposed to first level) expressions. Apparently, hypertext users are much more involved in shallow cognitive tasks (clicking, typing, reading, etc.) than in deep processing (understanding content and structure, monitoring plans etc.). This is too premature a conclusion, however. The uneven distribution of first level, second level and third level expressions may be caused by the online character of the thinking aloud method. Thinking aloud users have to conceptualize their thoughts immediately on the fly, which may incite them to verbalize the here and now of each and every screen, instead of stepping back and talk about global structure or task progress.

For the low number of second level (spatial) expressions, there may be an additional explanation, i.e., the narrow definition of second level (as opposed to first level) expressions. At the outset we decided to only code segments as second level when the understanding of the content of elements was verbalized. But much more expressions can be said to entail an awareness of a larger information structure on the part of the user. For example, when the user verbalizes the first level action "I am going back to the link on commissioners", there is a clear awareness of some structural organization in the hypertext, which may be seen as a second level attribute. Another reason for the low number of second level expressions may be the type of tasks used. Maybe their nature was too low-level (fact finding), which may have resulted in many low-level expressions.

In sum, it is unclear to what extent the online character of the thinking aloud method overstressed the attention for low-level actions. Furthermore, the analysis

showed that one expression can express different levels of cognitive actions, which again can be seen as a shortcoming of this method to answer the questions we were interested in. This explorative analysis showed that users indeed use spatial expressions to talk about their task, and this is the outcome of the thinking aloud method. Aside from its drawbacks, this method can be used to further our understanding of the use of space in hypertext in order to determine and explain the beneficial nature of spatially organized navigation help.

In this chapter, we presented an exploratory case study looking into multimodal information presentation from web site usability. In the next chapter, the second case study is discussed in which we look into multimodal information presentation from the perspective of cognitive and instructional psychology.

Footnotes

1   The difference between effectiveness and efficiency is that effectiveness measures are based on users' search accuracy as well as users' recall and understanding of the structure of the hypertext, whereas efficiency measures are based on speed and the number of steps taken to complete an information search.

2   http://www.techsmith.com/products/studio/default.asp

3   http://www.medicinfo.nl

4   http://europa.eu.int/comm/index_nl.htm

5   Unlike in English, the Dutch preposition bij has a clear locative interpretation.

6   http://www.maxqda.com/

7   In the final analysis, the analysts disagreed on cases which could be classified as first level or second level. These utterances were classified in a separate category, which may account for the high Kappa-score (see Results and Discussion).

4 | Modalities for procedural instructions
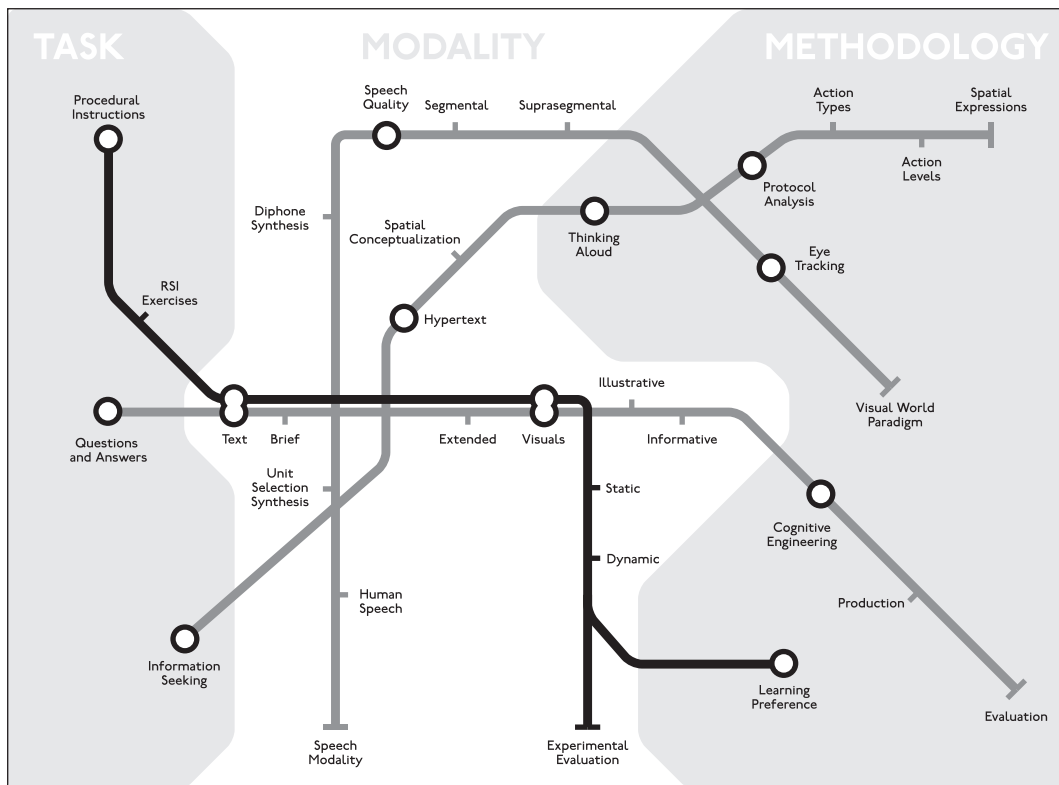
This chapter is based on Van Hooijdonk, C.M.J., & E.J. Krahmer (in press). Information modalities for procedural instructions: The influence of text, pictures, and film clips on learning and executing RSI exercises. *IEEE Transactions on Professional Communication* and partly based on Van Hooijdonk, C.M.J., & E.J. Krahmer (2006). De invloed van unimodale en multimodale instructies op de effectiviteit van RSI-preventieoefeningen [The influence of unimodal and multimodal instructions on the effectiveness of RSI prevention exercises]. *Tijdschrift voor Taalbeheersing, 28*(2), 73-87.

4.1 Introduction

In Chapter 2, we investigated whether people use and prefer multimodal information to express different types of information. In this chapter, we focus on the effectiveness of different information modalities (i.e., text vs. pictures vs. film clips) expressing a specific type of discourse, namely procedural instructions.

4.1.1 The effectiveness of different information modalities

The emergence of computer-based learning has led to new possibilities for presenting instructions and to a renewed interest in the effects of different information modalities. Instructions can be given in plain text, but also in the form of static visuals (e.g., diagrams, pictures) or in dynamic visuals (e.g., animations, film clips). Naturally, this raises the question which (combinations of) modalities are best for which learning task. This question has been addressed in a large number of recent studies (e.g., Lewalter, 2003; Lowe, 2004; Mayer 2001, 2003; Michas & Berry, 2000; Ploetzner & Lowe, 2004, among many others).

Each of the aforementioned presentation modes has its own unique characteristics, and its own advantages and disadvantages from an instructional perspective. Language is the basic form of human communication, and one of its main strengths is that it is expressive and explicit (both for concrete and for abstract subject matter). An additional advantage of its written form (as opposed to the spoken variant) is that it is not transient: written sentences remain visible on paper and can be re-read if desired, while spoken sentences are "gone" once they have been uttered. But, reading a text requires considerable skill and effort, moreover, text primarily facilitates linear information processing.

Text and pictures differ in the type of information that they can convey, due to the nature of their symbol system. The symbol system used for text is often abstract/linguistic, meaning that the relation between a word and its referent is arbitrary and symbolic (e.g., the word "cat" does not resemble the actual animal). The symbol system used for pictures is sensory, meaning that the relation between a picture and its referent is often analogous (e.g., a picture of a "cat" resembles the actual animal). This difference in the level of analogy between the presentation format and referent,

also called "articulatory distance" (Williams & Harkus, 1998), affects the type of information text and pictures can convey. For example, pictures can communicate perceptual information directly (e.g., spatial orientation and location). Moreover, pictures are not constrained by the linear structure of text, and are therefore argued to be more efficient at representing nonlinear relations among objects. In text these relationships often remain implicit (Larkin & Simon, 1987).

The focus of many recent studies has been on the instruction effects of dynamic visuals, presumably because such instructions only recently have become a real possibility due to the increased computing power of standard multimedia PCs. A number of reasons have been suggested to expect an advantage of dynamic visuals over other presentation possibilities, such as text and static visuals. For instance, it has been argued that dynamic visuals are beneficial for learning since they offer a "complete model" of a learning task (e.g., Lewalter, 2003; Park & Hopkins, 1993). In other words, they are "informationally complete" Schnotz et al., 1999: p.249). When static visuals or text is used, learners themselves will have to construct a mental representation of the temporal aspects in the instruction. It has been argued that dynamic illustrations offer a better representation of these temporal aspects, in addition reducing the level of abstraction, and supporting a deeper level of understanding than static visuals would do (Park & Hopkins, 1993). Arguably, this should facilitate learning, since it would reduce learning times, would require less practice, and would result in fewer errors.

A substantial number of studies have tried to demonstrate this presumed learning benefit, however with mixed results (e.g., Bétrancourt & Tversky, 2000; Lewalter, 2003; Michas & Berry, 2000; Tversky et al., 2002). An explanation for these mixed results is that dynamic visuals have a fixed duration, which viewers simply have to watch. This may lead to inherently longer learning times (Tversky et al., 2002). In a similar vein, it has been suggested that dynamic visuals may take more time to process than other presentation modes; the information in dynamic visuals changes continuously, and as a result learners could be overwhelmed with the amount of information (Ainsworth & VanLabeke, 2004; Lewalter, 2003). Alternatively, it has been suggested that dynamic visuals are processed somewhat superficially; they require little cognitive effort, as they can be watched rather passively (e.g., Schnotz et al., 1999; Schnotz & Rasch, 2005). Moreover, methodological problems in the

various experimental studies could also contribute to the mixed findings found for presumed leaning benefit of dynamic visuals. For instance, to be really beneficial dynamics should have some added value (e.g., Weiss, Knowlton & Morrison, 2002), which is not always the case. When it comes to temporal sequencing or indicating direction of motion, it has been claimed that arrows in static visuals may be just as effective as dynamic visuals (e.g., Tversky, Zacks, Lee & Heiser, 2000). Tversky et al. (2002) point out that in some comparative studies there is a lack of equivalence between dynamic and static visuals in content or procedures, for instance because the dynamic visuals convey more information or involve interactivity, which is absent in the "static" conditions. Some researchers have argued that even when dynamic visuals do not lead to improved learning, they are more attractive and motivating than other instruction forms, and should be preferred for that reason (e.g., Perez & White, 1985; Rieber, 1991; Tversky et al., 2002). However, this "subjective satisfaction" (Nielsen, 1993) of various instruction modes is often not addressed, but when it is the results are equivocally positive for the dynamic instructions (e.g., Pane et al., 1996).

A factor that might also have an influence on the relative benefits of information modalities is the type of task (Weiss et al., 2002). It seems reasonable to assume that different task types benefit from different kinds of instructions (see also Hegarty, 2004). In many studies, dynamic visuals are used as learning instructions for descriptive, "scientific" explanations (often describing causes and effects), for instance, of mechanical and electronic systems (e.g., brakes, pumps, generators: Mayer, 1989; Mayer & Gallini, 1990; electronic circuit systems: Park & Gittelman, 1995), mathematical concepts (e.g., algebra: Reed, 1985), or complex natural phenomena (e.g., electricity: Cheng, 2002; lightning: Mayer & Moreno, 2002; gravitational lensing: Lewalter, 2003; meteorological changes: Lowe, 2004). Arguably, some of these task types lend themselves better for dynamic visualization than others. Moreover, what holds for descriptive tasks (such as those above), may not apply to procedural ones, such as bandaging a hand (Michas & Berry, 2000), folding origami models (Caroll & Wiebe, 2004), operating a control panel (Boekelder & Steehouder, 1998), or tying nautical knots (Schwan & Riempp, 2004). These procedural tasks differ from descriptive ones in a number of respects. Not only is the nature of the task different (procedures are inherently more linear, one step following another), but also the learning goal is different (the focus is not only on understanding, but also

on acquiring certain capabilities or skills). One of the factors that might affect the processing of procedural information is the design format (Ganier, 2004). Procedural information is often presented in a text and / or picture format (e.g., route directions, maintenance instructions, assembly instructions). Although pictures may help users to form a mental model of the procedure (e.g., Schnotz, Picard & Hron, 1993; Winn, 1989), the combination of text and picture may not always be helpful. For example, the users' attention has to switch between the information presented in text and picture leading to the so-called split attention effect (Sweller & Chandler, 1994). Therefore, document designers have to understand the strengths and limitations of both text and pictures when designing procedural information (Williams & Harkus, 1998). Both the type of picture (e.g., overview vs. partial view: Gelleij, Van Der Meij, De Jong & Pieters, 1999; line drawing vs. photo: Michas & Berry, 20000; object-centered vs. body-centered: Krull, D'Souza, Roy & Sharp, 2004) as well as the type of textual instruction (e.g., user-centered vs. object-centered vs. environment-centered: Maes & Lenting, 1999) may influence users' performance of the procedure.

To further complicate the situation it may well be that besides variation between learning domains there is also variation within learning domains. Arguably, some learning tasks in a given domain are easier than others, and this may have an influence on the relative contribution of various instruction formats for those tasks. For example, in a series of experiments Marcus et al. (1996) systematically varied the complexity of a specific type of procedural instruction (i.e., connections of electrical resistors) and their presentation format (i.e., text vs. diagram). In these experiment, participants had to follow instructions on how to connect electrical resistors (i.e., single-series connections, multiple-series connections, and parallel connections). These instructions differed in the number of elements that needed to be considered in order to solve the connection problem. The instructions were either presented in a text or in a diagram. The results of the experiments showed that the participants needed more time to solve a "difficult" connection problem in which more elements needed to be considered than to solve an "easy" connection problem in which less elements needed to be considered. Moreover it was found that when the instructions were presented in a textual format the participants needed more time to solve the connection problem than when the instructions were presented in a diagrammatic format.

4.1.2 Expectations concerning the effectiveness of information modalities

In this study, the effects of task difficulty and information modality (comparing dynamic visuals with static visuals and text) are reported on learning a specific class of procedural tasks, namely exercises aiming at the prevention of Repetitive Strain Injury (RSI). RSI is a general term for disorders that are caused by repetitious use of hands, arms, and shoulders, often as a result of prolonged computer terminal work (e.g., Stone, 1983). It is generally assumed that taking regular breaks during computer work in combination with exercises is beneficial for the prevention of RSI (e.g., Balci & Aghazadeh, 2003; McLean et al., 2001; Williams et al., 1989). RSI exercises offer a new and understudied learning domain, with various interesting properties. Generally, these exercises involve little or no abstraction, do not consist of a complicated sequence of actions, and are relatively short. Moreover, the exercises are highly recognizable (almost everybody has two hands). It is interesting to observe that current RSI information web sites offer a large variety of prevention exercises, in many different presentation formats (see Figure 4.1 for a representative selection) which raises the natural question what the effectiveness of the various presentation formats is. Note also that there is substantial variation in the difficulty level of existing RSI exercises, so that the other factor of interest (variation in task difficulty) can be modelled in a fairly straightforward way in this domain.

The potential influence of both presentation modality and task difficulty on learning performance can be motivated from cognitive load theory (e.g., Marcus et al., 1996; Sweller & Chandler, 1991; Sweller, Van Merriënboer & Paas, 1998), a theory which aims to develop ptimal instructional designs while considering the limitations of the human mind. Cognitive load theory builds on the broadly accepted assumption that the mind combines a short term (or working) memory of very limited capacity (where all conscious activity and processing of information occurs; Baddely, 1992; Miller, 1956) and a long term memory with a virtually unlimited capacity (e.g., Sweller et al., 1998). According to cognitive load theory, learning amounts to the construction of new (or modification of existing) schemata (Chi, Glaser & Rees, 1982), which are subsequently stored in long term memory. Since the capacity of working memory is severely limited, the cognitive load of learners should be kept at a minimum during learning. In the current version of the theory (Sweller et al.,

1998; Sweller, 1999), three kinds of load are distinguished: intrinsic load, caused by the structure and intrinsic nature of the learning task, extraneous load, caused by the manner of presentation and its influence on working memory, and germane load, caused by the learners' effort to process and comprehend learning material. The sum of these three kinds of load should not exceed working memory capacity, in order to avoid cognitive overload. As argued above, the intrinsic load may vary both between and within task domains. Since the intrinsic load of a particular learning task is fixed, instruction design can only influence the extraneous and the germane load, and, obviously, when the intrinsic load of a particular task is high, there is relatively less room for extraneous and germane load. Germane load is a "positive" form of load, since it might direct learners' attention to processes that are relevant for learning, which may lead to improved schemata construction (Van Merriënboer, Schuurman, De Croock, & Paas, 2002; Sweller et al., 1998). This suggests that optimal instructions are those which minimize extraneous and maximize germane load, while simultaneously avoiding overload. However, Van Merriënboer et al. (2002) noticed that the distinction between extraneous and germane cognitive load, although intuitively plausible, cannot reliably be measured. In general, measuring cognitive load in realistic learning situations is not straightforward. While some recent attempts have been made to measure cognitive load directly (e.g., the dual task approach advocated by Brünken et al., 2003), many studies rely on indirect measures such as behavioral or learning outcome measures (see e.g., Brünken et al., 2003; Van Merriënboer et al., 2002). Although these measures only relate to cognitive load indirectly, they do assess learning behavior directly which is sufficient for current purposes. In this study, the learning behavior was measured through learning times, amount of practising during learning, execution times, and the amount of correctly executed exercises.

Arguably, the different information modalities of interest (text, picture, film clip) have different loading potentials, which may influence their performance on one or more of these measures. Of all three modalities, it seems likely that text imposes the highest load: it could be argued that reading a text requires more mental effort than watching a picture or a film clip, hence it seems likely that learning from text takes longer than learning from pictures or film clips. Potentially, an additional complication for learning RSI prevention exercises from text is that schemata

construction may be more involved than for pictures and film clips. In the text condition, learners have to "translate" the textual instructions to manual gestures. Notice that this is a concrete instance of Glenberg's Indexical Hypothesis, stating that readers associate words and phrases with objects and actions in "the real world", which should facilitate understanding (e.g., Glenberg, 1997; Glenberg & Robertson, 1999). An instruction in the form of a film clip and (probably to a lesser extent) a picture, depicts the hand and arm movements the learner should make, while the learner has to infer these gestures when presented in text. Hence, it is hypothesized that learners will practice more often while learning from text than while learning from visual presentations. An interesting question is whether the load imposed by text is only extraneous or also partly germane. It might be that the extra effort required to learn from text might pay off and may lead to good learning results (short execution times, few execution errors), especially when the intrinsic load is low (so that overload can be avoided).
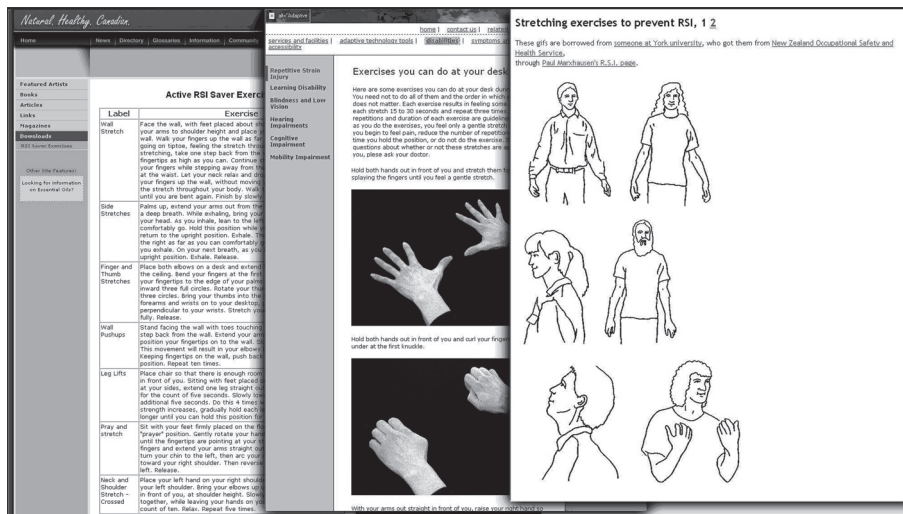
Figure 4.1

Three different web sites[1], which use different formats (text, picture and text, and animation) to illustrate RSI prevention exercises.

In this case, we expect pictures to impose the lowest load of the three presentation modalities; viewing a static picture requires little mental effort. Provided that a picture captures the essential information of a procedure, it is to be expected that learning times for pictures are relatively short. However, due to their static nature, pictures offer little information about the temporal structure of a procedure, and for more complicated exercises (i.e., exercises with a higher intrinsic load) this may hamper schemata construction and might result in suboptimal execution performance.

To what extent film clips impose cognitive load is uncertain: on the one hand, it can be argued that they may induce load, since the film clips continuously change and learners have to process this information which reduces the cognitive resources available for germane load, but they may also lessen cognitive load by relieving the learner in the translation process, which is required when reading text. The main strength of film clips might well be their "informational completeness"; learners do not have to infer the exact sequence of movements from text or from a single snapshot, the entire sequence of actions is visualized, which arguably facilitates schemata construction. This suggests that learners will not practice much during learning. Whether this will also result in few execution errors is not obvious: it may be that the large amount of external support relieves learners of cognitive load demands that they would be able to fulfil, but which might prevent them from performing beneficial cognitive actions for learning.

To find out what the actual strengths and weaknesses of the various information modalities are two experiments are described. In the first experiment (section 4.2), participants learn and execute 20 RSI prevention exercises in two degrees of difficulty. The influence of presenting an instruction in text, picture or film clip was measured through learning times, amount of practicing during learning, execution times, and number of correctly executed exercises. Besides these objective measures, participants are also asked for their subjective satisfaction. In the second experiment (section 4.3), participants are asked which instructional format they preferred in a forced choice experiment. This experiment basically tries to find learning preferences, in a manner somewhat similar to Leutner & Plass (1998), on the visualizer-verbalizer dimension (Kirby & Moore, 1988). The Chapter ends with a general discussion in section 4.4.

4.2 Effectiveness and subjective satisfaction of information modalities

4.2.1 Research method

Participants

Participants were 30 young adults, between 18 and 30 years of age. Of the participants, 15 were male and 15 were female.

Design

The experiment had a 3 (information modality) × 2 (difficulty degree) factorial design, with information modality (dynamic visual [film clip], static visual [picture], text) as between participants variable and difficulty degree (easy, difficult) as within participants variable, and with learning times, amount of practicing during learning, execution times, and number of correctly executed exercises as dependent variables. The participants were randomly assigned to an experimental condition.

Stimuli

A total of 20 RSI prevention exercises were chosen from web sites on RSI prevention and RSI injury prevention software[2]. The chosen exercises were all exercises to prevent RSI in hands and arms. Of the 20 exercises, ten exercises were assumed to be easy to perform, and ten were assumed to be difficult. The criterion for a difficult exercise was that it should be either a complex symmetrical movement or an asymmetrical movement. Complex symmetrical movements were classified as movements consisting of at least three sequential atomic movements, in the course of which both arms and hands make the same movements. Asymmetrical movements were classified as movements in which the participant should make a different movement with each arm or hand. Easy exercises were neither asymmetrical nor complex movements. Figure 4.2 and 4.3 show representative examples of an easy and a difficult exercise. Note that this operationalisation of easy and difficult tasks is based on the relative complexity of the sequence of motoric movements. To find out to what extent this objective criterion coincided with subjective perception of task difficulty a pre-test was carried out, in which 9 participants were asked to classify the

20 exercises (presented in text + picture format, and in random order). They were instructed to make two piles, the first consisting of the ten exercises they considered easiest to perform, the second pile consisting of the ten exercises they considered more difficult to perform. It turned out that of the 10 exercises classified as easy, 7 were indeed perceived as easy by the vast majority (> 75%) of the participants, while the remaining 3 were perceived as difficult by a majority of the participants (presumably because these consisted of gestures that are motorically simple, but not commonly used and hence with which participants may have been less familiar). Of the 10 exercises classified as difficult, 9 were indeed perceived as such by the vast majority (> 75%), while the remaining one was perceived as easy by most participants (interestingly, this was an exercise that was motorically complex, but familiar to most participants). In sum, for 16 of the 20 exercises there was a clear and consistent correlation between the objective and the perceived difficulty degree. Throughout this Chapter, the results relating to the original objective classification of the exercises will be reported (see also footnote [6]).

The 20 RSI prevention exercises were presented in three different formats. In the text condition, the exercises were explained to the participants in a purely textual format. The total amount of words did not differ between the 10 easy and 10 difficult exercises: both counted 268 words in total. Thus, the average length was almost 27 words per exercise. Since some exercises are a few words shorter and others a few words longer, only the mean *total* averages for the 10 exercises in each condition will be reported. Because natural language is often ambiguous, the text exercises were checked on their comprehensibility in a second pre-test with three participants (different from those in the first pre-test). It turned out that a few exercises led to misunderstandings and these exercises were reformulated. The new versions were presented to the same three participants, and they agreed that the reformulations solved the misunderstandings.

In the picture condition, each of the 20 exercises was displayed in a single photograph. For this, pictures were taken with a digital camera of a female who made the exercises. She wore black clothing and the movements were shot against a black background so that only her hands were visible in the picture. The photo depicted the "stroke" of the movement, which is that phase of the movement as it unfolds in time containing the "semantic content" of the movement (Kendon, 1980).

To indicate the direction of movement, arrows were added to the pictures of nine difficult and two easy exercises. The size of the pictures was 1536 * 1014 pixels.

Hold your hands in front of you with your palms facing downwards. Lift both your index fingers from the knuckles. Next, gently drop your index fingers[3].

Figure 4.2

A typical easy RSI exercise

Hold your left arm in front of you and drop the left hand down bending at the wrist. Place your right hand on the knuckles of the left hand. Press right your right hand towards you[4].

Figure 4.3

A typical difficult RSI exercise

For the film clip condition, the same female in an identical surrounding was filmed with a digital film camera (25 frames per second). The total number of frames did not differ between easy and difficult exercises: both counted 1097 frames (average film length was thus 4.39 seconds). Again, since some film clips are somewhat

shorter, and others somewhat longer, only the mean *total* averages will be reported for the 10 easy and the 10 difficult exercises.

The exercises were presented to the participants via a web site: one web site for each condition. The experiment was run on a multimedia PC, with a 17-inch colour monitor. In all three conditions, the presentations of the RSI exercises appeared at the centre of the display. In the film clip web site the movements were shown in a film clip with a start and a stop button and a slider. The participants had the possibility to view the film as often as they desired, but not much use was made of this option. Exercises were presented in one of two random orders to control for possible learning effects.

Procedure

The experiment included three parts: a practice session, the central part in which the 20 RSI exercises were presented, and finally, a questionnaire. The participants took part one at a time. Each participant was invited to an experimental laboratory, and took a seat behind the computer. Participants were told that they would receive 20 exercises which they had to learn and perform to assess to what extent they suffered from RSI. In addition, they were led to believe that their hand and arm movements were filmed because a doctor would later look at the recordings of the participants to determine to what extent they suffered from RSI. The procedure of the experiment is depicted in figure 4.4. After the participants had read the instructions, they could click on the hyperlink "start", and the first trial exercise appeared. Depending on their experimental condition, the participants read or viewed the trial exercise until they thought that they could properly execute the exercise. Subsequently they clicked the hyperlink "next" at the bottom of the page. A new webpage appeared with the text "execute trial exercise 1" at the centre of the page. During the execution of an exercise, participants could not see the instruction. When they had executed the exercise, they clicked the hyperlink "next exercise" at the bottom of the page. After completing the second trial exercise, participants were asked if they had any questions regarding the experimental procedure. If not, the participant could start the actual experiment, proceeding in the same way as during the trial session. During the experiment, there was no further interaction between participants and experiment leader.

After the participants completed the 20 experimental exercises, they received a questionnaire. In this questionnaire the subjective satisfaction regarding the content and structure of the web site as well as the comprehensibility and the attractiveness of the exercises were measured. The questionnaire consisted of 16 bi-polar 7-point semantic differentials addressing structure and content of the web sites as well as comprehensibility and attractiveness of the exercises (see appendix B). The presentation order of the adjectives was randomized. For processing the positive adjectives were mapped to 1 (= very positive) and the negative ones to 7 (= very negative). The participants were debriefed at the end of the experiment.

Data processing
The following data were collected: learning times, number of practiced exercises during the learning time, execution times, number of correctly executed exercises, and the results of the questionnaire. The time it took the participants to learn and execute the exercises was computed from the data of the log program ProxyPlus[5]. This program registered the times associated with participants' mouse clicks during the experiment. The time period between clicking the hyperlink "next" which preceded the instruction of an exercise and the hyperlink "next" which followed the instruction of an exercise was defined as the *learning time* (see figure 4.4). The time period between clicking the hyperlink "next" which followed the instruction of an exercise and the hyperlink "next" that preceded a new instruction of an exercise was defined as the *execution time* (see figure 4.4).

A digital video camera was used to record the movements the participant made during the experiment. These video recordings of the participants' hands and arms were used to assess whether the participants practised the exercise during the learning period and to assess their performance while executing the RSI exercises. Occasionally, a participant performed an exercise in a correct but not intended way. These cases were counted as correctly executed exercises. One judge did the assessment: scoring was easy, and the few difficult cases that did arise were resolved after discussion.

Tests for significance were performed using a repeated measures analysis of variance (ANOVA), with a significance threshold of .05. For post hoc tests the Bonferroni method was used. The internal consistency of the four item sets of the questionnaire was measured using Cronbach's α.
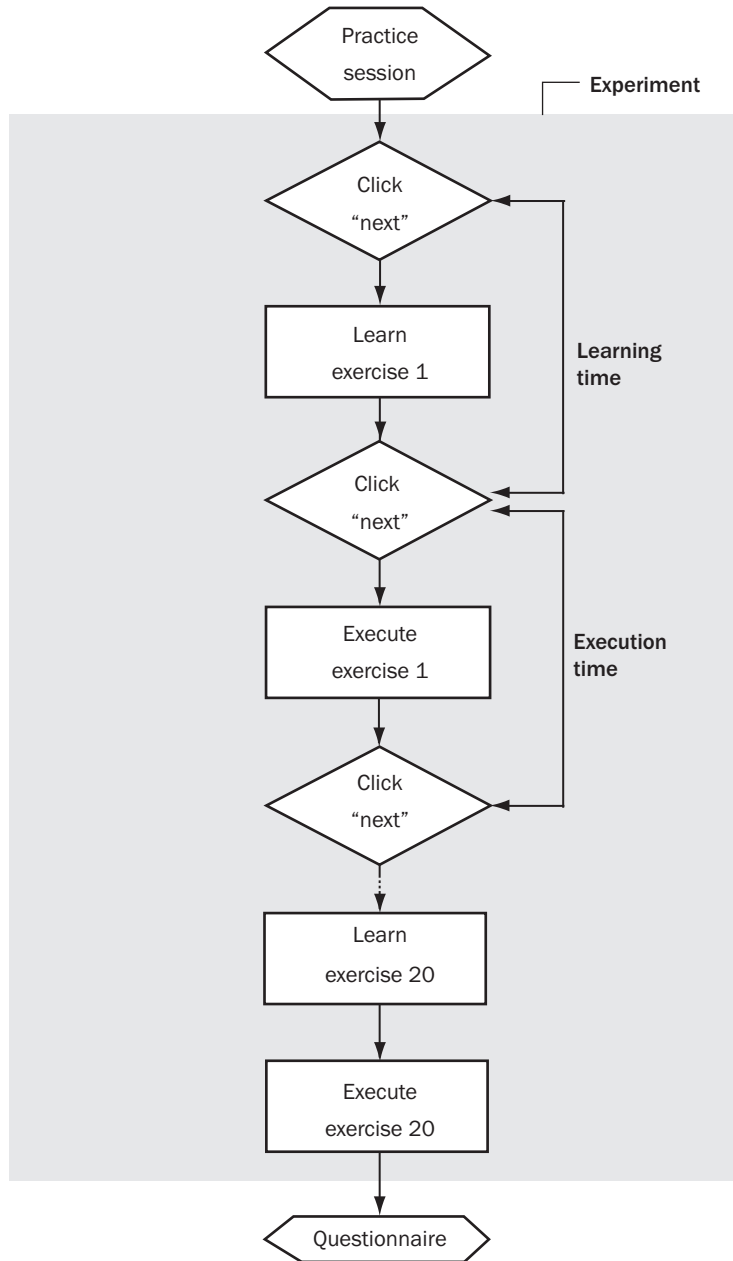
Figure 4.4

Procedure of the experiment

4.2.2 Results

Table 4.1

Mean total time in seconds taken to learn and execute the exercises, average number of exercises for which participants practised during the learning period, and the number of correctly executed exercises as a function of difficulty degree for the three information modalities of interest (standard deviations in parenthesis).

Factor	Difficulty degree	Text		Picture		Flim clip		Average	
Learning time	Easy	96.30	(45.04)	55.10	(17.51)	76.10	(17.07)	75.83	(33.28)
	Difficult	108.10	(40.64)	69.80	(29.04)	91.20	(15.42)	89.70	(33.20)
Practicing	Easy	4.70	(4.79)	1.50	(1.84)	0.00	(0.00)	2.07	(3.48)
	Difficult	5.30	((4.62)	2.20	(2.30)	0.40	(0.70)	2.63	(3.56)
Execution time	Easy	164.20	(48.86)	53.80	(15.48)	77.90	(13.42)	98.63	(56.52)
	Difficult	196.10	(71.28)	77.00	(27.95)	99.60	(31.41)	124.23	(69.89)
Correctly executed exercises	Easy	8.40	(0.70)	9.20	(0.63)	9.60	(0.70)	9.10	(0.84)
	Difficult	7.90	(1.10)	7.60	(1.35)	9.40	(0.70)	8.27	(1.44)

Learning times

Table 4.1 summarizes the results. First consider the learning time. It was found that the difficulty degree had an effect on the amount of time to learn the exercises, $F[1,27] = 37.35$, $p < .001$, $\eta^2_p = .58$. Overall, the participants needed more time to learn the difficult exercises than the easy ones. Also the information presentation had an effect on the learning time, $F[2,27] = 4.53$, $p < .025$, $\eta^2_p = .26$. Participants in the picture condition required the shortest learning times, participants in the text condition had the longest learning times, with learning times from film clips in between these two. Post-hoc tests indicated that the instruction in text differed significantly from the instruction in a picture ($p < .025$). There was no significant difference between the instruction in text and film clip ($p = .35$). Also, the instruction in a picture did not differ significantly from the instruction in a film clip ($p = .25$). No significant interactions between difficulty degree and information modality were found.

Amount of practicing during the learning period

Table 4.1 also reveals that the participants practiced the difficult exercises more often than the easy ones during the learning period, and this difference was found to be statistically significant, F [1,27] = 9.00, p< .01, η^2_p = .25. Also information presentation had an effect on the amount of practising during the learning period, F [2,27] = 6.76, p< .005, η^2_p = .33. In the film clip condition, participants almost never practiced, in the picture condition they practiced for about a fifth of the exercises, while in the text condition participants practiced about half of the exercises. Post-hoc tests showed that the difference between the instruction in text and the instruction in a picture approached significance (p= .06). Text differed significantly from the instruction in a film clip (p< .005). There was no significant difference between the instruction in a picture and a film clip (p= .43), presumably because of the relatively high standard deviation. No significant interaction effects were found.

Execution times

The difficulty degree had a main effect on the amount of time needed to perform the exercises, F [1,27] = 20.84, p< .001, η^2_p = .44. The participants required more time to execute the difficult exercises than the easy ones. There was also a main effect of information modality on the execution times, F [2,27] = 26.78, p< .001, η^2_p = .67. Participants in the text condition had much longer execution times than those in the picture and film clip conditions. Participants in the picture condition needed somewhat less time for execution than the participants in the film clip condition, but the differences are relatively small and the standard deviations are relatively high. Post hoc tests indicated that there was a significant difference between the instruction in text and the instruction in a picture (p< .001). Also, the instruction in text significantly differed from the instruction in a film (p< .001). There was no significant difference between the instruction in a film clip and in a picture (p = .35). There was no significant interaction between difficulty degree and information modality.

Number of correctly executed exercises

A main effect of difficulty degree on the performance was found, F [1,27] = 11.76, p< .005, η^2_p = .26. As can be seen in Table 4.1, the participants executed on average

9.1 easy exercises correctly, as opposed to 8.3 difficult ones. There was also a main effect of information modality: the participants who watched the film clip executed the most movements correctly F [2,27] = 11.68, p< .001, η^2_p = .46. A two-way interaction between difficulty degree and information modality was found, F [2,27] = 3.62, p< .05, η^2_p = .21. This interaction effect can be explained as follows: for the instruction in text, F [1,9] = 1.22, p = .30, η^2_p = .12, and for the instruction in a film clip, F [1,9] = 1.00, p = .34, η^2_p = .10, no significant differences were found in the number of correctly executed easy and difficult exercises. However, for the instruction in a picture, F [1,9] = 12.52, p< .01, η^2_p = .58, a significant difference was found in the number of correctly executed easy and difficult exercises. The participants in the picture condition executed fewer difficult RSI exercises correctly than easy exercises (respectively 7.6 difficult exercises versus 9.2 easy exercises).

Table 4.2

Mean results of the subjective satisfaction questionnaire regarding the structure of the web site, the comprehensibility and attractiveness of the exercises, and the content of the web site's exercises in relation to the 3 experimental conditions (scores range from 1 = "very positive" to 7 = "very negative"; standard deviations in parenthesis). Since the α for Content was below the threshold, the four components are reported separately.

Factor	Subjective satisfaction regarding	Text	Picture	Film clip
	Structure Content	3.35 (0.99)	1.85 (0.58)	2.23 (1.11)
Web site	*Informative*	4.30 (1.95)	3.40 (1.51)	4.20 (1.40)
	Clear	2.50 (1.18)	2.70 (1.49)	2.70 (1.06)
	Comprehensible	2.90 (1.29)	3.30 (1.77)	2.30 (1.43)
	Understandable	3.70 (1.70)	3.00 (2.00)	2.40 (1.42)
Exercises	Comprehensibility	3.80 (1.05)	3.93 (1.32)	2.80 (1.44)
	Attractiveness	3.57 (1.08)	4.07 (1.19)	3.50 (.93)

Subjective satisfaction

The internal consistency on the four items sets in the questionnaire was measured using Cronbach's α. Following standard practice, α was qualified as adequate if its value was higher than .70 (Van Wijk, 2000). For the structure of the web site the α was 0.78, and for the content of the web site the α was 0.56. The α for the comprehensibility of the exercises was 0.82, and for the attractiveness of the exercises 0.83. Table 4.2 gives

an overview of the results of the subjective satisfaction questionnaire. Information modality had no effect on the subjective satisfaction regarding the web site and the exercises. No effects were found between the three conditions for the web site ($F < 1$) and for the exercises ($F < 1$).

4.2.3 Conclusion

The results showed that an instruction in text led to the longest learning times, the most practising of the exercises during the learning phase, and the longest execution times. However, an instruction in text led to a fairly good learning performance, both for easy and difficult exercises. An instruction in a picture led to the lowest learning and execution times. Moreover, the participants in the picture condition engaged in a moderate amount of practicing of the exercises during the learning phase. For easy exercises, learning from pictures led to a good learning performance, but the performance dropped for the difficult exercises, where as many errors were made as in the text condition. Finally, the instruction in a film clip led to medium length learning and execution times. The participants in the film clip condition hardly engaged in practicing the exercises during the learning phase, but they had the highest learning performance, both for easy and for difficult exercises. The subjective satisfaction of the participants regarding the web site and the exercises revealed no differences between the three information modalities. An explanation for this result could be the between-subjects design: participants only saw one realization of each exercise, and arguably could not form an informed preference for one of the three information modalities. Therefore, a second experiment was conducted to find out whether participants preferred one of these three information modalities.

4.3 Subjective preference for information modalities

4.3.1 Research method

Participants

Participants were 26 young adults, between 18 and 25 years old. Of the participants 13 were male and 13 were female. None participated in the first study.

Design

The experiment had a 3 (information modality) × 2 (difficulty degree) factorial design, with information modality and difficulty degree as within participants variables and preference as the dependent variable. The participants were randomly assigned to an experimental condition.

Stimuli

In the second experiment, participants did not have to execute the RSI exercises, but were asked which instructional format (text, picture, or film clip) they preferred for a given exercise. Eight exercises (four easy and four difficult ones) were randomly selected from the 20 exercises from experiment 1. Two of these exercises (one easy and one difficult) were used to instruct the participant during the practice period, the other six were used in the actual experiment. To control for possible learning effects, the exercises were presented in four random orders, i.e., two random orders for the presentation of the information modality and two random orders for the exercises.

Procedure

Participants took part one at a time. They were invited to an experimental laboratory, and were seated behind the computer. They were told that they would receive six RSI exercises to learn in three versions (i.e., text, picture, and film clip). After learning the exercise, their task was to indicate (by forced choice) which of the three information modalities they preferred for that exercise. Following the instructions, participants could proceed with two trial exercises to make them acquainted with the stimuli and

the task. For each exercise, the three presentation formats were presented beneath each other (see figure 4.5). When the participants had observed the three instructional formats, they were instructed to fill in their preferred realization for that exercise on an answer sheet. The sequence in which the information modalities were presented on the answer sheet corresponded to the sequence in which they were presented at the computer screen. After the trial exercises, the experiment leader asked if the participants had any questions regarding the procedure of the experiment. If the procedure was clear participants could start the actual experiment and select their preferred mode of presentation in the same manner as during the practice session. There was no further interaction between participants and experiment leader during the experiment.

A. Film clip

B. Picture

C. Text

Leg uw handen voor u met de handpalmen naar beneden.
Buig uw polsen nu zover mogelijk omlaag,
totdat uw vingertoppen naar beneden wijzen
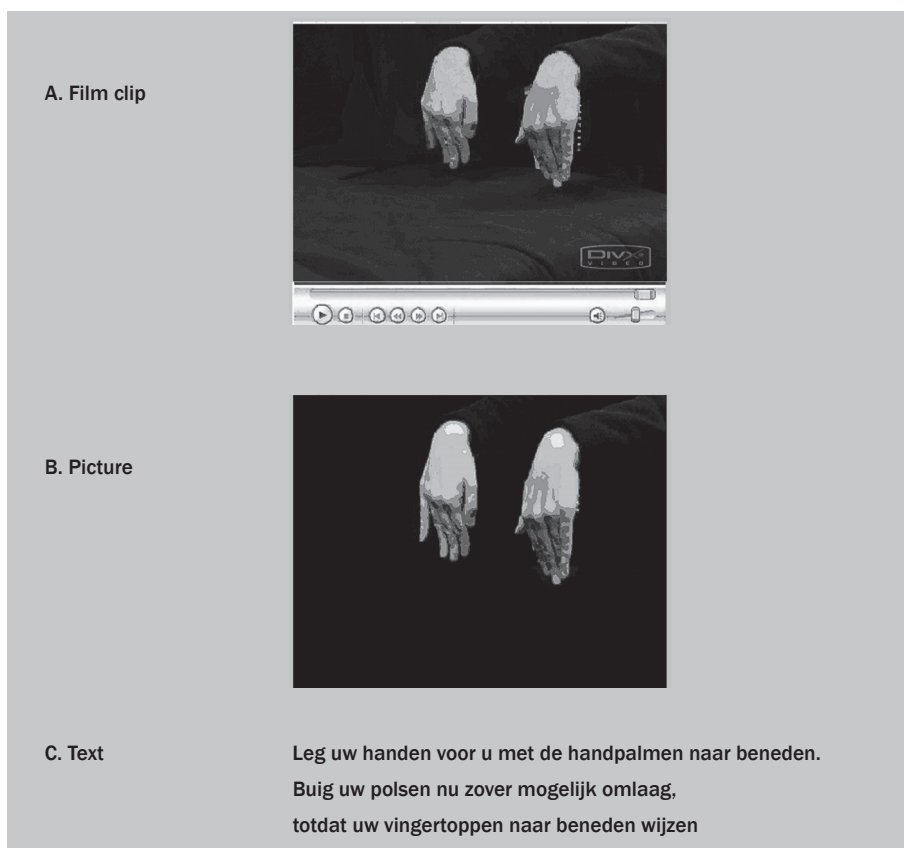
Figure 4.5

Presentation of the exercises in the second experiment

4.3.2 Results

The data were analysed with χ^2-tests to check for significant differences in participants' preferences of presentation formats for RSI exercises. Table 4.3 shows the result: for all exercises the majority of the participants preferred the film clip to text and picture. The overall distribution was significantly different from chance, χ^2 (2) = 81.03, p< .01. There was no effect of difficulty degree on the preference of the participants, χ^2 (5) = 1.61, n.s. Interestingly, there were some notable differences in the distribution of preferences for the first two easy exercises. For the first easy exercise, the participants preferred the film clip and text to the instruction in a picture. For the second easy exercise, the participants preferred the film clip and picture to the instruction in text. This was a significant difference, χ^2 (5) = 12.76, p< .05.

Table 4.3

The distribution of the participant's preferences for text, picture, and film clip in presenting easy and difficult RSI exercises.

	Text	Picture	Film clips	Totals
Easy 1	10	3	13	**26**
Easy 2	1	12	13	**26**
Easy 3	3	0	23	**26**
Difficult 1	6	1	19	**26**
Difficult 2	3	4	19	**26**
Difficult 3	3	5	18	**26**
Totals	**26**	**25**	**105**	**156**

4.3.3 Conclusion

A second experiment was conducted to find out whether participants preferred one of the three information modalities. The results of this study showed that for all exercises most participants preferred the film clips for illustrating the RSI exercises. Thus, there was a significant difference of the instructional format on the subject's preference. Difficulty degree did not influence this preference pattern.

4.4 Discussion

In this Chapter, the learnability of RSI prevention exercises in different presentation formats was investigated. What effective ways of learning such exercises are, is an important research question, in view of the growing awareness of the importance of RSI prevention and the bewildering array of presentation formats for RSI prevention exercises currently being employed on the internet and in RSI prevention software.

Two experiments were performed looking at the effect of offering RSI prevention exercises in three different formats (film clips, pictures, or text) on learning time, amount of practicing, execution time, learning results, and subjective satisfaction. To model variation within a domain, twenty RSI prevention exercises were selected from different sources in such a way that 10 exercises were motorically easy (symmetric and simple) and 10 exercises more difficult (asymmetric or complex), and a pre-test with 9 participants revealed that there was strong connection between objective and perceived difficulty degree[6]. The results of the first experiment indeed reveal that the exercises assumed to be easy were "easier" to perform than the assumed difficult ones, since significant main effects of difficulty degree were found on all dimensions of interest. Thus, irrespective of presentation format, the easy exercises are associated with shorter learning times, less practicing, shorter execution times and fewer execution errors than the difficult exercises. It is worth repeating that the summed length (in terms of frames and number of words) of the 10 easy exercises was exactly the same as the summed length of the 10 difficult ones.

4.4.1 Which information modality was most effective?

Of the three presentation formats under investigation, text was expected to impose the highest load. Overall, text indeed led to the longest learning times, which can in part be ascribed to the fact that it takes more mental effort to read a text than to watch a picture or film clip. But during the learning phase, people not only read the text, but also engage in a substantial amount of practicing which takes time as well. People in the text condition did by far the most practicing, which is consistent with the Indexical Hypothesis (e.g., Glenberg 1997): to foster understanding, participants "translated" the textual instructions into actual movements during learning for

many exercises. Participants must thus engage in fairly deep processing of the textual instructions, but arguably this is to some extent a form of germane load: the deep processing and practicing appear to be beneficial for learning. Contrary to what one might expect, the relatively long learning phase, does not lead to shorter execution times. However, it does lead to a good learning performance. For the easy exercises, participants in the text condition make a few more errors than participants in the other two conditions, but for the difficult exercises, performance is even slightly better than for pictures, which suggests that the germane activities pay off.

Pictures were expected to impose the lowest load. The average learning times were indeed lowest in the picture condition, as were the average execution times. For easy exercises, learning from pictures led to a good performance. In fact, participants made as few errors for these exercises as participants in the film clip condition. But the performance dropped for the difficult exercises, where as many errors were made as in the text condition. An explanation for these results would be that the pictures did not offer a complete model of the difficult exercises. The pictures only depicted the stroke of the movement, with arrows indicating motion where this is applicable. The expectation was that people generally would be able to derive the complete exercise on the basis of this information. However, it turned out that for the difficult RSI exercises the participants lacked information about the exact temporal sequence of movements of the exercise. This might explain why a moderate amount of practicing took place in this condition (more than for film clips).

Concerning the load of film clips, two contrasting hypotheses were mentioned. The first was that film clips might induce load because a learner is presented with continuously changing images, and has to remember the relevant stages of the RSI exercise (Ainsworth & VanLabeke, 2004; Lewalter, 2003). The second hypothesis was that film clips might reduce load, freeing the learner by simply presenting a complete, physical model of the task to be carried out (Tversky et al., 2002). It was found that film clips led to medium length learning times (between picture and text). In part this can be attributed to the fact that watching a clip takes a fixed amount of time. But it is interesting to see that difficult exercises require longer learning times than easy ones, even though they are of the same average duration (and it is not the case that learners played difficult exercises more often), which is probably due to the higher average intrinsic load of the difficult tasks. There was virtually no practicing in the

film clip condition, as expected, since the clips offer an informationally complete model of the task. Contrary to expectation, the execution times were not the shortest, which suggests that participants still had to engage in cognitive activity during the execution phase. Film clips did lead to a consistently high learning performance, both for easy and for difficult exercises. Hence, despite the apparent lack of cognitive effort during learning (no practicing), learners do construct the necessary scheme based on germane cognitive processes. Apparently, germane cognitive processes are not only invested in the learning phase of the exercises, but also in the execution of the exercises. This could explain the relatively longer execution times in the film clip condition.

It is interesting to observe that none of the presentation formats appears to be superior on all dimensions of interest, each has some disadvantage (less efficient for learning, relatively many errors, etc.). In view of the fact that no single modality outperformed the others on all dimensions, it is perhaps not surprising that the first study did not reveal any significant differences on the subjective preference dimension. Hence, a second experiment was performed, in which participants had to select their learning preference via forced choice. They had to do so for 6 randomly selected exercises (3 easy and 3 difficult ones). No significant differences were found for the difficulty degree of the exercises, which might be explained by the fact that the participants of the second study only observed the exercises: they did not have to learn and execute them, and therefore might have processed these exercises on a more superficial level. Interestingly, a general and consistent preference for film clips was found (contrary to the results of the first experiment). What causes this apparent discrepancy between the effectiveness of information modalities and the subjective learning preference is not entirely clear. Part of the explanation may be that the film clips are the most "visually appealing". In addition, it may be that participants of the second experiment recognize that film clips offer a complete action representation, but do not realize that learning from text or a picture may lead to good results as well (and perhaps even quicker than for film clips, see above).

4.4.2   Research limitations

RSI prevention exercises offer a new and ecologically valid learning situation with a number of interesting properties. These exercises are quite brief, and arguably relatively easy to learn. A downside of this is that with respect to learning performance (number of errors) there may be a ceiling effect, in that easy exercises are learned very well for all three modalities. It would be interesting to redo the experiment with more complex RSI related tasks, and see whether this would lead to more differentiation between the different modalities where errors are concerned.

It turned out that it was not always straightforward, to make sure that the exercises in the three conditions offered comparable information, as recommended by Tversky et al. (2002). While a static picture combined with an arrow indicating direction or motion can be very informative, it does not make the entire intended movement explicit as a dynamic picture does. In the former, but not in the latter case, the learner has to infer the full movement, which may lead to errors, especially for the difficult exercises. Still, it is interesting to see that the efficiency of pictures (learning and execution times) is higher than that of the other two modalities, and leads to nearly optimal results for easy exercises. This indicates that a particularly efficient method for illustrating more complex exercises might be via a series of pictures, depicting key stages of the procedure. One would expect that this could lead to both a high efficiency and good learning results, for easy as well as for difficult exercises.

4.4.3   Recommendations for further research

In a somewhat similar vein, it was found that certain RSI exercises seem to be inherently easier to represent than others, and especially that this ease-of-representation may vary across different presentation formats. Some movements can be very concisely described in language, because the entire movement can be coded in a fixed expression (e.g., "Make fists"), whereas other movements can be rather cumbersome to describe (see for example the description of the exercises depicted in figure 4.3) Also expressing how a particular movement "feels" (i.e., "Spread your fingers until a mild stretch between the fingers is felt") is obviously easier in language than in static or dynamic visuals. For such exercises, a textual presentation might

have had an added value over other presentation formats. It would be interesting to systematically vary the presence or absence of linguistic short cuts (describing complex movements in a few words) and investigate how this influences the effectiveness of textual instructions.

In this Chapter the effectiveness of unimodal instructions on learning and executing of RSI exercises was investigated. It would also be interesting to study the effects of multimodal instructions. There are different views on the effectiveness of multimodal information presentations. For example, Mayer's multimedia principle states that people learn better from text and picture rather than from text alone (Mayer, 2005). However, Tindall-Ford, Chandler & Sweller (1997) stated that when one information modality information is intelligible by itself, adding a second modality may be unnecessary. Hence, a small-scale follow-up experiment was performed investigating the effectiveness of multimodal instructions on learning and executing RSI exercises (Van Hooijdonk & Krahmer, 2006). Twenty participants (age between 18 and 25 years old, 10 males and 10 females) had to learn and execute 20 (10 easy and 10 difficult) RSI exercises presented in two different formats: text + picture and text + film clip. The text and picture or film clip were presented together, with both visuals presented above the text. The experiment was conducted under the same circumstances as described in section 4.2. The results showed that these multimodal presentations of RSI exercises did not lead to shorter learning and execution times, nor did they lead to a good learning performance. A possible explanation for these results could be the split attention effect (Sweller et al., 1998). Participants had to switch their attention between the information presented in the text and also to the picture or film clip. Moreover, the information presented in the text was not adjusted to the information presented in the picture or film clip. It would be interesting to replicate this follow-up experiment with multimodal information presentation (text + picture and text + film clip) in which the information presented in text, static and dynamic visuals are adjusted to each other in such way that they complement each other.

In summary, the results of both experiments showed that no single modality outperformed the others on all learning dimensions. This implies that there is no single, straightforward design recommendation on how to present information using different modalities. The goal of the information presentation influences the

type of presentation. For example, if the amount of practicing is considered to be the most important factor, procedural instructions are best described in text. However, if quick leaning is most important, procedural instructions are best illustrated with a picture. Finally, when the goal of the information presentation is a good overall execution, procedural instructions are best visualized with a film clip.

In this chapter, the learning behaviour of the procedural instructions presented in different information modalities was measured indirectly. Although the results indicated that the information presented in different information modalities had different learning outcomes, it is unclear how the information modalities under investigation were actually processed. A research method that could provide an answer to how information modalities are processed is eye tracking. In the next chapter, eye movements are used to investigate the incremental processing of the speech modality.

Footnotes

1   http://naturalhealthcare.ca/rsi_saver_exercises.phtml

    http://web.mit.edu/atic/www/disabilities/rsi/exercises.html

    http://busy-bee.net/rsi/

2   http://web.mit.edu/atic/www/disabilities/rsi/exercises.html

    http://www.workpace.com

3   English translation of Dutch original, in Dutch this exercise contains 29 words.

4   English translation of Dutch original, in Dutch this exercise contains 30 words.

5   http://www.proxyplus.com

6   We redid all statistical analyses omitting the few exercises for which the subjective assessments in the pre-test did not coincide with the objective classification. This led to essentially the same results as those reported above.

Appendix B: Questionnaire addressing the structure and the content of the website as well as comprehensibility and attractiveness of the exercises (in Dutch)

U heeft zojuist een aantal oefeningen uitgevoerd op de RSI Diagnose website. Graag zouden we van u willen weten wat u van de website én van de oefeningen vond. Hieronder staan een aantal stellingen over de website en over de oefeningen. Geef aan wat u van de website én van de oefeningen vond.

De opbouw van de website was:

overzichtelijk	1	2	3	4	5	6	7	onoverzichtelijk
onduidelijk	1	2	3	4	5	6	7	duidelijk
makkelijk te doorgronden	1	2	3	4	5	6	7	moeilijk te doorgronden
moeilijk te doorzien	1	2	3	4	5	6	7	makkelijk te doorzien

De inhoud op de website was:

informatief	1	2	3	4	5	6	7	niet informatief
onduidelijk	1	2	3	4	5	6	7	duidelijk
begrijpelijk	1	2	3	4	5	6	7	onbegrijpelijk
helder	1	2	3	4	5	6	7	vaag

Ik geef de website het rapportcijfer:

1	2	3	4	5	6	7	8	9	10

Begrijpelijkheid van de oefeningen

De oefeningen waren:

moeilijk	1	2	3	4	5	6	7	makkelijk
eenvoudig	1	2	3	4	5	6	7	ingewikkeld
onduidelijk	1	2	3	4	5	6	7	duidelijk
helder	1	2	3	4	5	6	7	ambigue

Aantrekkelijkheid van de oefeningen

De oefeningen waren:

afwisselend	1	2	3	4	5	6	7	eentonig
oninteressant	1	3	3	4	5	6	7	interessant
aansprekend	1	2	3	4	5	6	7	afstandelijk
saai	1	2	3	4	5	6	7	boeiend

Ik geef de oefeningen het rapportcijfer:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Bedankt voor uw medewerking!

5 | Evaluating the speech modality with eye movements

TASK MODALITY METHODOLOGY

Procedural Instructions

Speech Quality Segmental Suprasegmental

Action Types Spatial Expressions

Action Levels

Diphone Synthesis

Protocol Analysis

Spatial Conceptualization

Thinking Aloud

Eye Tracking

RSI Exercises

Hypertext

Illustrative

Visual World Paradigm

Questions and Answers

Text Brief Extended Visuals Informative

Unit Selection Synthesis

Static

Cognitive Engineering

Dynamic

Production

Human Speech

Information Seeking

Learning Preference

Evaluation

Speech Modality

Experimental Evaluation

A journal paper based on this chapter is submitted for publication. Earlier versions of this chapter appeared as Van Hooijdonk, C.M.J., Commandeur, E., Cozijn, R., Krahmer, E.J., & Marsi, E. (2007). The online evaluation of speech synthesis using eye movements. *Proceedings of the sixth ICSA workshop on speech synthesis Workshop*, Bonn, Germany, pp. 385-390 and as Van Hooijdonk, C.M.J., Commandeur, E., Cozijn, R., Krahmer, E.J., Marsi, E. (2007). Using eye movements for online evaluation of speech synthesis. *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2007)*, Antwerp, Belgium, pp. 1346-1349

5.1 Introduction

Developments in speech technology have led to a frequent use of synthetic speech in computer applications, like computer-aided instructions and consumer products (e.g., navigational aids and mobile telephones) (Paris et al., 2000). However, research has shown that synthetic speech is harder to comprehend than human speech. First, synthetic speech is less intelligible than human speech as the acoustic signals of synthesized speech are impoverished (e.g., Luce et al., 1983; Nusbaum & Pisoni, 1985; Reynolds & Givens, 2001). Second, synthetic speech sounds unnatural compared to human speech due to the limited modeling of prosodic cues, like intonation, stress, and durational patterns (Nusbaum et al., 1995).

The evaluation of synthetic speech in terms of intelligibility has primarily been done with offline research methods. For example, the Modified Rhyme Test (House et al., 1965) has been used to investigate the segmental intelligibility of synthetic speech (Pisoni, 1987). In this test, listeners are presented with spoken words and are instructed to select the word they heard from a set of alternatives that differ only in one phoneme. Another example is the Mean Opinion Score (Schmidt-Nielsen, 1995) in which listeners have to rate the quality of spoken sentences on scales (i.e., excellent - bad).

A disadvantage of offline research methods is that no insight is obtained in how listeners process synthetic speech. Online research methods, like eye tracking, give a direct insight in how speech is processed incrementally. In the "visual world paradigm", participants are asked to follow spoken instructions to look up or pick up objects within a visual display (e.g., Altmann & Kamide, 2004; Tanenhaus & Spivey-Knowlton, 1996; Tanenhaus et al., 1995). The fixation patterns on the objects within the display are used to draw inferences about the processing of spoken instructions. Eye tracking might give an idea of how similar the processing of synthetic speech is, compared to the processing of human speech. This idea was first explored by Swift, Campana, Allen, and Tanenhaus (2002) in a study concentrating on acoustically confusable words (e.g., beetle, beaker, and speaker) to see if the "disambiguation" point was processed at comparable time windows for two instances of synthetic speech and human speech. The results showed that both human speech instructions and synthetic speech instructions were indeed processed incrementally. Moreover,

when hearing the onset of the target noun (e.g., *beaker*), the listeners were more likely to look at the cohort competitor (i.e., a noun that shares the same initial phonemes with the target noun, such as *beetle*) than at the rhyme competitor (i.e., a noun that shares the same final phonemes with the target noun, like *speaker*). Finally, the listeners identified the target more rapidly in the human speech condition than in the two synthetic speech conditions.

The intelligibility of speech does not only depend on its segmental quality but also on the quality and the appropriateness of the prosodic information in the speech signal (i.e., suprasegmental quality) (Sanderman & Collier, 1997). The visual world paradigm has more recently been used to investigate how humans process prosodic information (e.g., Chen, Den Os & De Ruiter, 2007; Weber, Braun, & Crocker, 2006). For example, Weber et al. (2006) used eye tracking to investigate how prosodic information influences the processing of spoken referential expressions. In two experiments, participants followed two consecutive instructions to click on an object within a visual display. The first instruction mentioned the referent (e.g., purple scissors). The second instruction either mentioned a target of the same type but with a different color (red scissors) or of a different type and a different color (red vase). The instructions were either realized with an accent on the adjective (e.g., Click on the PURPLE scissors, Click now on the RED scissors) or on the noun (e.g., Click on the purple SCISSORS, Click now on the red SCISSORS). The results showed that the listeners were affected by this prosodic difference. When the first instruction was realized with an accent on the adjective (e.g., Click on the PURPLE scissors), listeners anticipated the upcoming target, i.e., before the onset of the target noun, listeners looked at the target of the same type as the referent but with a different color (red scissors). When both instructions were realized with an accent on the adjective (e.g., Click on the PURPLE scissors, Click now on the RED scissors) this anticipation only increased. However, when the instructions were realized with an accent on the noun (e.g., Click on the purple SCISSORS, Click now on the red SCISSORS), listeners did not anticipate the upcoming target.

Both segmental and suprasegmental quality are important factors in processing synthetic speech. In this paper, we therefore extend on the work by Swift et al. (2002) by focusing on both segmental and suprasegmental aspects of speech. In our evaluation experiment, the participants were given two consecutive spoken

instructions to look at a certain object within the visual display. These instructions were presented in three speech conditions: diphone synthesis, unit selection synthesis, and human speech. Diphone synthesis is based on concatenating pre-recorded diphones (i.e., phoneme transitions), followed by signal processing to obtain the required pitch and duration. Unit synthesis is also based on concatenation and is realized by segmenting recorded human speech in units of variable sizes (e.g., sentences, constituents, words, morphemes, syllables, and diphones). As larger units of natural speech are exploited, requiring less concatenation, the segmental quality of unit synthesis is in general significantly higher than that of diphone synthesis. At the same time, the prosody may be inadequate, because the intended realization of, for example, pitch accents, may not be available in the speech database. Thus, while quality of diphone synthesis is in general inferior to that of unit synthesis, it has the advantage that it can always produce contextually appropriate prosody (albeit by human intervention). In this experiment, we investigated this trade-off between segmental quality on the one hand and contextually appropriate prosody (i.e., suprasegmental quality) on the other from the perspective of humans processing synthetic speech. The human speech condition was added as a baseline to compare processing of natural and synthetic speech.

5.2 Research method

5.2.1 Participants

Thirty-eight native speakers of Dutch (13 male and 25 female, between 18 and 33 years of age) were paid to participate. They had normal or corrected-to-normal vision and normal hearing. None of the participants were color-blind and none had any involvement in speech synthesis research.

5.2.2 Stimuli

Fifteen pairs of Dutch monosyllabic picturable nouns were chosen as stimuli (see appendix C). These nouns shared the same initial phonemes (e.g., *vork - vos*, fork

- fox). The nouns were depicted using the pictures from the corpus of Snodgrass & Vanderwart (1980) or were retrieved from the Internet when this corpus did not contain a suitable picture. Subsequently, the pictures were colored blue and pink using Adobe Photoshop.
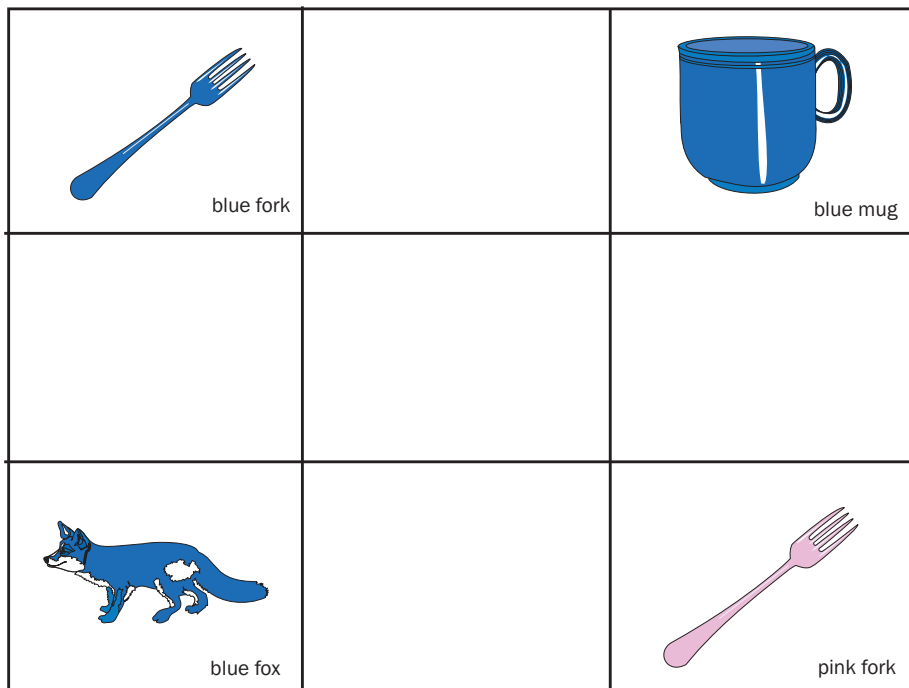
Figure 5.1

Example of a visual display

Each experimental trial consisted of a 3x3 grid with four objects in the corner cells, see Figure 5.1[1]. For every grid, the participants were given two consecutive spoken instructions each referring to a certain object within the grid. In both instructions, the nouns were modified with a color adjective (blue or pink). The first instruction mentioned the **referent** (e.g., *Kijk naar de roze vork*, Look at the pink fork). The second instruction mentioned the **target**. The target could either be of the **same type** as the referent modified with a different color adjective (e.g., *Kijk nu naar de blauwe vork*, Now look at the blue fork), or of a **different type** as the referent modified with

a different color adjective (e.g., *Kijk nu naar de blauwe vos*, Now look at the blue fox). A fourth object was added as a **distractor** (e.g., *blauwe mok*, blue mug). The distractor did not share the type of the other objects, but did share the color with the two targets. The distractor was never mentioned in the experimental trials. The colors blue and pink could occur in both instructions and were randomized across the trials.

The first instruction was realized with a standard, neutral intonation contour meaning that none of the words in the instruction were accented. In the second instruction, the adjective and noun were both accented (e.g., BLAUWE VOS, BLUE FOX). In half of the cases the second instruction had a contextually appropriate double accent pattern while the other half had not, see Table 5.1. The second instruction had an appropriate accent pattern when it mentioned a different color adjective and a different object type as the referent in the first instruction. The second instruction had an inappropriate accent pattern when it mentioned a different color adjective but had the same object type as the referent in the first instruction (Nooteboom & Kruyt, 1987; Terken & Nooteboom, 1987). Note that the choice of a double accent pattern was forced by the output of the unit selection synthesizer, as it typically produces these double accents.

Table 5.1

Example of the instructions accompanying the visual display in Figure 5.1

First instruction	*Kijk naar de roze vork* Look at the pink fork
Second instruction Contextually appropriate double accent pattern	*Kijk nu naar de BLAUWE VOS* Now look at the BLUE FOX
Second instruction Contextually inappropriate double accent pattern	*Kijk nu naar de BLAUWE VORK* Now look at the BLUE FORK

The instructions were realized in three speech conditions, i.e., diphone synthesis, unit selection synthesis, and human speech. A female voice was used for all three speech conditions. The diphone stimuli were produced using the Nextens[2] TTS system for Dutch, which is based on the Festival TTS system (Black, Taylor & Caley, 2002). The input consisted of words and prosodic markup. Pitch accents

were phonetically realized with a rule-based implementation of the Gussenhoven & Rietveld model for Dutch intonation (Gussenhoven & Rietveld, 1992). For the unit selection synthesis a commercially available synthesizer was used. The instructions were obtained through an interactive web interface of the synthesizer. The output that was given by the interface was stored. Note that it was not possible to control the accent patterns of the instructions, as this type of synthesis is dependent on the intonation of the selected units in the database of the synthesizer. The instructions in the human speech condition were recorded by a native speaker of Dutch in a quiet room at Tilburg University. The instructions were digitally recorded, sampling at 44 kHz, using Sony Sound Forge™ and a Sennheiser™ microphone[3]. The instructions were recorded multiple times and the best realizations were chosen. An independent intonation expert checked the utterances using PRAAT (Boersma & Weenink, 1996) to make sure that the intended accents in the second instructions were properly realised. All instructions in the three speech conditions were normalized at -16 dB, using Sony Sound Forge™, and stored in stereo format.

Table 5.2 shows the average length of the instructions for the three speech conditions and the two target object types mentioned in the second instruction. The second instruction was on average 160 milliseconds longer than the first instruction, which is due to the presence of an additional word (i.e., *nu*, now) in the second instruction. We made sure that there were no durational differences in the second instruction[4] realized in the various conditions: speech condition did not affect the duration (F< 1), nor did the target object type (same object type vs. different object type) mentioned in the second instruction. Also, there was no interaction for duration between speech condition and target object type (F< 1).

In addition to the 90 experimental trials (15 stimuli × 3 speech conditions × 2 target object types), 20 filler trials were constructed to add variety to the visual display, and the accent pattern of the second instruction. In the filler trials, either the adjective or the noun mentioned in the second instruction was accented (i.e., ROZE mok, PINK mug or *roze MOK*, pink MUG), and they were only realized in human speech and diphone synthesis. Moreover, all objects within the visual display had the same color (pink or blue) in the filler trials.

Three lists were constructed in a semi-Latin square design, each containing 90 experimental and 20 filler trials. Experimental trials and filler trials appeared in the

same sequential position in all three lists. Participants were randomly assigned to each list, with an equal number of participants assigned to each list.

Table 5.2

The average length of the first instruction and the second instruction for the three speech conditions and target object types mentioned in the second instruction (Average length in milliseconds; standard deviations in parenthesis).

	First instruction		Second instruction			
			Target object of the same type		Target object of a different type	
Diphone synthesis	1904,52	(80,25)	1947,64	(116,87)	1932,91	(99,06)
Unit selection synthesis	1697,38	(132,50)	1932,61	(115,25)	1917,58	(85,94)
Human speech	1719,25	(56,92)	1939,87	(107,845)	1934,61	(76,80)

5.2.3 Procedure

The procedure of the experiment is illustrated in Figure 5.2. Each participant was invited to an experimental laboratory, and was seated in front of a computer monitor. First, the participants were familiarized with the objects that occurred within the visual display during the experiment to ensure that they identified them as intended. This was done by asking them to describe the thirty depicted objects and their color (pink or blue) aloud. The objects were shown in the middle of the computer screen. Participants could view each object at their own pace by clicking on a button, and they were corrected when an object was described incorrectly. This object was viewed again until it was described correctly.

Subsequently, the instructions of the actual experiment were read to the participants, and the eye-tracking system was mounted and calibrated. Participants' eye movements were monitored using an SR Research EyeLink II eye-tracking system, sampling at 250 Hz. Only the right eye of the participant was tracked. The instructions were presented to the participants binaurally through headphones. Next, the participants were presented with a practice session in which the procedure
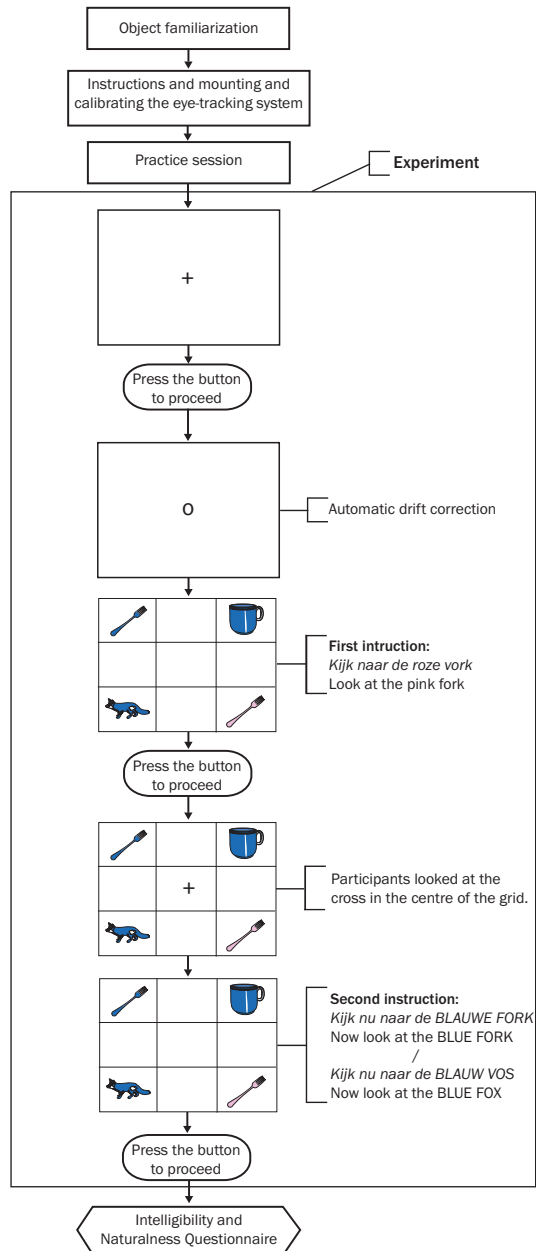
Figure 5.2

Procedure of the experiment

of the experiment was illustrated. This practice session consisted of six trials (3 speech conditions × 2 target object types). The structure of a trial was as follows. First, participants saw a white screen with in the middle a little black cross, and they pressed a button to continue. Next, a white screen appeared with in the middle a central fixation point, and the participants were instructed to look at this point. The experimenter then initiated an automatic drift correction to correct for shifts of the head-mounted tracking system. After the automatic drift correction, the visual display appeared. The first instruction was given after 50 milliseconds. The participants had to look at the object that was mentioned, after which they pushed a button. Subsequently, a little black cross appeared in the centre of the grid and the participants were instructed to look at this cross. After 2000 milliseconds, the cross disappeared and the second instruction was given. Again, the participants had to look at the object that was mentioned, after which they pushed a button. Subsequently, the white screen with in the middle a little black cross appeared again and the participants pressed on a button indicating the start of the next trial. After completing the practice session, the actual experiment started, proceeding in the same way as the practice session. During the experiment, there was no interaction between the participant and the experiment leader.

After the participants completed the experiment, they were asked to listen to an instruction (i.e., *Kijk nu naar de* BLAUWE BLOEM, Now look at the BLUE FLOWER) realised in diphone synthesis, unit selection synthesis, and human speech. Next, they were asked to fill out a questionnaire. This questionnaire consisted of four items about the intelligibility (i.e., audibility, comprehensibility, perceptibility, and distinctness) and four items about the naturalness (i.e., intonation, pleasantness to listen, speech rate, and naturalness) of the three speech conditions. Each question was accompanied with a seven-point Likert scale on which the participants could indicate how much they agreed or disagreed with the content of each item.

5.2.4   Coding procedure and data processing

Eyelink software parsed the eye-movement data into fixations, saccades, and blinks. Fixations were automatically mapped (using the program Fixation[5]) on the objects presented in each trial, and this mapping was checked by hand. The fixations

occurring in the first and second instruction of a trial were analyzed. In the first instruction, trials in which less than 50% of the sample points after the onset of the referent noun belonged to fixations on the referent object were excluded from further analysis. In the second instruction, trials in which less than 50% of the sample points before the onset of the target noun belonged to fixations on the centre of the grid were excluded from further analysis. These trials were excluded because the instructions were not followed. The data of one participant was excluded, as she did not meet the above-mentioned criteria in any of the trials. The total amount of data that was excluded from further analysis was 7.7%, including the data discarded for the above-mentioned participant.

Fixation proportions were averaged over three time windows[6] for each participant F_1 and each item F_2 and analyzed with a 3 (diphone synthesis, unit selection synthesis, human speech) $\times$ 2 (same target object type, different target object type) repeated measures analysis of variance (ANOVA)[7] with a significance threshold of .05. For post hoc tests, the Bonferroni method was used. The dependent variables were the mean proportions of fixations to the target and to the competitor. The first time window began 200 ms after the onset of the target noun, because this is the earliest point at which fixations driven by information from the target noun were expected (e.g., Altmann & Kamide, 2004; Matin, Shao & Boff, 1993). The time window extended over 400 ms, which roughly corresponded to the mean duration of the target noun. The second time window extended from 600 to 1000 ms after the target noun onset. Finally, the third time window extended from 1000 to 1500 ms after the target noun onset. Note, that we checked whether the target noun duration affected the results found and this turned out not to be the case[8].

The results of the questionnaire were processed by mapping them to scores ranging from 1 (= disagree) to 7 (=agree). Next, these scores were analyzed with a 3 (speech condition) $\times$ 4 (items) repeated measures analysis of variance (ANOVA), with a significance threshold of .05. For post hoc tests, the Bonferroni method was used.

5.3 Results

First, we report on the results found for the three speech conditions and the two target object types mentioned in the second instruction in the mean proportions of fixations to the target and the competitor over three time windows: 200-600 ms, 600-1000 ms, and 1000-1500 ms. Subsequently, we report on the results of the questionnaire on the intelligibility and naturalness of the three speech conditions.

5.3.1   Results of the eye movement data

Figure 5.3 shows the fixation patterns for the various conditions. The six figures all show a similar pattern: the mean proportions of fixations to the *target* increased, whereas the mean proportions of fixations to the *competitor* decreased.

The pattern of the mean proportions of fixations to the *competitor* differed between the two target object types mentioned. When the second instruction mentioned a target object of the *same type* as the referent, the fixations to the competitor increased slightly from 200 ms till approximately 450 ms for all three speech conditions. However, when the second instruction mentioned a target object of a *different type* as the referent, the fixations to the competitor increased rapidly for all three speech conditions.

The three figures on the bottom row of Figure 5.3 illustrate that the participants anticipated the upcoming target object when the acoustical information of the target noun became available (i.e., the participants expected that the upcoming target object was of the *same type* as the referent). For the unit selection synthesis and human speech, it was found that the fixations to the competitor increased from 200 ms till approximately 500 ms after which they decreased. For these two speech conditions, the participants revised their anticipation as the acoustic information of the target noun became fully available: at the end of the second instruction, the participants looked at the target (i.e, BLUE FOX) and not at the competitor (i.e., BLUE FORK). However, for the diphone synthesis it was found that the fixations to the competitor increased prior to the target noun onset till approximately 450 ms after which they decreased slightly. At the end of the second instruction approximately 70% of the fixations went to the target (i.e, BLUE FOX), whereas approximately 20% of

the fixations went to the competitor (i.e., BLUE FORK). Thus, the participants found it hard to revise their anticipation on the upcoming target object in the diphone synthesis condition.

Results found within the first time window: 200-600 ms

Table 5.3 summarizes the results found within the time window 200 to 600 ms for the three speech conditions. The statistics showed that the mean proportions of fixations to the *target* did not differ significantly between the three speech conditions. In all three speech conditions, the mean proportions of fixations to the target were approximately 20%. Table 5.3 also shows that there was a significant difference between the three speech conditions in the mean proportions of fixations to the *competitor*. The mean proportions of fixations to the competitor were the highest in the diphone synthesis condition and the lowest in the unit selection synthesis condition. The mean proportions of fixations to the competitor in the human speech condition fell between these two. Post-hoc tests indicated that diphone synthesis differed significantly from unit selection (p< .001) and human speech (p<. 005). Also, unit selection synthesis differed significantly from human speech (p< .05). Table 5.4 shows the results found within the time window 200 to 600 ms for the two target object types mentioned in the second instruction. The mean proportions of fixations to the *target* were significantly higher when the second instruction mentioned a target object of the same type as the referent than when it mentioned a target object of a different type as the referent. Conversely, the mean proportions of fixations to the *competitor* were significantly higher when the second instruction mentioned a target object of a different type as the referent than when it mentioned a target object of the same type as the referent. Finally, an interaction was found between speech condition and target object type mentioned in the second instruction for both the mean proportions of fixations to the *target*, F_1 [2,72] = 18.93, p< .001, η^2_p = .35; F_2 [2,28] = 11.18, p< .001, η^2_p = .44, and to the *competitor*, F_1 [2,72] = 21.95, p< .001, η^2_p = .38; F_2 [2,28] = 9.73, p< .005, η^2_p = .41. Table 5.5 reveals that for all three speech conditions, the mean proportions of fixations to the *target* were significantly higher when the second instruction mentioned a target object of the same type than when it mentioned a target object of a different type. Conversely, for all three speech conditions the mean proportions of fixations to the *competitor* were significantly

higher when the second instruction mentioned a target object of a different type than when it mentioned a target object of the same type.
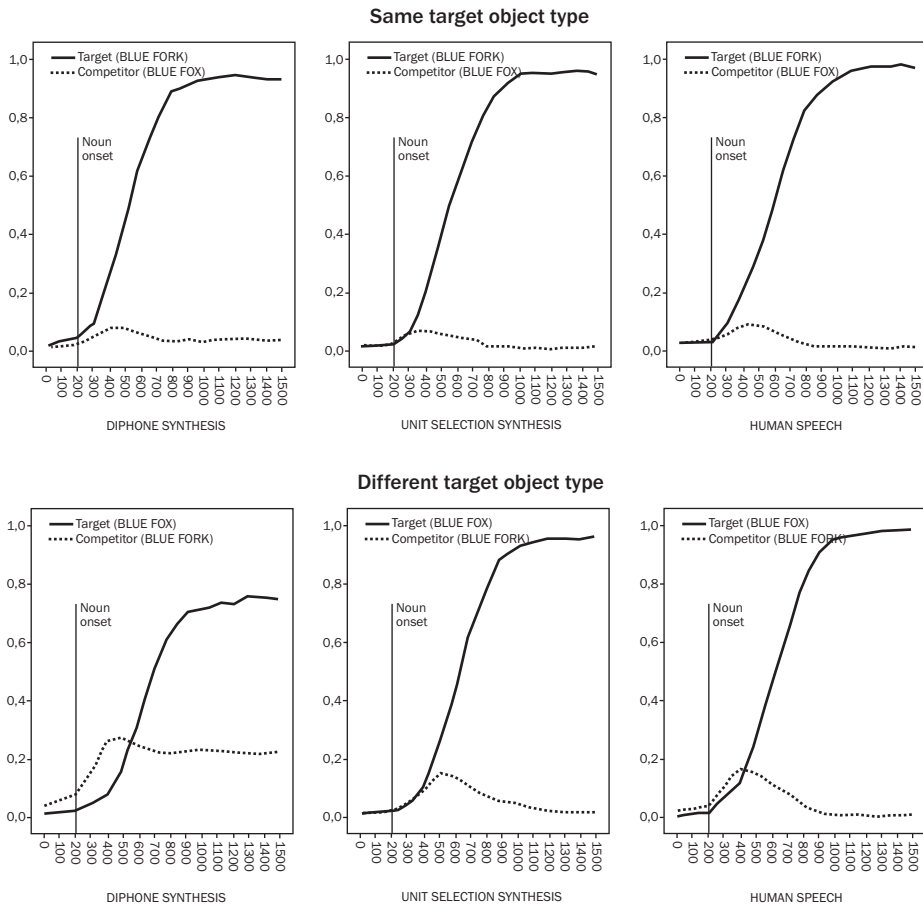
Figure 5.3

Proportions of fixations to the target and the competitor for diphone synthesis, unit selection synthesis, and human speech for the second instruction mentioning a same target object type (top row) and different target object type (bottom row).

Results found within the second time window: 600-1000 ms

In the second time window similar results were found compared to the first time window. These results included the effects of target object type in the mean proportions of fixations to the *target* and to the *competitor* (see Table 5.4). Moreover, a similar interaction was found between speech conditions and target object types in the mean proportions of fixations to the *competitor*, F_1 [2,72] = 53.45, p< .001; η^2_p = .60; F_2 [2,28] = 3.70, p< .05, η^2_p = .21 (see Table 5.5).

Also, different results were found in the second time window. Contrary to the first time window, a significant effect was found of speech condition in the mean proportions of fixations to the *target*, although not by items. The mean proportions of fixations to the target were the highest for unit selection synthesis and human speech and low for diphone synthesis (see Table 5.3). Post-hoc tests showed that diphone synthesis differed significantly from unit selection synthesis (p< .001) and human speech (p< .005). However, there was no significant difference between unit selection synthesis and human speech (p = 1.0). Also, a significant effect was found of speech condition in the mean proportions of fixations to the *competitor* (see Table 5.3). Although this effect was also found in the first time window, the pairwise comparisons between the three speech conditions differed from those found in the first time window. The mean proportions of fixations to the competitor were the highest in the diphone synthesis condition and the lowest in the human significantly higher when the second instruction mentioned a target object of a different type than when it speech condition. The mean proportions of fixations to the competitor in the unit selection synthesis condition fell between these two. Post-hoc tests revealed that diphone synthesis differed significantly from unit selection (p< .001) and human speech (p< .001). However, there was no significant difference between unit selection synthesis and human speech (p = .38). Moreover, an interaction was found between speech condition and target object type in the mean proportions of fixations to the *target*, F_1 [2,72] = 57.20, p< .001; η^2_p = .61; F_2 [2,28] = 5.96 p< .01, η^2_p = .30. This interaction differed from the one found in the first time window. Only in the diphone synthesis and in the unit selection synthesis, the mean proportions of fixations to the target were significantly higher when the second instruction mentioned a target object of the same type than when it mentioned a target object of a different type. For human speech, no difference was found between

Table 5.3

The mean proportions of fixations to the target and the competitor and the corresponding F_1 and F_2 statistics in relation to the three speech conditions and the three time windows

Time window	Object within the grid	Speech condition	Mean proportions of fixations	F_1 and F_2 statistics
200-600 ms	Target	Diphone synthesis	.22	$F_1 < 1$
		Unit selection synthesis	.20	$F_2 < 1$
		Human speech	.21	
	Competitor	Diphone synthesis	.15	$F_1 [2,72] = 20.68, p < .001, \eta^2_p = .37$
		Unit selection synthesis	.09	$F_2 [2,28] = 10.13, p < .001, \eta^2_p = .42$
		Human speech	.11	
600-1000 ms	Target	Diphone synthesis	.72	$F_1 [2,27] = 12.70, p < .001, \eta^2_p = .26$
		Unit selection synthesis	.78	$F_2 [2,28] = 1.32, p = .28$
		Human speech	.78	
	Competitor	Diphone synthesis	.14	$F_1 [2,72] = 57.16, p < .001, \eta^2_p = .61$
		Unit selection synthesis	.06	$F_2 [2,28] = 5.28, p < .025, \eta^2_p = .27$
		Human speech	.04	
1000-1500 ms	Target	Diphone synthesis	.83	$F_1 [2,72] = 117.45, p < .001, \eta^2_p = .77$
		Unit selection synthesis	.95	$F_2 [2,28] = 7.11, p < .005, \eta^2_p = .34$
		Human speech	.96	
	Competitor	Diphone synthesis	.13	$F_1 [2,72] = 126.55, p < .001, \eta^2_p = .78$
		Unit selection synthesis	.02	$F_2 [2,28] = 6.64, p < .005, \eta^2_p = .32$
		Human speech	.01	

Table 5.4

The mean proportions of fixations to the target and the competitor and the corresponding F_1 and F_2 statistics in relation to the two target object types mentioned in the second instruction and the three time windows.

Time window	Object within the grid	Target object type	Mean proportions of fixations	F_1 and F_2 statistics
200-600 ms	Target	Same object type	.26	$F_1 [1,36] = 48.82, p< .001, \eta^2_p = .58$
		Different object type	.16	$F_2 [1,14] = 34.08, p< .001, \eta^2_p = .71$
	Competitor	Same object type	.07	$F_1 [1,36] = 44.40, p< .001, \eta^2_p = .55$
		Different object type	.16	$F_2 [1,14] = 21.67, p< .001, \eta^2_p = .61$
600-1000 ms	Target	Same object type	.82	$F_1 [1,36] = 72.92, p< .001, \eta^2_p = .67$
		Different object type	.70	$F_2 [1,14] = 19.93\ p< .005, \eta^2_p = .59$
	Competitor	Same object type	.03	$F_1 [1,36] = 83.13, p< .001, \eta^2_p = .70$
		Different object type	.12	$F_2 [1,14] = 10.87, p< .01, \eta^2_p = .44$
1000-1500 ms	Target	Same object type	.95	$F_1 [1,36] = 42.84, p< .001, \eta^2_p = .54$
		Different object type	.88	$F_2 [1,14] = 4.91, p< .05, \eta^2_p = .26$
	Competitor	Same object type	.02	$F_1 [1,36] = 55.60, p< .001, \eta^2_p = .61$
		Different object type	.09	$F_2 [1,14] = 5.49, p< .05, \eta^2_p = .28$

Table 5.5

The mean proportions of fixations to the target and the competitor and the corresponding F_1 and F_2 statistics for each time window as a function of speech condition and the target object type mentioned in the second instruction.

Time window	Object within grid	Condition	Mean proportions of fixations	F_1 and F_2 statistics
200-600 ms	Target	Diphone synthesis	Same object type	.31	$F_1 [1,36] = 68.70, p < .001, \eta_p^2 = .66$
			Different object type	.12	$F_2 [1,14] = 69.78, p < .001, \eta_p^2 = .83$
		Unit selection synthesis	Same object type	.24	$F_1 [1,36] = 27.53, p < .001, \eta_p^2 = .44$
			Different object type	.16	$F_2 [1,14] = 18.18, p < .005, \eta_p^2 = .57$
		Human speech	Same object type	.24	$F_1 [1,36] = 7.23, p < .025, \eta_p^2 = .17$
			Different object type	.18	$F_2 [1,14] = 3.23, p = .09$
	Competitor	Diphone synthesis	Same object type	.07	$F_1 [1,36] = 53.78, p < .001, \eta_p^2 = .60$
			Different object type	.23	$F_2 [1,14] = 29.95, p < .001, \eta_p^2 = .68$
		Unit selection synthesis	Same object type	.06	$F_1 [1,36] = 8.46, p < .01, \eta_p^2 = .19$
			Different object type	.11	$F_2 [1,14] = 4.39, p = .06$
		Human speech	Same object type	.08	$F_1 [1,36] = 16.26, p < .001, \eta_p^2 = .31$
			Different object type	.14	$F_2 [1,14] = 6.60, p < .025, \eta_p^2 = .32$
600-1000 ms	Target	Diphone synthesis	Same object type	.85	$F_1 [1,36] = 129.89, p < .001, \eta_p^2 = .78$
			Different object type	.59	$F_2 [1,14] = 13.18, p < .005, \eta_p^2 = .49$
		Unit selection synthesis	Same object type	.81	$F_1 [1,36] = 17.12, p < .001, \eta_p^2 = .32$
			Different object type	.75	$F_2 [1,14] = 7.26, p < .025, \eta_p^2 = .34$
		Human speech	Same object type	.79	$F_1 [1,36] = 1.52, p = .23$
			Different object type	.77	$F_2 < 1$
	Competitor	Diphone synthesis	Same object type	.04	$F_1 [1,36] = 104.41, p < .001, \eta_p^2 = .74$
			Different object type	.23	$F_2 [1,14] = 6.59, p < .025, \eta_p^2 = .32$
		Unit selection synthesis	Same object type	.03	$F_1 [1,36] = 18.46, p < .001, \eta_p^2 = .34$
			Different object type	.08	$F_2 [1,14] = 4.85, p < .05, \eta_p^2 = .26$
		Human speech	Same object type	.03	$F_1 [1,36] = 8.71, p < .01, \eta_p^2 = .20$
			Different object type	.05	$F_2 [1,14] = 2.19, p = .16$

Time window	Object within grid	Condition	Mean proportions of fixations	F_1 and F_2 statistics
1000-1500 ms	Target	Diphone synthesis	Same object type .93	$F_1 [1,36] = 106.29, p< .001, \eta^2_p = .75$
			Different object type .73	$F_2 [1,14] = 5.18, p< .05, \eta^2_p = .27$
		Unit selection synthesis	Same object type .96	$F_1 < 1$
			Different object type .95	$F_2 < 1$
		Human speech	Same object type .96	$F_1 < 1$
			Different object type .97	$F_2 < 1$
	Competitor	Diphone synthesis	Same object type .04	$F_1 [1,36] = 72.10, p< .001, \eta^2_p = .67$
			Different object type .22	$F_2 [1,14] = 4.83, p< .05, \eta^2_p = .26$
		Unit selection synthesis	Same object type .01	$F_1 [1,36] = 6.23, p< .025, \eta^2_p = .15$
			Different object type .03	$F_2 [1,14] = 2.16, p= .16$
		Human speech	Same object type .01	$F_1 < 1$
			Different object type .01	$F_2 < 1$

the target object types in the mean proportions of fixations to the target (see Table 5.5).

Results found within the third time window: 1000-1500 ms

In the third time window similar results were found compared to first and the second time window. As in the second time window, significant effects were found of speech condition in the mean proportions of fixations to the *target* and to the *competitor* (see Table 5.3). Also, significant effects were found of target object type in the mean proportions of fixations to the *target* and to the *competitor* (see Table 5.4). These effects were also similar to the ones found within the first and second time window.

In contrast to the first and the second time window, different interactions were found between speech condition and target object type in the mean proportions of fixations to the *target*, F_1 [2,72] = 68.84, p< .001; η^2_p = .66; F_2 [2,28] = 5.03 p< .025, η^2_p = .36, and to the *competitor*, F_1 [2,72] = 61.96, p< .001; η^2_p = .63; F_2 [2,28] = 4.33, p< .025, η^2_p = .24. These interactions can be explained as follows: for the *target* it was the case that only for the diphone synthesis the mean proportion of fixations were significantly higher when the second instruction mentioned a target object of the same type than when it mentioned a target object of a different type. For unit selection synthesis and human speech no significant differences were found between the target object types mentioned in the second instruction (see Table 5.5). For the *competitor* it was the case that for the diphone synthesis and the unit selection synthesis (although not by items) the mean proportions of fixations were mentioned a target object of the same type. For human speech, no significant difference was found between the target object types mentioned in the second instruction (see Table 5.5).

5.3.2   Intelligibility and naturalness of the three speech conditions

Figure 5.4 illustrates the mean scores found for the questionnaire on the intelligibility and the naturalness of the three speech conditions. A significant effect was found of speech condition for both intelligibility: F [2,72] = 42.52, p< .001, η^2_p = .54 and naturalness: F [2,72] = 49.83, p< .001, η^2_p = .58. Post-hoc tests showed that all pairwise comparisons were significant at p< .001. For both intelligibility and

naturalness diphone synthesis was rated lowest followed by unit selection synthesis. Human speech was rated highest. Figure 5.4 also shows that the participants were homogeneous in their ratings on the intelligibility and the naturalness of human speech, but they were heterogeneous in their ratings of unit selection synthesis and even more for diphone synthesis.
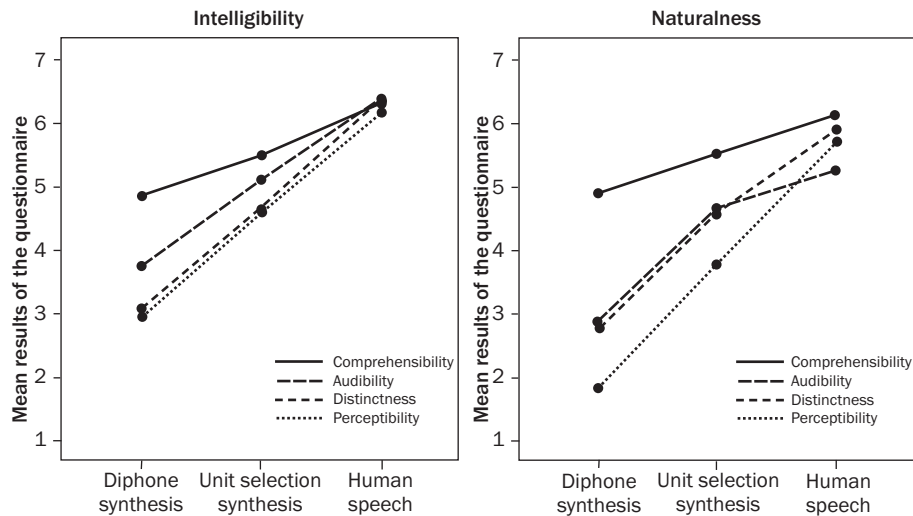
Figure 5.4

Results of the questionnaire on the intelligibility and the naturalness of diphone synthesis, unit selection synthesis, and human speech (Mean scores on a seven point Likert scale: scores range from 1 = "disagree" to 7 = "agree").

5.3.3 Conclusion

The eye movement data showed a similar pattern in all six conditions: the fixations to the target increased, whereas the fixations to the competitor decreased. Also, significant differences were found between the three speech conditions. The performance was best for human speech and worst for diphone synthesis. The performance for unit selection synthesis fell between these two. Moreover, the participants anticipated the upcoming target object. When a different target object type was mentioned in the second instruction this anticipation was hard to overrule

for diphone synthesis as the fixations to the competitor remained high (see figure 5.3, bottom row). For unit selection synthesis and human speech this anticipation was overruled as fixations to the competitor decreased (see figure 5.3, bottom row). Finally, the results of the questionnaire corresponded with eye-movement data. Human speech was rated most intelligible and natural followed by unit selection synthesis and diphone synthesis.

5.4 Discussion

This chapter described an eye tracking experiment to study the intelligibility of diphone synthesis, unit selection synthesis, and human speech taking both segmental and suprasegmental speech quality into account. Diphone synthesis is based on concatenating diphones (i.e., phoneme transitions) after which signal processing is done to obtain the required pitch. Arguably, diphone synthesis has a relatively poor segmental quality, whereas its suprasegmental quality is relatively good. Unit selection synthesis is based on the concatenation of prerecorded human speech of variable sizes (i.e., from sentences to diphones). This type of concatenation leads to a relatively good segmental quality, but its suprasegmental quality is relatively poor. In the experiment, we investigated the trade-off between segmental quality on the one hand and suprasegmental quality on the other. The human speech condition was added to compare processing of natural and synthetic speech.

In the experiment, we used the visual world paradigm (e.g., Tanenhaus et al., 1995) to obtain insight in how synthetic and natural speech is incrementally processed. The participants were presented with a visual display and heard two consecutive spoken instructions. The first instruction was realized with a neutral intonation contour. The second instruction either had a contextually appropriate or inappropriate double accent pattern. After the participants had completed the eye tracking experiment, they filled out a questionnaire on the intelligibility and naturalness of the three speech conditions. This was done to check whether the objective perception corresponded with the subjective perception of the three speech conditions.

5.4.1 Comparing the intelligibility of synthetic and natural speech

The eye movement data revealed differences in the incremental processing of synthetic and natural speech. As expected, the participants identified the target most rapidly in the human speech condition. In the second and third time window, the mean proportions of fixations to the target were high, whereas the mean proportions of fixations to the competitor were low. Similar results were found for unit selection synthesis: in the second and third time window, the mean proportions of fixations to the target were equally high, whereas the mean proportions of fixations to the competitor were equally low. For the diphone synthesis different results were found. The mean proportions of fixations to the target were relatively low, and the mean proportions of fixations to the competitor were relatively high in the second and third time window. These results show that the differences in segmental intelligibility were reflected in the incremental processing of the three speech conditions: participants identified the mentioned target object most rapidly in the human speech condition (having the best segmental intelligibility) and least rapidly in the diphone synthesis condition (having the worst segmental intelligibility). The performance of the participants in the unit selection synthesis condition fell between these two. Interestingly, differences between the three speech conditions in the mean proportions fixations to the *target* were not found in the first time window (i.e., 200 to 600 ms). An explanation for this result could be type of target nouns (i.e., monosyllabic nouns that shared the same initial phonemes). Possibly, the disambiguation point of the target noun could not be perceived in the first time window as the target and competitor nouns only differed in their last phonemes.

We also found that the participants anticipated the upcoming target object mentioned in the second instruction. When the second instruction mentioned a target object of the *same type* as the referent, the mean proportions of fixations to the target increased rapidly, while the mean proportions fixations to the competitor remained relatively low (see figure 5.3, top row). However, when the second instruction mentioned a target object of a *different type* as the referent, both the mean proportions of fixations to the target and to the competitor increased (see figure 5.3, bottom row). Thus, when the acoustical information of the target noun became available, the participants expected the target to be of the same type as the

referent. This expectation was met when the second instruction mentioned a target object of the same type. However, the participants' expectation needed to be revised when the second instruction mentioned target object of a different type, resulting in a decrease in the mean proportions of fixations to the competitor and in an increase in the mean proportions of fixations to the target. A possible explanation for this result could be that participants interpreted the accent on the adjective in the second instruction contrastively to the first instruction.

Note that the second instruction was realized with a double accent pattern. This accent pattern was contextually appropriate when the second instruction mentioned a different color adjective and a different target object type as the referent mentioned in the first instruction (i.e., first instruction: *Kijk naar de roze vork*, Look at the pink fork; second instruction: *Kijk nu naar de BLAUWE VOS*, Now look at the BLUE FOX). When the second instruction was realized with a contextually appropriate double accent pattern, the participants found it difficult to correctly identify the target object as the mean proportions of fixations to the competitor increased (see Figure 5.3, bottom row). The accent pattern was contextually inappropriate when the second instruction mentioned a different color adjective but the same target object type as the referent (i.e., first instruction: *Kijk naar de roze vork*, Look at the pink fork; second instruction: *Kijk nu naar de BLAUWE VORK*, Now look at the BLUE FORK). When the second instruction was realized with a contextually inappropriate double accent pattern, the participants rapidly identified the mentioned target object (see Figure 5.3, top row). Note that these results do not correspond with the results found by Nooteboom and Kruyt (1987) and Terken and Nooteboom (1987) who demonstrated that accenting 'given' information (i.e., information given in preceding utterances) slowed reaction times. Arguably, the perceived prominence of the accent on the adjective might have been higher than the perceived prominence of the accent on the noun which in turn might have led to a contrastive interpretation of the accented adjective.

Moreover, interactions between speech condition and target object type were found in the mean proportions of fixations to the target and to the competitor in the three time windows. In the first time window, the mean proportions of fixations to the target were significantly higher when the second instruction mentioned a same target object type as the referent for all three speech conditions. The reverse was found

for all three speech conditions in the mean proportion of fixations to the competitor. A different interaction was found in the second time window: only in the diphone synthesis and in the unit selection synthesis the mean proportions of fixations to the target were significantly higher when the second instruction mentioned a same target object type as the referent. However, the interaction between speech condition and target object type was significant for all three speech conditions in the mean proportions of fixations to the competitor. Also, in the third time window the interaction between speech condition and target object type was different. Only for diphone synthesis the mean proportions of fixations to the target were significantly higher when the second instruction mentioned a same target object type. Moreover, the mean proportions of fixations to the competitor were significantly higher when the second instruction mentioned a different same target object type for diphone synthesis and unit selection synthesis. Thus, in all three time windows an interaction between speech condition and target object type was found. The results showed that the participants anticipated the upcoming target object. Moreover, the results showed that when the second instruction mentioned a different target object type, the participants could revise this anticipation for human speech and unit selection synthesis. However, this anticipation was hard to revise for the diphone synthesis. Arguably, the relatively poor segmental intelligibility of the diphone synthesis made it harder for the participants to determine the disambiguation point of nouns sharing the same initial phonemes. Due to the relatively poor segmental intelligibility of the diphone synthesis, participants could have relied relatively more on the prosodic information for the interpretation of the upcoming target object.

Finally, the results of the questionnaire corresponded with eye-movement data. For both intelligibility and naturalness, human speech was rated and diphone synthesis was rated lowest. Unit selection synthesis fell between these two. Interestingly, the participants were homogeneous in their assessment on the intelligibility and naturalness of human speech. However, their assessment was heterogeneous on the intelligibility and naturalness of unit selection synthesis and diphone synthesis. Possibly, the notions 'intelligibility' and 'naturalness' are rather complex especially when the speech itself has a relatively poor quality.

5.4.2   Research limitations and directions for future research

In this chapter, we described an experiment in which online and offline research methods were used to evaluate the intelligibility of synthetic and human speech. One might wonder what the advantage is of using an online research method (i.e., eye tracking) above using an offline research method (i.e., a questionnaire) as both research methods had similar outcomes: human speech was more intelligible than unit selection synthesis and diphone synthesis. The results of the questionnaire showed that there were differences in the intelligibility of the three speech conditions. However, the eye movement data enabled us to make fine-grained comparisons between the incremental processing of human and synthetic speech. For instance, the eye movements showed how soon listeners identified the target object in the different speech conditions. Also, the eye movements showed when differences in the processing of human speech and synthetic speech occurred and under which circumstances. For example, the eye movements clearly illustrated participants' confusion when a different target object was mentioned in the diphone synthesis. Thus, eye tracking, and the visual world paradigm in particular, offers a new way of evaluating speech synthesis as it provides a direct insight in how listeners incrementally process speech. However, a disadvantage of eye tracking is that is rather time consuming research method.

In this experiment, the type of task that was presented to the participants was relatively easy. The visual display was small and the instructions were brief. It would be interesting to redo the experiment with a bigger visual display and more complex instructions (e.g., Dahan, Tanenhaus & Chambers, 2002). Moreover, the suprasegmental quality of the instructions was limited in the use of either a contextually appropriate or inappropriate double accent pattern. Research by Weber et al. (2006) and Chen et al. (2007) has shown that different accent patterns influence the online processing of human and synthetic speech. Therefore, it would be interesting to investigate how different accent patterns influence the intelligibility of synthetic speech.

Footnotes

1. The textual descriptions in Figure 5.1 are only added for illustrative purposes, they did not occur in the actual experiment.

2. http://nextens.uvt.nl

3. Type SKM 135 G2

4. We also checked whether there were durational differences between the target nouns within the various conditions. It turned out that speech condition affected the duration of the target nouns, F_2 [2,28] = 5.13, p< .025, η^2_p= .27. Post-hoc tests indicated that there was only a significant difference in target noun duration between unit selection synthesis and human speech (p< .025). The target nouns realized in human speech were on average 81 milliseconds longer than the ones realized in unit selection synthesis. Target object type did not affect the target noun duration, F_2 [1,14] = 1.17, p = .30, nor was there an interaction between speech condition and target object type for the target noun duration, F_2 < 1.

5. http://www.tilburguniversity.nl/faculties/humanities/people/cozijn/research

6. There is a variety in the amount and the length of time windows analyzed, for instance Chen et al. (2007) analyzed 12 time windows of 33 ms, while Dahan & Tananhaus (2004) used one time window of 300 ms. In this experiment, we did not formulate expectations on when possible effects of speech condition and target object type would occur. Therefore, we divided the time interval 200 to 1500 ms in three time windows that extended over 400 ms.

7. Mauchly's test of sphericity was significant for some main effects and interactions. For these cases, we looked both at Greenhouse-Geisser and Huynh-Feldt corrections on the degrees of freedom, which gave similar results. For the sake of transparency, we report on the normal degrees of freedom.

8. The results of the ANCOVA showed that for the first time window (i.e., 200 to 600 ms) and third time window (i.e., 1000 to 1500 ms), the target noun duration did not affect the results found for speech condition and target object type in the mean proportions of fixations to the target and to the competitor. For the second time window (i.e., 600 to 1000 ms), the target noun duration only affected the results found for speech condition in the mean proportions of fixations to the target (F_2 < 1) and the results found for the interaction between speech condition and target object type in the mean proportions of fixations to the competitor (F_2 [2,83] = 2.94, p = .06).

Appendix C: Thirty Dutch monosyllabic nouns with their English translations

	Dutch		English	
1	Bloem	Bloes	Flower	Blouse
2	Boom	Boot	Tree	Boat
3	Hark	Harp	Rake	Harp
4	Kam	Kan	Comb	Jug
5	Krans	Krant	Wreath	Paper
6	Mok	Mot	Mug	Moth
7	Pen	Pet	Pen	Cap
8	Plank	Plant	Shelf	Plant
9	Pijl	Pijp	Arrow	Pipe
10	Schaal	Schaar	Bowl	Scissors
11	Tak	Tas	Branch	Bag
12	Tol	Ton	Top	Barrel
13	Vaan	Vaas	Banner	Vase
14	Vlag	Vlam	Flag	Flame
15	Vork	Vos	Fork	Fox

6 | General conclusion and discussion
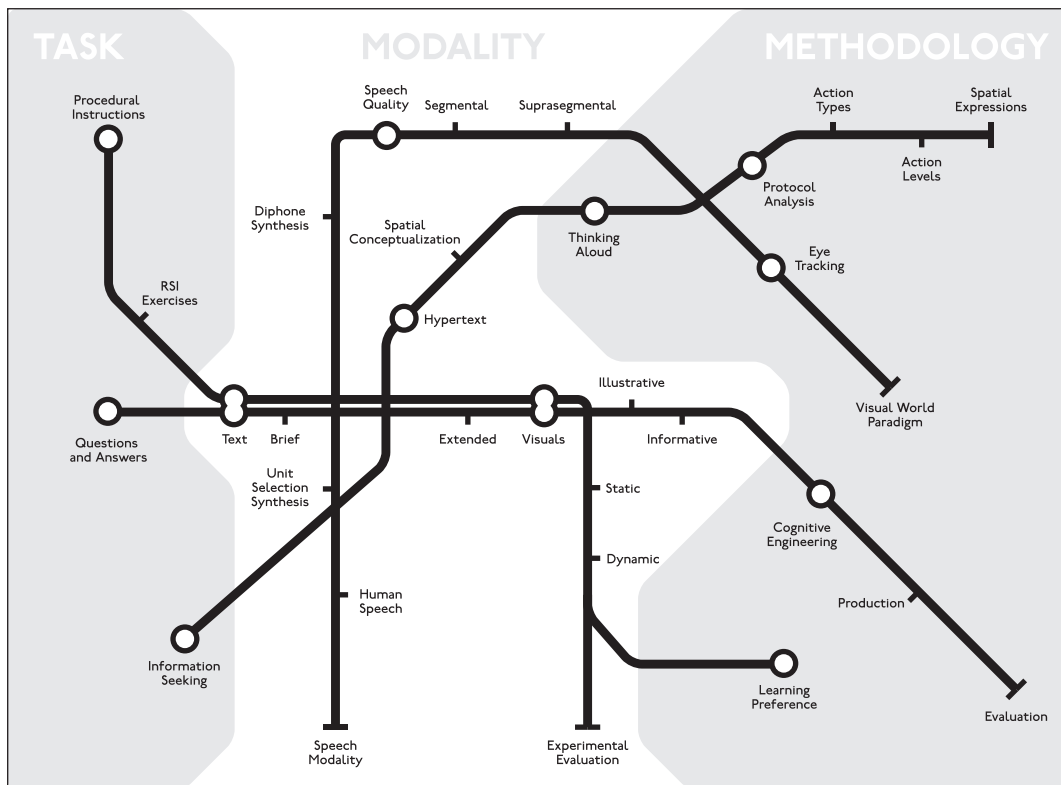
The four studies in this dissertation attempt to contribute to our understanding of the production, processing, and evaluation of multimodal information presentations. In this dissertation different research areas were addressed, each having a specific view on multimodal information presentations and different research methodologies to evaluate them. In this chapter, the main findings of the four chapters are summarized. In addition, we discuss possibilities for future research and end with some final remarks.

6.1 Conclusion

In Chapter 2, we presented two explorative studies in which we looked at the basic issues around multimodal information presentation. In the first experiment, the following research question was posed:

- **When and how do people present information in a multimodal way?**

To answer this research question a production experiment was carried out to determine which modalities people choose to answer different types of questions (definition vs. procedural). The participants had to create potentially multimodal presentations of answers (brief and extended ones) to general medical questions. The collected answer presentations were coded on the presence of visual media (i.e., photos, graphics, and animations) and their degree of informativity (i.e., decorational < representational < informative). The results of the production experiment showed that almost one in four answers contained one or more visual media. The design of these presentations was affected by the answer length: informative visuals occurred most often in brief answers while representational visuals occurred most often in extended answers. Arguably, when an answer does not contain much text, it is more likely that a visual easily contains additional information with regard to the text. And conversely, when the answer contains much text, it is likely that a visual will represent the information already present in the text. Also, the question type influenced the design of the answer presentations: representational visuals were most frequent in the answers to definition questions, whereas informative visuals were most frequent

in the answers to procedural questions. A possible explanation for this result could be that the textual answers to definition questions (e.g., "How many molars does a human have?") often explained an element of the question, like 'molars', which was represented with a visual. Visuals in the answers of procedural questions were often used to explain the steps within the procedure and therefore added information to the textual answer.

In the second study the following research question was investigated:

- **How do people evaluate unimodal and multimodal information presentations?**

An evaluation experiment was conducted in which participants had to assess the informativity and attractiveness of answer presentations for different types of medical questions (i.e., definition vs. procedural). The answer presentations originated from the production experiment and were manipulated in their answer length (brief vs. extended) and their type of visual (i.e., no visual vs. visuals with a low informative value vs. visuals with a high informative value). The results showed that answer presentations having a visual with a high informative value were evaluated as most informative and most attractive.

In Chapter 3, we looked at multimodal information presentation from the perspective of web site usability. When users are navigating in large multimodal information environments, such as web sites, it is important that they find the information they need. However, users often experience problems when searching for information in web sites, like 'disorientation' and 'cognitive overload' (e.g., Ahuja & Webster, 2001; Conklin, 1987; Elm & Woods, 1985). In order to help users finding the information they need, we have to investigate how they conceptualize their actions when navigating web sites. There are several indications that users conceptualize web sites spatially (e.g., Boechler, 2001; Maglio & Matlock, 2003). For instance, people talk about the Internet using spatial metaphors, such as 'jumping from page to page'. Chapter 3 discussed an explorative study in which we investigated the following research question:

- **How do users conceptualize their actions when navigating in multimodal information environments? And what is the role of real and virtual space in this conceptualization?**

Ten thinking aloud protocols were collected that originated from two different usability studies. In both studies, users were asked to perform simple search tasks in a web site (looking up the answers to factual questions), and to think aloud while executing these tasks. The ten protocols were analysed on the types of actions (executions vs. evaluations) and the levels of actions (first level vs. second level vs. third level) users were involved in when navigating a web site. In particular, we coded which action types and levels were expressed in spatial terms. The results of the protocol analysis showed that verbalizations were mostly referring to evaluations (e.g., "I cannot click on this item"). For the levels of actions, it was found that verbalizations referring to the first action level occurred most often (e.g., "I am double clicking on this object"). Moreover, the results indicated that spatial expressions were most frequent when users described executions on the first action level (e.g., "I am going back to the homepage"). In general, the research results confirmed that people use spatial expressions when navigating a web site. However, the difference between a spatial or a non-spatial expression was not always clear. For example, the verbalization "I am in the main menu" could be interpreted as an expression indicating some awareness of the web site's structural organization, or it could refer to something which the user merely perceives.

Chapter 4 focussed on the effectiveness of different information modalities through the following research question:

- **Which presentation modes are most effective for learning and executing procedural instructions?**

To answer this question an experiment was conducted to investigate which information modality (text vs. picture vs. film clip) was most effective for learning and executing RSI prevention exercises which differed in their complexity (easy exercises: simple symmetrical movements vs. difficult exercises: complex symmetrical movements or asymmetrical movements). The influence of presenting an instruction in text,

picture, or film clip was measured through learning times, amount of practicing during learning, execution times, and number of correctly executed exercises. Participants were also asked for their subjective satisfaction. The results showed that no single modality outperformed the others on all dependent variables. Also, the subjective satisfaction of the participants revealed no differences between the three information modalities. An explanation for this result could be the between-subjects design of the experiment. Therefore, a second study was conducted in which the following research question was posed:

- **Which presentation modes do people prefer when learning procedural instructions?**

In this second experiment participants did not have to execute the RSI exercises, but were asked which information modality (text, picture, or film clip) they preferred for six RSI prevention exercises. The results showed that overall participants preferred film clips to learn the RSI exercises. However, we also found that for some exercises (e.g., "make fists"), users preferred an instruction in text to an instruction in a picture and film clip. In sum, the research results of both experiments showed that no single modality outperformed the others. This could imply that the effectiveness of presentation modes depends on several factors, like the communicative goal of the information presentation (i.e., merely observing RSI exercises vs. executing RSI exercises) or the qualities of presentation modes (i.e., expressing how a particular movement "feels" will be easier to represent in a textual instruction, while expressing that each hand should make a different movement will be easier to represent in an instruction with a static or dynamic visual).

In Chapter 5 we took a closer look at the speech modality and in particular we investigated quality differences between diphone synthesis, unit selection synthesis, and human speech. Although the segmental quality of diphone synthesis is in general inferior to that of unit selection synthesis, the suprasegmental quality of diphone synthesis is potentially better than that of unit selection synthesis. Thus, we investigated the trade-off between segmental quality on the one hand and suprasegmental quality on the other through the following research question:

- **How do quality differences within the speech modality influence its incremental processing and how can we assess these quality differences?**

An eye tracking experiment was conducted in which we used the visual world paradigm (e.g., Tanenhaus et al., 1995) to evaluate the incremental processing of diphone synthesis, unit selection synthesis, and human speech. Participants were presented with a visual display in which four objects were shown (see Figure 5.1 in Chapter 5). For every visual display, the participants were given two consecutive spoken instructions each referring to an object within the display. The first instruction was realized with a neutral accent pattern (e.g., *Kijk naar de roze vork*, Look at the pink fork). The second instruction either had a contextually appropriate (e.g., *Kijk nu naar de blauwe vos*, Now look at the blue fox) or inappropriate (e.g., *Kijk nu naar de blauwe vork*, Now look at the blue fork) double accent pattern. Note that an instruction with a contextually appropriate accent pattern mentioned a different object than the one mentioned in the first instruction. Conversely, an instruction with a contextually inappropriate accent pattern mentioned the same object as the first instruction. In addition, participants had to fill out a questionnaire on the intelligibility and the naturalness of the three speech conditions.

The research results showed that the differences in segmental intelligibility were reflected in the incremental processing of the three speech conditions: participants identified the mentioned target object most rapidly in the human speech condition (having the best segmental intelligibility) and least rapidly in the diphone synthesis condition (having the worst segmental intelligibility). The performance of the participants in the unit selection synthesis condition fell between these two. Also, differences in the suprasegmental intelligibility affected the incremental processing of the three speech conditions: when the second instruction had a contextually appropriate accent pattern (e.g., *Kijk nu naar de blauwe vos*, Now look at the blue fox) fixations to the competitor increased rapidly. Apparently, participants interpreted the accent on the adjective in the second instruction contrastively to adjective in the first instruction. Consequently, the participants expected that the second instruction would mention an object of the same type as the one mentioned in the first instruction. Moreover, there was an interaction between segmental and suprasegmental intelligibility: participants anticipated the upcoming target

mentioned in the second instruction, however, when the second instruction mentioned a different object type this anticipation was hard to overrule for diphone synthesis, and was easier to overrule for unit selection synthesis and human speech. Finally, the results of the questionnaire corresponded with eye-movement data. Human speech was rated most intelligible and most natural followed by unit selection synthesis and diphone synthesis. Thus, evaluating quality differences between synthetic and human speech using eye tracking provides us with a detailed insight in what the differences between diphone synthesis, unit selection synthesis, and human speech are and when they occur.

6.2 Discussion

In Chapter 1, we argued that presenting information in a multimodal way is not straightforward as the effectiveness of multimodal information presentation is the result of a complicated mixture of characteristics of communicative tasks, qualities of the presentation modes, characteristics of research methodology, and user characteristics. Each of these factors could be addressed in more detail in future research.

6.2.1 Characteristics of the task

The research presented in this thesis indicated that task characteristics can affect the type of (multimodal) information presentation. For example, the research presented in Chapter 2 showed that the type of information question (definition vs. procedural) one has, influences the type of visual occurring in the answer presentations: visuals with a low informative value occurred most often in definition questions whereas visuals with a high informative value occurred most often in procedural questions. The characteristics of a task not only refer to the task type, but also refer to the complexity of the task. For instance, in Chapter 4 we found that learning RSI exercises from a picture led to a good performance for easy exercises, but the performance dropped for the difficult exercises. It would be interesting to investigate the interplay between the type and the complexity of a task on the one hand and the type of (multimodal) information presentation on the other in future research.

6.2.2 Characteristics within the same information modality

In this thesis we have investigated the effectiveness of different information modalities as well as the effectiveness of differences within the same information modality. For instance, Chapter 5 illustrated that different speech qualities (i.e., segmental and suprasegmental qualities) influenced the incremental processing of the speech modality. Another example can be found in Chapter 2 in which the results of the production experiment showed that photographs often had a representational function and graphics often an informative function. Both visuals belong to the category of static pictures, but clearly they express information differently: photographs typically represent reality whereas graphics schematize it (Tversky et al., 2006). In general, it would be interesting to investigate to what extent differences within the same information modality influence their effectiveness.

6.2.3 Characteristics of the research methodology

In this thesis, different research methodologies were used to evaluate multimodal information presentations which can be categorized in various dimensions: qualitative and quantitative research approaches, online and offline research methods, and objective and subjective measures.

The characteristics of the research methodology can influence which aspects of an information presentation are investigated. For example, online research methods, such as eye tracking, enable us to investigate the incremental processing of a presentation mode. Offline research methods, like experimental evaluation, provide us with the end results of processing a presentation mode. Another example can be found in Chapter 2 in which we used both qualitative and quantitative research approaches to investigate the production and evaluation of multimodal information presentations. In the first qualitative study, we analyzed the functions of visuals in relation to text. The coding scheme (decorational < representational < informative) we formulated was based on whether the visual gave an answer to the medical question. However, we could also have looked at the relation between the visual and the textual answer which might have led to different results. In the second quantitative study, we analyzed the perceived informativity and attractiveness of

unimodal and multimodal answer presentations. However, different dependent measures, like study times, might have given us a different insight in how people perceive unimodal and multimodal answer presentations. Both qualitative and quantitative research methods have their own advantages and disadvantages, but when complementing each other (like in Chapter 2) they provide us with useful insight in multimodal information presentations.

Also, the characteristics of the dependent variables can reveal differences in the observed effectiveness of a multimodal information presentation. In Chapter 4, we have seen that each dependent variable (e.g., learning times and number of correctly executed exercises) shed a different light on the effectiveness of the information modalities under investigation (e.g., while an instruction in a text led to the longest learning and execution times, learning from an instruction in text led to a fairly good learning performance). Also, the type of dependent variable can influence the effectiveness of the information modalities. In Chapter 4, we found that no single presentation mode outperformed the others on all objective dependent variables. However, participants preferred film clips to learn RSI exercises. What causes this apparent discrepancy between the effectiveness of information modalities and the subjective satisfaction is not entirely clear. Arguably, film clips are more 'visually appealing' than pictures. Moreover, it may be that participants recognize that film clips offer a complete action representation, but do not realize that learning from text or a picture may lead to good results as well. This discrepancy between the effectiveness of information modalities and the subjective learning preference could be investigated in more detail in future research.

6.2.4 Characteristics of the user

A limitation of the research presented in this thesis could be that user characteristics were not taken into account. Mayer (2001) argues that individual differences could influence the effectiveness of multimodal information presentations. User characteristics, such as prior knowledge, spatial ability, and learning preferences might all play a role in the effectiveness of (multimodal) information presentation. For example, research has shown that when listeners are trained in synthetic speech their performance on recognizing words produced in synthetic speech improves

(Swab, Nusbaum & Pisoni, 1985). Another example comes from research on individual differences in hypertext use which has indicated that users with high spatial ability interact more efficiently with a hypertext than users with lower spatial ability (e.g., Campagnoni & Ehrlich, 1989; Vincente & Williges, 1988). Arguably, it is possible that there is a difference between users with a high and low spatial ability in how they conceptualize their actions when navigating a website. In sum, future research on the effectiveness of multimodal information presentations could take user characteristics into account. For example, it could be possible that people who are highly knowledgeable on medical topics, like medical students, produce and evaluate multimodal medical answer presentationss differently than users who are not highly knowledgeable on medical topics, like the participants in our experiment.

6.3 Studying multimodal information presentation: pitfalls and caveats

6.3.1 Comparing apples and oranges?

When comparing different information modalities, it is important that they offer comparable information (Tversky et al., 2002). However, it turned out that this was not always as straightforward as it may seem. For example, a picture can not express how a certain movement feels (e.g., "spread your fingers until a mild stretch between the fingers is felt"). Moreover, a static picture combined with an arrow indicating the motion does not make the entire intended movement as explicit as in film clip. Therefore, one could argue that comparing the effectiveness of different information modalities is comparing apples and oranges. However, research by Sandford (1995) showed that this is possible and has interesting results. Tversky et al. (2002) are right when they argue that when comparing different information modalities, it is important that they offer comparable information. However, some information can not be 'translated' from one information modality to another. Therefore it might be an illusion to think that the same amount of information can be presented in different information modalities.

6.3.2 The redundancy of multimodal information presentations

When presenting information in a multimodal way, it is possible that the same information is presented in different modalities which results in redundancy. Figure 1.1 in Chapter 1 gives an example of a multimodal information presentation in which redundancy occurs. In this answer presentation both textual and pictorial representations are used to explain how a workspace can be ergonomically organized. However, presenting redundant information may interfere with learning. According to the cognitive load theory (Sweller & Chandler, 1991) processing redundant information increases the working memory load, which interferes with the information transfer to the long-term memory.

Figure 2.5 in Chapter 2 gives an example of an answer presentation to a procedural question in which the text and the (informative) visual explain how to organize an ergonomical workspace. It can be argued that the answer presentation illustrated in Figure 2.5 contains redundant information which decrements learning. However, users could also benefit from a textual and visual representation of organizing a ergonomical workspace. The evaluation experiment discussed in Chapter 2 showed that answer presentations were evaluated as more informative when they contained an informative visual.

The redundancy principle as Mayer (2001) formulated it, may need some refining. For example, the type of information presentation (procedures vs. definitions) could influence the amount of redundancy between the textual and visual answer presentation. It would be interesting to investigate to what extent the amount of redundancy affects learning in different types of answer presentation.

References

Agrawala, M., & Stolte, C. (2001). Rendering effective route maps: Improving usability through generalization. *Proceedings of SIGGRAPH 2001,* 241-250.

Ahuja, S.J. & Webster, J. (2001) Perceived disorientation an examination of a new measure to assess web design effectiveness, *Interacting with Computers*, *14***,** 15-29.

Ainsworth, S.E. & Van Labeke (2004). Multiple forms of dynamic representation. *Learning and Instruction, 14*, 241-255

Altmann, G.T., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp 347-386). New York: Psychology Press.

André, E. (2000). The generation of multimedia presentations. In R. Dale, H. Moisl, and H. Somers (Eds.), *A handbook of natural language processing: techniques and applications for the processing of language as text* (pp. 305-327). New York: Marcel Dekker Inc.

Arens, Y., Hovy, E. & Vossers, M. (1993). On the knowledge underlying multimedia presentations. In M. T. Maybury (Ed.), *Intelligent Multimedia Interfaces* (pp. 280-306). Menlo Park: AAAI Press.

Baddely, A. (1992). Working memory. *Science, 255,* 556-559.

Balci, R., & Aghazadeh, F. (2003). The effect of work-rest schedules and type of task on the discomfort and performance of VDT users. *Ergonomics, 46*, 455-465.

Bateman, J.A., Kamps, T., Kleinz, J., & Reichenberger, K. (2001). Constructive text, diagram and layout generation for information presentation: the DArt$_{bio}$ system. *Computational Linguistics*, *27*, 409-449.

Bernsen, N. (1994). Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers*. 6, 347-371.

Bétrancourt, M., & Tversky, B. (2000). Effects of computer animation on user's performance: a review. *Le travail humain, 63*, 311-329.

Black, A.W., Taylor, P., & Caley. R., (2002). The Festival Speech Synthesis System, System documentation. Centre for Speech Technology Research University of Edinburgh.

Boechler, P. M. (2001). How spatial is hyperspace? Interacting with hypertext documents: cognitive processes and concepts. *CyberPsychology & Behavior*, *4*, 23-46.

Boekelder, A. & Steehouder, M. (1998). Selecting and switching: some advantages of diagrams for presenting instructions. *IEEE Transactions on Professional Communication*, *41*, 229-241.

Boersma, P., & Weenink, D. (1996). Praat, a system for doing phonetics by computer, version 3.4. Institute of Phonetic Sciences of the University of Amsterdam, Report 132.

Bosma, W. (2005). Extending answers using discourse structure. In H. Saggion and J. L. Minel (Eds.), *Bulgaria Borovets RANLP Workshop on crossing barriers in text summarization research*. (pp. 2-9). Borovets: Incoma Ltd.

Brünken, R., Plass, L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning, *Educational Psychologist, 38*, 53-61.

Campagnoni, F.R., & Ehrlich, K. (1989). Information retrieval using a hypertext-based help system. *ACM transactions on information systems, 7*, 271-291.

Carney, R., & Levin, J. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, *14*, 5-26.

Carroll, L., & Wiebe, E. (2004). Static versus dynamic presentation of procedural instruction: Investigating the efficacy of video-based delivery. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Chen, A., Os, E., den, Ruiter, J. P., de. (2007). Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech. *Linguistic Review*, *24*, 317-344.

Chen, C., & Rada, R. (1996). Interacting with hypertext: a meta-analysis of experimental studies. *Human-Computer Interaction, 11*, 125-156.

Cheng, P. (2002). Electrifying diagrams for learning: principles for complex representational systems. *Cognitive Science, 26*, 658-736.

Chi, M., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In Sternberg, R. (Eds.), *Advances in the Psychology of Human Intelligence* (pp. 7-75). Hillsdale, NJ: Erlbaum.

Conklin, J. (1987). Hypertext: an introduction and survey. *IEEE Computer, 20*, 17-41.

Dahan, D., & Tanenhaus, M.K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 498-513.

Dahan, D., Tanenhaus, M.K., & Chambers, C.G., (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*, 292-314.

Danielson, D.R. (2002). Web navigation and the behavioural effects of constantly visible site maps. *Interacting with Computers, 14*, 601-618.

Dee-Lucas, D., & Larkin, J. H. (1995). Learning form electronic texts: effects of interactive overviews for information access. *Cognition and Instruction, 13*, 431-468.

Dias, P., & Sousa, A.P. (1997). Understanding navigation and disorientation in hypermedia learning environments. *Journal of Educational Multimedia and Hypermedia, 6*, 173-185.

Dillon, A. (2004). Designing usable electronic text (2nd ed.). London: CRC Press.

Downs, R.M., & Stea, D. (1973). Cognitive maps and spatial behaviour: process and products. In R. M. Downs & D. Stea (Eds*.), Image and Environment* (pp. 8-27). Chicago: Aldine Publishing.

Downs, R. M., & Stea, D. (1977). *Maps in minds: reflections on cognitive mapping.* London: Harper & Row.

Dumas, J. & Redish, J. (1993). *A Practical Guide to Usability Testing*, Ablex, Norwood, NJ.

Edwards, D.M., & Hardman, L. (1989). "Lost in hyperspace": cognitive mapping and navigation in a hypertext environment. In R. MacAleese (Eds.), *Hypertext: theory into practice* (pp. 90-105). Exeter: Intellect.

Elm, E.C., & Woods, D.D. (1985). Getting lost: a case study in interface design. *Proceedings of the 29th Annual Meeting of the Human Factor Society* (pp. 927 - 931). Santa Monica, CA: Human Factors Society

Ericsson, K., & Simon, H. (1993). *Protocol analysis: verbal reports as data.* Cambridge: MITT Press.

Farris, J.S., Jones, K.S., & Elgin, P.D. (2002). Users' schemata of hypermedia: what is so "spatial" about a website? *Interacting with Computers, 14*, 487-502.

Fletcher, C.R., & Chrysler, S.T. (1990). Surface forms, textbases, and situation models: recognition memory for three types of textual information. *Discourse Processes, 13*, 175-190.

Ganier, F. (2004). Factors affecting the processing of procedural instructions: implications for document design. *IEEE Transactions on Professional Communication*, *47*, 15-26.

Gellevij, M., Meij, H., van der, Jong, T., de, & Pieters J. (1999). The effects of screen captures in manuals: a textual and two visual manuals compared. *IEEE Transactions on Professional Communication*, *42*, 277-291.

Gentner, D., & Boroditsky, L. (2001). Individuation, relational relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), Language acquisition and conceptual development. Cambridge, England: Cambridge University Press.

Glenberg, A. (1997). What memory is for. *Behavioural and Brain Sciences, 20*, 1-19.

Glenberg, A., & Robertson, D. (1999). Indexical understanding of instructions. *Discourse Processes, 28*, 1-26.

Gibbs, R.W.J. (2005). Embodiment in metaphorical imagination. In D. Pecher & R. A. Zwaan (Eds.), *Grounding Cognition: The role of perception and action in memory, language and thinking* (pp. 65-92). Cambridge: Cambridge University Press.

Gibson, J. (1979). *The ecological approach to visual perception.* Hillsdale: Lawrence Erlbaum Associates.

Golledge, R.G. (1999). Human wayfinding and cognitive maps. In R. G. Golledge (Eds.), *Wayfinding behaviour: cognitive mapping and other spatial processes* (pp. 5-45). Baltimore: John Hopkins University Press.

Gupta, M., & Gramopadhye, A.K. (1995). An evaluation of different navigational tools in using hypertext. *Computers and Industrial Engineering, 29*, 437-441.

Gussenhoven, C., & Rietveld T. (1992). A target-interpolation model for the intonation of Dutch. *Proceedings of the ICSLP*, Banff, Canada, 1235-1238.

Heller, R., Martin, C., Haneef, N., & Gievka-Krliu, S. (2001). Using a theoretical multimedia taxonomy framework. *ACM Journal of Educational Resources in Computing. 1,* 1-22.

Hegarty, M. (2004). Dynamic visualizations and learning: getting to the difficult questions. *Learning and Instruction, 14*, 343-351.

Heiser, J., Phan, D., Agrawala, M., Tversky, B., & Hanrahan, P. (2004). Identification and validation of cognitive design principles for automated generation of assembly instructions. *Proceedings of Advanced Visual Interfaces*, 311-319.

Hofman, R., & Oostendorp, H., van. (1999). Cognitive effects of a structural overview in hypertext. *British Journal of Educational Technology*, *30*, 129-140.

Hooijdonk, C.M.J., van & Krahmer, E.J. (2006). De invloed van unimodale en multimodale instructies op de effectiviteit van RSI-preventieoefeningen [The influence of unimodal and multimodal instructions on the effectiveness of RSI prevention exercises]. *Tijdschrift voor Taalbeheersing*, *28*, 73-87.

Hooijdonk, C.M.J., van & Krahmer, E.J. (In press). Information modalities for procedural instructions: The influence of text, pictures, and film clips on learning and executing RSI exercises. *IEEE Transactions on Professional Communication.*

House, A.S., Williams, C.E., Hecker, M.H., & Kryter, K.D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *Journal of the American Statistical Association, 37*, 158-166.

Johnson-Laird, P.N. (1983). *Mental models*. Cambridge: Cambridge University Press.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M.R. Key (Eds.), *The relationship of verbal and nonverbal communication* (pp. 207-227). The Hague: Mouton.

Kim, J. (1999). An empirical study of navigation aids in customer interfaces. *Behaviour & Information Technology, 18*, 213-224.

Kintsch, W., & Dijk, T. A., van. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Kirby, J., Moore, P. & Schofield, N. (1988). Verbal and visual learning styles, *Contemporary Educational Psychology, 13*, 169-184.

Kostelnick, C., & Roberts, D. (1998). *Designing visual language. Strategies for professional communicators*. Boston: Allyn and Bacon.

Knapp, M.L. (1978). *Nonverbal communication in human interaction*. New York: Holt, Rinehart and Winston.

Krahmer, E.J., & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*, *47*, 105-117.

Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Beverly Hills: Sage Publications.

Krull, R., D'Souza, S., Roy, D., & Sharp, D. (2004). Designing procedural illustrations: Special issue on acquiring procedural knowledge of a technology interface, *IEEE Transactions on Professional Communication*, *47*, 27-33.

Kuhn, W. (1996). Handling Data Spatially: Spatializing User Interfaces. In M.J. Kraak & M. Molenaar (Eds.), *Proceedings of 7th International Symposium on Spatial Data Handling (*pp: 13B.1 - 13B.23)*,* Delft, The Netherlands: IGU.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Larkin, J. & Simon H. (1987). Why a diagram is (sometimes) worth a thousand words, *Cognitive Science, 11,* 65-99.

Leutner, D. & Plass, J. (1998). Measuring learning styles with questionnaires versus direct observation of preferential choice behaviour in authentic learning situations: The Visualizer/Verbalizer Behaviour Observation Scale (VV-BOS), *Computers in Human Behaviour, 14*, 543-557.

Lewalter, D. (2003). Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction, 13*, 177-189.

Lida, B., Hull, S., & Pilcher, K. (2003, February). Breadcrumb navigation: a exploratory study of usage Usability News, 5. Retrieved 13 May, 2004, from http://psychology.wichita.edu/surl/usabilitynews/51/breadcrumb.htm

Lowe, R. (2004). Interrogation of a dynamic visualization during learning. *Learning and Instruction, 14*, 257-274.

Luce, P., Feustel, T., & Pisoni, D. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, *25*, 17-32.

Maes, A. (2005). Een multimodale kijk op informatie [A multimodal view on information]. In H. Van Driel (Ed.), *Digitaal Communiceren* (pp. 219-258). Amsterdam: Boom.

Maes, A., Geel, A., van, & Cozijn, R. (2006). Signposts on the digital highway. The effect of second level and third level hyperlink previews. *Interacting with Computers, 18*, 265-282

Maes, A., & Lenting, H. (1999). How to put the instructive space into words? *IEEE Transactions on Professional Communication*, *42,* 100-113

Maglio, P., & Matlock, T. (2003). The conceptual structure of information space. In K. Höök & D. Benyon & A. J. Munro (Eds.), *Designing information spaces: the social navigation approach* (pp. 385-404). London: Springer.

Marchionini, G. (1989). Information seeking strategies of novices using a full text electronic encyclopedia. *Journal of the American Society for Information Science, 40*, 54-66.

Marchionini, G., & Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer, 21,* 70-80.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology*, *88*, 49-62.

Matin, E., Shao, K. & Boff, K. (1993). Saccadic overhead: information processing time with and without saccades, *Perceptual Psychophysics*, *53*, 372-380.

Marsh, E. & White, M. (2003). Taxonomy of relationships between images and text. *Journal of Documentation*, *59*, 647-672.

Maybury, M.T. (1993). *Intelligent multimedia interfaces*. Menlo Park, CA: AAAI Press.

Maybury, M., & Lee, J. (2000). Multimedia and multimodal interaction structure. In M. Taylor, F. Néel, & D. Bouwhuis (Eds.), *The structure of multimodal dialogue II* (pp. 295-308). Amsterdam: John Benjamins.

Mayer, R. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology*, *81*, 240-246.

Mayer, R. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction, 13*, 125-139.

Mayer, R. (2005). *The Cambridge handbook of multimedia learning*. Cambridge: Cambridge University Press.

Mayer, R., & Gallini, J. (1990). When is an illustration worth a thousand words? *Journal of Educational Psychology*, *82*, 715-726.

Mayer, R. & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90,* 312-320

Mayer, R., & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning & Instruction*, *12*, 107-119.

Mayer, R., Sobko, K., & Mautone, P. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*, 419-425.

McDonald, S., & Stevenson, R. J. (1999). Spatial versus conceptual maps as learning tools in hypertext. *Journal of Educational Multimedia and Hypermedia, 8*, 43-64.

Mclean, L., Tingley, M., Scott, R., & Rickards, J. (2001). Computer terminal work and the benefit of micro-breaks. *Applied Ergonomics, 32*, 225-237.

Merriënboer, J., van, Schuurman, J., Croock, M., de, & Paas, F. (2002). Redirecting learners' attention during training: effects on cognitive load, transfer test, performance and training efficiency. *Learning and Instruction, 12,* 11-37.

Michas, I., & Berry, D. (2000). Learning a procedural task: effectiveness of multimedia presentations. *Applied Cognitive Psychology*, *14*, 555-575.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Moreno, R., & Mayer, R. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*, 358-368.

Newell, A., & Simon, H. (1972). *Human Problem Solving*, Englewood Cliffs, NJ: Prentice Hall.

Nielsen, J. (1993). *Usability Engineering*, Cambridge MA: Academic Press

Nooteboom, S.G., & Kruyt, J.G. (1987). Accent, focus distribution, and perceived distribution of given and new information: An experiment. *Journal of the American Statistical Association*, *82*, 1512-1524.

Norman, D. A. (1998). *The Design of everyday things*. London: The MITT Press.

Nusbaum, H., Francis, A., & Henly, A. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 1,* 7-19

Nusbaum, H., & Pisoni, D. (1985).Some constraints on the perception of synthetic speech, *Behavior Research Methods, Instruments, & Computers*, *17*, 235-242

Otter, M., & Johnson, H. (2000). Lost in hyperspace: metrics and mental models. *Interacting with Computers (13)*, 1-40.

Oviatt, S. (1999). Ten myths of multimodal interaction. Communications of the ACM, *42*, 74-81.

Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., et al. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction. *Proceedings of the 5th International Conference on Multimodal Interfaces*, Vancouver, Canada, 2003, 44-51.

Pane, J., Corbett, A., & John, B. (1996). Assessing dynamics in computer-based instruction. *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, Vancouver, CA, 197- 204

Paris, R., Thomas, M., Gilson, R., & Kincaid, J.(2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors*, *42*, 421-431.

Park, O., & Gittelman, S. (1995). Dynamic characteristics of mental models and dynamic visual displays. *Journal of Instructional Science*, *23*, 303-320.

Park, O., & Hopkins, R. (1993). Instructional conditions for using dynamic visual displays: a review. *Journal of Instructional Science, 21*, 427-449.

Perez, E., & White, M. (1985). Student evaluation of motivational and learning attributes of microcomputer software. *Journal of Computer-Based Instruction, 12*, 39-43.

Pisoni, D.B. (1987). Some measures of intelligibility and comprehension. In J. Allen, M.S. Hunnicutt, & D.H. Klatt (Eds.), *From Text to Speech: the MITalk System* (pp.151-171). Cambridge: Cambridge University Press.

Ploetzner, R., & Lowe, R. (2004). Dynamic visualisations and learning. *Learning and Instruction, 14*, 235-240.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Reed, S. (1985). Effects of computer graphics on improving estimates to algebra word problems. *Journal of Educational Psychology, 77*, 285-298.

Reynolds, M.E., & Givens, J. (2001). Presentation rate in comprehension of natural and synthesized speech. *Perceptual and Motor Skills, 92*, 958–968.

Rieber, L. (1991). Animation, incidental learning, and continuing motivation. *Journal of Educational Psychology, 83,* 318-328.

Rietveld, T., & Van Hout. R. (1993). *Statistical techniques for the study of language and language behaviour.* Berlin: Mouton de Gruyter.

Roberts, M.J. (2005). *Underground maps after Beck.* Harrow: Capital Transport Publishing

Sanderman, A.A., & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, *40*, 391-409.

Sandford, S.A. (1995). On the comparison of apples and oranges. *Annals of Improbable Research*, *1*, 2-3.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett & S. Greenspan (Eds.), *Applied Speech Technology* (pp. 194-231). Boca Raton: CRC.

Schnotz, W., Böckheler, J., & Grzondziel, H. (1999). Individual and co-operative learning with interactive animated pictures. *European Journal of Psychology of Education*, *14*, 245-265.

Schnotz, W., Picard, E., & Hron, A. (1993). How do successful and unsuccessful learners use text and graphics?, *Special issue of Learning and Instruction*, 181-199.

Schnotz, W., & Rasch, T. (2005), Enabling, facilitating, and inhibiting effects of animations in multimedia learning: Why reduction of cognitive load can have negative results on learning. *Educational Technology Research & Development, 53,* 47-58.

Schwab, E.C., Nusbaum, H.C., & Pisoni, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, *27*, 395-408.

Schwan, S., & Riempp, R. (2004). The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning & Instruction*, *14*, 293-305.

Shneiderman, B. (1998). *Designing the user interface: strategies for effective human-computer interaction*. 3rd ed. Reading, Mass: Addison-Wesley.

Shum, S. (1990). Real and virtual spaces: mapping from spatial cognition to hypertext. *Hypermedia, 2*, 133-158.

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 174-215.

Stone, W. (1983). Repetitive strain injuries. *Medical Journal of Australia 10*, 616-618.

Sutcliffe, A. (1997). Task-related information analysis. *Int. Journal of Human Computer Studies*, *47*, 223-257.

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction, 8,* 351-362.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, 12(3), 185 - 223.

Sweller, J., Merriënboer, J., van, Paas, F. (1998), Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296.

Swift, M.D., Campana, E., Allen, J.F., & Tanenhaus, M.K. (2002). Monitoring eye movements as an evaluation of synthesized speech. *Proceedings of the IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA, 19-22.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language and Cognitive Processes, 11,* 583–588.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Taylor, H. A., & Tversky, B. (1992a). Descriptions and depictions of environments. *Memory & Cognition, 20*, 483-496.

Taylor, H. A., & Tversky, B. (1992b). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language, 31*, 261-292.

Terken, J., & Nooteboom, S.G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information, *Language and Cognitive Processes*, *2*, 145-163.

Theune, M., Schooten, B., van, Akker, R., op den, Bosma, W., Hofs, D., Nijholt, A., et al. (2007). Questions, pictures, answers: introducing pictures in question-answering systems. In L. Ruiz Miyarez, A. Munoz Alvarado & C. Alvarez Moreno (Eds.), ACTAS-1 of X Symposio Internacional de Comunicacion Social, Santiago de Cuba, 450-463,

Tindall-Ford, S., Chandler, P., and Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied, 3*, 257–287.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79-112).Cambridge, MA: The MIT Press.

Tversky, B. (2003). Structures of mental spaces: how people think about space. *Environment and Behaviour, 35*, 66-80.

Tversky, B., Agrawala, M., Heiser, J., Lee, P., Hanrahan, P., Phan, D., Stolte, C., & Daniel, MP. (2006). Cognitive design principles for automated generation of visualizations. In G.L. Allen (Ed.), Applied spatial cognition : from research to cognitive technology (pp. 53-75). New York :Lawrence Erlbaum Associates

Tversky, B., Morrison, J., & Bétrancourt, M. (2002). Animation; can it facilitate? *Int. J. Human-Computer Studies, 57*, 247-262.

Tversky, B., Zacks, J., Lee, P., & Heiser, J. (2000). Lines, blobs, crosses, and arrows: diagrammatic communication with schematic figures. In M. Anderson, P. Cheng, V. Haarslev (Eds.), *Theory and application of diagrams* (pp. 221-230). Berlin: Springer.

Twyman, M. (1987). A schema for the study of graphic language. In O. Boyd-Barrett & Braham, P. (Eds.), *Media, Knowledge, and Power* (pp. 201-225). London: Croom Helm (Reprinted from *Processing of Visible Language*, pp.117-150, by P.A. Kohlers, M.E. Wrolstad & H. Bouma (Eds.), 1979, New York: Plenum Press.

Veenker, T. (2005). *WWStim: A CGI script for presenting web-based questionnaires and experiments*. Website:
http://www.let.uu.nl/Theo.Veenker/personal/projects/wwstim/doc/en/

Vincente, K.J., & Williges, R.C. (1988). Accommodating individual differences in searching a hierarchical file system. *International Journal Man Machine Studies, 29*, 647-668.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*, 367-392.

Weidenmann, B. (1988). When good pictures fail. An information processing approach to the effect of illustrations. In H. Mandl, J. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 157-171.), Amsterdam: Elsevier.

Weiss, R., Knolton, D., & Morrison, G. (2002). Principles for using animation in computer-based instruction: theoretical heuristics for effective design. *Computers in Human Behaviour, 18,* 465-477.

Wijk, C., van, (2000). *Toetsende statistiek : basistechnieken. Een praktijkgerichte inleiding voor onderzoekers van taal, gedrag en communicatie*. Bussum: Coutinho

Williams, T., & Harkus, D. (1998). Editing visual media. *IEEE Transactions on Professional Communication*, *41*, pp. 33-45.

Williams, T., Smith, L, & Herrick, R. (1989). Exercise as a prophylactic device against carpal tunnel syndrome. In *Proceedings of the 33rd Annual Meeting of the Human Factors Society*, Santa Monica, CA, 723-727.

Winn, W. (1989). The design and use of instructional graphics. In H. Mandl, J. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 125-144), Amsterdam: Elsevier.

Summary

This dissertation attempts to contribute to our understanding of the production, processing, and evaluation of multimodal information presentations. There are reasons to believe that in some cases presenting information using multiple modalities is more effective than presenting information using a single modality. However, presenting information in a multimodal way implicates a complicated mixture of characteristics of communicative tasks and goals, characteristics of sensory modalities, and qualities of presentation modes. Moreover, evaluating multimodal information presentations can be done using different research methodologies.

In this thesis, four exploratory studies are discussed each addressing different research areas, including human-computer interaction, web usability, instructional psychology, and speech technology, with a specific focus on multimodal information presentation. Moreover, different research methodologies were used to evaluate multimodal information presentations, ranging from eye tracking and protocol analysis, to corpus research and experimental evaluation studies.

Chapter 2 described a production experiment that was carried out to determine which modalities people choose to answer different types of questions. In this experiment, participants had to create (potentially multimodal) presentations of answers to general medical questions. In total 1775 answer presentations were collected. The collected corpus was coded on the presence of visual media (i.e., photos, graphics, and animations) and their function.

The results showed that almost one in four answers contained one or more visual media. Moreover, the design of the answer presentations was affected by the answer length: visuals with a high informative value occurred more often in brief answers while visuals with a lower informative value occurred more often in extended answers. Arguably, it is likely that a visual added less information to the textual answer when the answer contains much text and vice versa. Also, the question type influenced the design of the answer presentations: visuals with a low informative value were more frequent in the answers of definition questions, whereas visuals with a high informative value were more frequent in the answers of procedural questions. A possible explanation for this result could be that the textual answers to definition

questions often explained an element of the question, which was represented with a visual. Visuals in the answers of procedural questions were often used to explain the steps within the procedure and therefore added information to the textual answer.

Next, Chapter 2 described an evaluation experiment that concentrates on how users evaluate unimodal and multimodal answer presentations. The participants had to assess the informativity and attractiveness of answer presentations for different types of medical questions. These answer presentations, originating from the production experiment, were manipulated in their answer length (brief vs. extended) and their type of visuals (i.e., visuals with a low or high informative value). In a post-test, participants they had to indicate how much they had recalled from the presented answer presentations.

The results showed that answer presentations having a visual with a high informative value were evaluated as most informative and most attractive. The results for the post-test suggested that learning from answer presentations with informative visuals led to a better learning performance than learning from purely textual answer presentations, although the differences were not statistically significant.

Chapter 3 described an explorative thinking aloud study that investigated how users verbalize their actions when navigating in a hypertext. Moreover, we studied which actions were expressed in spatial terms. Ten thinking aloud protocols were collected that originated from two different usability studies. In both studies, users were asked to perform simple search tasks in a hypertext (i.e., looking up the answers to factual questions) and to think aloud while executing these tasks. The total corpus consisted of 694 coded segments which were analyzed on the types of actions and the levels of actions users were involved in when navigating a web site. We distinguished two action types: executions and evaluations. Moreover, each action type could be described in three action levels: first, second, and third level. Also, we investigated which action types and actions levels were expressed in spatial terms (e.g., "I am going back to the homepage").

The results of the protocol analysis showed that verbalizations were mostly referring to evaluations (e.g., "I cannot click on this item"). For the levels of actions, it was found that verbalizations referring to the first action level occurred most often (e.g., "I am double clicking on this object"). Moreover, the results indicated that

spatial expressions were most frequent when users described executions on the first action level (e.g., "I am going back to the homepage"). In general, the research results confirmed that people use spatial expressions when navigating a web site. However, the difference between a spatial and a non-spatial expression was not always clear.

In Chapter 4, we first described an experiment studying a specific kind of procedural instructions, namely exercises for the prevention of Repetitive Strain Injury (RSI), taking information modality (text vs. picture vs. film clip) and difficulty degree of the exercises (easy vs. difficult) into account. In the experiment, participants had to learn ten easy (simple symmetrical movements) and ten difficult (complex symmetrical movements or asymmetrical movements) RSI exercises and were asked to execute them. The influence of presenting an instruction in text, picture, or film clip was measured through learning times, amount of practicing during learning, execution times, and number of correctly executed exercises. Participants were also asked for their subjective satisfaction.

The results showed that no single modality outperformed the others on all dependent variables. Also, results for the subjective satisfaction of the participants revealed no differences between the three information modalities.

Next, Chapter 4 describes a preference study that investigated which presentation mode (i.e., text vs. picture vs. film clip) people prefer when learning RSI exercises. In this second experiment, participants had to study six RSI exercises in three versions (i.e., text, picture, and film clip), after which they had to indicate (by forced choice) which of the three presentation modes they preferred for a particular exercise. The results showed that overall participants preferred film clips to learn RSI exercises. However, for some exercises (e.g., "make fists") it was found that users preferred an instruction in text to an instruction in a picture and film clip.

Chapter 5, finally, described an eye tracking experiment that was conducted to study the incremental processing of diphone synthesis, unit selection synthesis, and human speech taking segmental and suprasegmental speech quality into account. Fifteen pairs of Dutch monosyllabic picturable nouns were used as stimuli. These nouns shared the same initial phonemes (e.g., *vork - vos*, fork - fox). The instructions were realized in three speech conditions, i.e., diphone synthesis, unit selection synthesis,

and human speech. The diphone stimuli were produced with a Dutch TTS system based on the Festival TTS system. The unit selection stimuli were obtained through a commercially available synthesizer. The human speech stimuli were recorded by a native speaker of Dutch.

In the experiment, participants were presented with a visual display in which four objects were shown. For every visual display, the participants were given two consecutive spoken instructions each referring to an object within the display. The first instruction was mentioned the *referent* and was realized with a neutral accent pattern (e.g., *Kijk naar de roze vork*, Look at the pink fork). The second instruction mentioned the *target* and was either realized with a contextually appropriate double accent pattern (e.g., *Kijk nu naar de BLAUWE VOS*, Now look at the BLUE FOX) or contextually inappropriate double accent pattern (e.g., *Kijk nu naar de BLAUWE VORK*, Now look at the BLUE FORK). In addition, participants had to fill out a questionnaire on the intelligibility and the naturalness of the three speech conditions.

The results showed that participants identified the target most rapidly in the human speech condition (having the best segmental intelligibility) and least rapidly in the diphone synthesis condition (having the worst segmental intelligibility). The performance of unit selection synthesis fell between these two. We also found that when the second instruction had a contextually appropriate accent pattern (e.g., *Kijk nu naar de BLAUWE VOS*, Now look at the BLUE FOX) fixations to the competitor (i.e., the blue fork) increased rapidly. Apparently, participants interpreted the accent on the adjective in the second instruction contrastively to adjective mentioned in the first instruction. This implies that participants anticipated the upcoming target mentioned in the second instruction. Moreover, we found that this anticipation was hard to overrule for diphone synthesis, but easier to overrule for unit selection synthesis and human speech. Finally, the results of the questionnaire corresponded with eye-movement data. Human speech was rated most intelligible and most natural followed by unit selection synthesis and diphone synthesis.

Chapter 6 presented the main results of the four studies and ended with some final remarks.

When comparing different information modalities, it is important that they offer comparable information. However, it turned out that this was not always as

straightforward as it may seem. For example, while it is possible to express how a certain movement feels in a textual instruction, this is not possible in an instruction with a visual. This implies that some information can not be 'translated' from one information modality to another. Therefore it might be an illusion to think that the same amount of information can be presented in different information modalities.

When presenting information in a multimodal way, it is possible that the same information is presented in different modalities which results in redundancy. However, presenting redundant information may interfere with learning. The research discussed in Chapter 2 indicated that people present information in multimodal way (i.e., the use a combination of text and visuals). This visual could several functions: it could be merely decorative, it could represent an element mentioned in the textual answer, or it could add information to the textual answer. Thus, presenting information in a multimodal way, also implies a certain amount of redundancy in the information presentation. However, it remains unclear to what extent the amount of redundancy affects learning.

Samenvatting

Dit proefschrift gaat over multimodale informatiepresentatie en levert een bijdrage aan onze kennis over de productie, verwerking en evaluatie ervan. Er zijn redenen om aan te nemen dat in sommige gevallen het presenteren van informatie met meerdere modaliteiten effectiever is dan het presenteren van informatie met slechts één modaliteit. Echter, het presenteren van multimodale informatie impliceert een gecompliceerde mix van eigenschappen van communicatieve taken en doelen, eigenschappen van zintuiglijke modaliteiten, en kwaliteiten van informatiemodaliteiten zelf. Daarnaast kunnen multimodale informatiepresentaties met verschillende onderzoeksmethodes geëvalueerd worden.

In dit proefschrift worden vier verkennende studies besproken, waarbij ieder studie een ander onderzoeksgebied belicht, namelijk: mens-computer interactie, website usability, leerpsychologie en spraaktechnologie. Ieder hoofdstuk heeft hierdoor een specifieke kijk op multimodale informatiepresentatie. Daarnaast zijn in de vier studies verschillende onderzoeksmethodes toegepast om multimodale informatiepresentaties te evalueren, variërend van oogbewegingsregistratie tot protocol analyse en van corpus onderzoek tot experimentele evaluatiestudies.

Hoofdstuk 2 beschrijft een productie-experiment dat werd uitgevoerd om te bepalen welke modaliteiten gebruikers kiezen om verschillende soorten medische vragen te beantwoorden. In het experiment werd aan de proefpersonen gevraagd om (potentieel multimodale) antwoordpresentaties te creëren op algemene medische vragen. In totaal werden er 1775 antwoordpresentaties verzameld. Vervolgens werd het verzamelde corpus geanalyseerd op de aanwezigheid van afbeeldingen (bijv. foto's, lijntekeningen en animaties) en hun functie. De resultaten toonden aan dat één op de vier antwoorden één of meerdere afbeeldingen bevatte. Daarnaast werd het ontwerp van de antwoordpresentaties beïnvloed door de lengte van het antwoord: afbeeldingen met een hoog informatiegehalte kwamen vaker voor in korte antwoorden terwijl afbeeldingen met een laag informatiegehalte vaker voorkwamen in lange antwoorden. Een verklaring voor dit resultaat kan zijn dat een afbeelding minder informatie toevoegt aan een antwoord naarmate het antwoord zelf meer tekst bevat en vice versa. Ook het vraagtype had een effect op het ontwerp

van de antwoordpresentaties: afbeeldingen met een laag informatiegehalte kwamen vaker voor in antwoorden op definitievragen terwijl afbeeldingen met een hoog informatiegehalte vaker voorkwamen in antwoorden op procedurele vragen. Een mogelijke verklaring voor dit resultaat kan zijn dat illustraties in definitievragen vaak niet meer doen dan illustreren wat in tekst al wordt uitgelegd terwijl afbeeldingen in procedurele antwoorden vaak gebruikt worden om de stappen in een proces uit te leggen, waardoor ze informatie toevoegen aan het tekstuele antwoord.

Vervolgens beschrijft Hoofdstuk 2 een evaluatie-experiment waarin gebruikers unimodale en multimodale antwoordpresentaties beoordeelden. De antwoordpresentaties waren afkomstig uit het productie-experiment en werden gemanipuleerd in antwoordlengte (kort versus lang) en in het type afbeelding dat in het antwoord voorkwam (afbeeldingen met een hoog of laag informatiegehalte). De proefpersonen moesten de informativiteit en aantrekkelijkheid van antwoordpresentaties op verschillende medische vraagtypes beoordelen. Daarnaast moesten ze in een posttest aangeven hoeveel ze zich nog konden herinneren van de gepresenteerde antwoordpresentaties. Uit de resultaten bleek dat antwoordpresentaties met afbeeldingen met een hoog informatiegehalte informatiever en aantrekkelijker werden beoordeeld dan antwoordpresentaties met afbeeldingen met een laag informatiegehalte. De resultaten van de posttest suggereerden verder dat het leren van multimodale antwoordpresentaties tot betere leerresultaten leidde dan het leren van unimodale antwoordpresentaties.

Hoofdstuk 3 beschrijft een exploratieve hardopdenkstudie die onderzocht hoe gebruikers hun acties verbaliseren wanneer ze in een website navigeren. Bovendien werd onderzocht welke acties in spatiële termen werden uitgedrukt. Tien hardopdenkprotocollen werden verzameld, afkomstig uit twee verschillende usability studies. In beide studies werd aan de proefpersonen gevraagd om eenvoudige zoektaken uit te voeren op een website (d.w.z. het zoeken van antwoorden op feitenvragen) en hierbij hardop te denken. Het verzamelde corpus bestond het 694 gecodeerde segmenten die geanalyseerd werden op het type en het niveau van de acties waarin gebruikers verwikkeld waren tijdens het navigeren op een website. We onderscheidden twee actietypes: uitvoerende en evaluatieve acties. Daarnaast kon iedere actie nader beschreven worden in drie actieniveaus. Daarnaast onderzochten

we welke actietypes en actieniveaus werden uigedrukt in spatiële termen (bijv. "Ik ga terug naar de homepage."). De resultaten van de protocolanalyse toonden aan dat uitingen voornamelijk betrekking hadden op evaluatieve acties (bijv. "Ik kan hierop niet klikken."). Daarnaast hadden de meeste uitingen betrekking op acties op het eerste niveau (bijv. "Ik dubbelklik hierop."). Bovendien kwamen spatiële uitdrukkingen het meeste voor wanneer gebruikers uitvoerende acties beschreven op het eerste actieniveau (bijv. "Ik ga terug naar de homepage."). In het algemeen bevestigden de resultaten dat gebruikers spatiële termen gebruiken wanneer ze navigeren in een website. Echter, het verschil tussen een spatiële en een niet-spatiële uitdrukking was niet altijd eenduidig vast te stellen.

Hoofdstuk 4 beschrijft eerst een experiment waarin de effecten werden onderzocht van drie modaliteiten (tekst vs. foto vs. filmclip) en de moeilijkheidsgraad (eenvoudig vs. moeilijk) van een speciaal type procedurele instructies, namelijk RSI-preventieoefeningen. In het experiment moesten de proefpersonen tien eenvoudige (eenvoudige symmetrische bewegingen) en tien moeilijke (complexe symmetrische bewegingen of asymmetrische bewegingen) RSI-oefeningen leren en uitvoeren. De effectiviteit van het presenteren van een instructie in een tekst, foto en filmclip werd bepaald met de leertijd, het aantal geoefende bewegingen tijdens de leertijd, de uitvoeringstijd en het aantal correct uitgevoerde oefeningen. De proefpersonen moesten ook hun subjectieve satisfactie aangeven. De resultaten toonden aan dat er geen enkele modaliteit was die de andere modaliteiten overtrof op de afhankelijke variabelen. Ook de resultaten voor de subjectieve satisfactie lieten geen verschil zien tussen de drie modaliteiten.

In een voorkeurstudie werd verder onderzocht welke modaliteit gebruikers prefereren wanneer men RSI-oefeningen moet leren. Proefpersonen moesten zes RSI-oefeningen bestuderen, waarna ze moesten aangeven welke realisatie (d.w.z. tekst vs. foto vs. filmclip) van de oefeningen hun voorkeur had. Uit de resultaten bleek dat de proefpersonen over het algemeen de voorkeur hadden voor de filmclip. Echter, voor sommige oefeningen (bijv. "Maak van beide handen een vuist.") gaven de proefpersonen aan dat ze een voorkeur hadden voor een instructie in een tekst.

Hoofdstuk 5 beschrijft een experiment waarin oogbewegingsregistratie werd gebruikt om de incrementele verwerking van difoonsynthese, unitsynthese en menselijk spraak te bestuderen. In dit experiment, keken we zowel naar de segmentele als naar de suprasegmentele kwaliteit van de spraak. Als stimuli werden dertig Nederlandse zelfstandige naamwoorden gebruikt die uit één syllabe bestonden en konden worden afgebeeld. Daarnaast hadden deze zelfstandige naamwoorden dezelfde eerste fonemen (bijv. *vo-rk en vo-s*). De instructies werden in drie spraakcondities gerealiseerd: difoonsynthese, unitsynthese en menselijk spraak. De difoonsynthese werd gecreëerd met een Nederlands TTS systeem dat gebaseerd is op het Festival TTS systeem. De unitsynthese werd verkregen via een commercieel beschikbare unitsynthesizer. De stimuli voor de menselijke spraak werden opgenomen door een vrouw die Nederlands als moedertaal had. In het experiment kregen de proefpersonen een scherm te zien waarop vier objecten werden getoond. Bij ieder scherm kregen de proefpersonen twee opeenvolgende gesproken instructies te horen, die verwezen naar een object op het scherm. In de eerste instructie werd de referent genoemd (bijv. *roze vork*). De eerste instructie had een neutraal accentpatroon (bijv. *Kijk naar de roze vork*). In de tweede instructie werd het doelobject genoemd (bijv. *blauwe vos* of *blauwe vork*). De tweede instructie had een contextueel gepast dubbel accentpatroon (bijv. *Kijk nu naar de* BLAUWE VOS) of een contextueel ongepast dubbel accentpatroon (bijv. *Kijk nu naar de* BLAUWE VORK). Daarnaast moesten de proefpersonen een vragenlijst invullen over de begrijpelijkheid en de natuurlijkheid van de drie spraakcondities.

De resultaten toonden aan de proefpersonen het doelobject het snelst identificeerden in de menselijke spraakconditie. Het doelobject werd het minst snel geïdentificeerd in de difoonsynthese. De resultaten voor de unitsynthese vielen tussen de resultaten van de andere twee spraakcondities in. Daarnaast lieten de resultaten zien dat wanneer de tweede instructie een contextueel gepast dubbel accentpatroon had (bijv. *Kijk nu naar de* BLAUWE VOS), de fixaties naar het concurrerende object (*blauwe vork*) toenamen. Blijkbaar interpreteerden de proefpersonen het accent op het adjectief in de tweede instructie (BLAUWE) contrasterend ten opzichte van het adjectief in de eerste instructie (*roze*). Dit impliceert dat de proefpersonen anticipeerden op het doelobject dat genoemd werd in de tweede instructie. Bovendien bleek dat deze anticipatie moeilijk te corrigeren was voor de difoonsynthese maar

gemakkelijker was voor de unitsynthese en menselijke spraak. Tenslotte kwamen de resultaten van de vragenlijst overeen met de oogbewegingsdata: menselijke spraak was begrijpelijker en kwam natuurlijker over dan de unitsynthese en de difoonsynthese.

Hoofdstuk 6 presenteert de belangrijkste resultaten van de vier studies en eindigt met enkele bevindingen over multimodale informatiepresentatie.

Wanneer men verschillende modaliteiten met elkaar vergelijkt, is het belangrijk dat ze dezelfde hoeveelheid informatie weergeven. Echter, dit is niet zo eenvoudig als het misschien lijkt. Zo is het bijvoorbeeld mogelijk om uit te drukken hoe een bepaalde beweging voelt in een tekstuele instructie. Maar dit is niet mogelijk in een visuele instructie. Dit impliceert dat sommige informatie niet van de ene naar de andere modaliteit 'vertaald' kan worden. Het is daarom misschien een illusie om te denken dan dezelfde hoeveelheid informatie gepresenteerd kan worden in verschillende modaliteiten.

Wanneer informatie door meerdere modaliteiten wordt gepresenteerd, is de kans aanwezig dat meerdere modaliteiten dezelfde informatie weergeven. Er is dan sprake van redundantie. Echter het weergeven van redundante informatie kan een negatief effect hebben op het leren van de informatie. Het onderzoek dat in Hoofdstuk 2 werd gepresenteerd, gaf aan dat gebruikers informatie op een multimodale manier weergeven (d.w.z. men gebruikte een combinatie van tekst en afbeeldingen). Deze afbeeldingen hadden meerdere functies: ze waren slechts decoratief of ze representeerden een element dat in het tekstuele antwoord werd genoemd of ze voegden informatie toe aan het tekstuele antwoord. Het presenteren van multimodale informatie impliceert dus een zekere mate van redundantie. Het is echter nog onduidelijk hoe de mate van redundantie het leren van multimodale informatie presentaties beïnvloedt.

Curriculum Vitae

Charlotte van Hooijdonk was born on October 30th 1980 in Hulst, The Netherlands. She studied Communication and Information Sciences at Tilburg University and graduated cum laude in 2003, specializing in Business Communication and Digital Media. After her graduation, Charlotte started her PhD research at the Communication and Cognition Group at Tilburg University. She was involved in the IMOGEN (Interactive Multimodal Output GENeration) project within the NWO research programme on Interactive Multimodal Information Extraction (IMIX). Charlotte currently works at the Department of Language and Communication at the Vrije Universiteit Amsterdam as an Assistant Professor, where she combines research and teaching.