On the Role of Visuals in Multimodal Answers to Medical Questions

Charlotte van Hooijdonk
Tilburg University
C.M.J.vanHooijdonk@uvt.nl

Jurry de Vos
Tilburg University
jurrydevos@xs4all.nl

Emiel Krahmer
Tilburg University
E.J.Krahmer@uvt.nl

Alfons Maes
Tilburg University
Maes@uvt.nl

Mariët Theune
University of Twente
M.Theune@utwente.nl

Wauter Bosma
University of Twente
W.E.Bosma@utwente.nl

Abstract

This paper 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 was carried out to determine which modalities people choose to answer different types of questions. In this experiment, participants had to create (multimodal) presentations of answers to general medical questions. The collected answer presentations were coded on the presence of visual media (i.e., photos, graphics, and animations) and their function. The results indicated that participants presented the information in a multimodal way. Moreover, significant differences were found in the presentation of different answer and question types. Next, an evaluation experiment was conducted to investigate how users evaluate different types of multimodal answer presentations. In this second experiment, 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 picture (illustrative vs. informative). After the participants had assessed the answer presentations, they received a post-test in which they had to indicate how much they had recalled from the presented answer presentations. The results showed that answer presentations with an informative picture were evaluated as more informative and more attractive than answer presentations with an illustrative picture. The results for the post-test tentatively indicated that learning from answer presentations with an informative picture leads to a better learning performance than learning from purely textual answer presentations.

Keywords: multimodal information presentation, cognitive engineering, document design.

Introduction

In this paper, we describe ongoing work in the context of a medical question answering system created within the Dutch research programme IMIX. A question answering (QA) system is an automatic system that can answer a user's question posed in natural language (e.g., "What is the capital of the Netherlands?") with an answer formulated in natural language (e.g., "Amsterdam"). The IMIX QA system has been designed to help users find information they need in the medical domain and covers so-called "encyclopedic questions". These are general medical questions of which the answers do not require expert medical knowledge.

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 [1]. 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 [2]. This raises the question how to determine for a given question, what the best

1-4244-1243-9/07/$25.00 ©2007 IEEE

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 psychology on the influence of (combinations of) different modalities on the users' understanding, recall and processing efficiency of the presented material (e.g., [3] - [5]). 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) [4]. 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., [6] - [8]) and procedural information describing how to acquire a certain skill (e.g., [9] - [11]). Yet, these guidelines do not tell us which modalities are most suited for which information types, as each learning domain has its own characteristics [12].

Several researchers have tried to make an overview of the characteristics of modalities, information types, and the matches between them. For example, Bernsen 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"* [13, 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., [14], [15]).

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?"* [16, p. 280]. According to Arens et al. [16] 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 communication theory as a whole (e.g., [16] - [19]).

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. [20]. 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 paper describes two experiments carried out in order to investigate the role of visuals in multimodal answer presentations for a medical question answering system. We first describe a production experiment 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, we describe an evaluation experiment 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.

Experiment I: Production

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 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. Our coding criteria for these variables are discussed below. To determine the reliability of the coding system, Cohen's κ [21] was calculated.

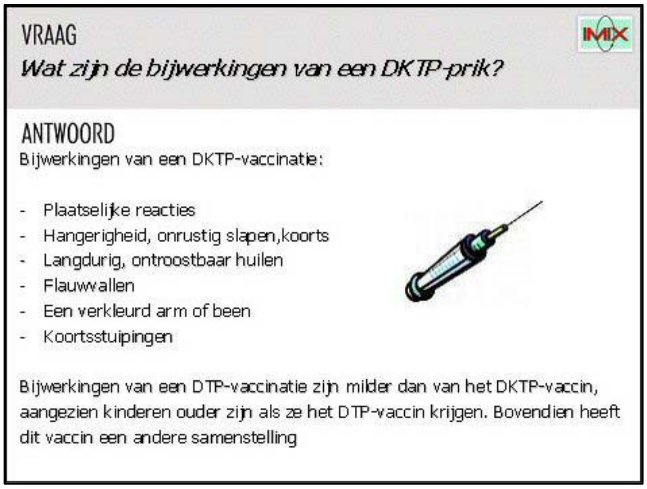
Photos. We distinguished whether the answer contained no photo, one photo or several photos.

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 [3]:

- *Decorational function*: a visual medium has a decorational function if removing it from the answer presentation does not alter the informativity of the answer in any way. Figure 1 shows an example of answer presentations in which the visual medium has a decorational function. 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.
- *Representational function*: a visual medium 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 shows two examples of answer presentations in which the visual medium has a representational function. 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.
- *Informative function*: a visual medium has an informative function if removing it from the answer presentation decreases the informativity of the answer. If an answer consists only of a visual medium, it automatically has an informative function. Figure 3 shows two examples of answer presentations in which the visual medium has an informative function. 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?"



VRAAG

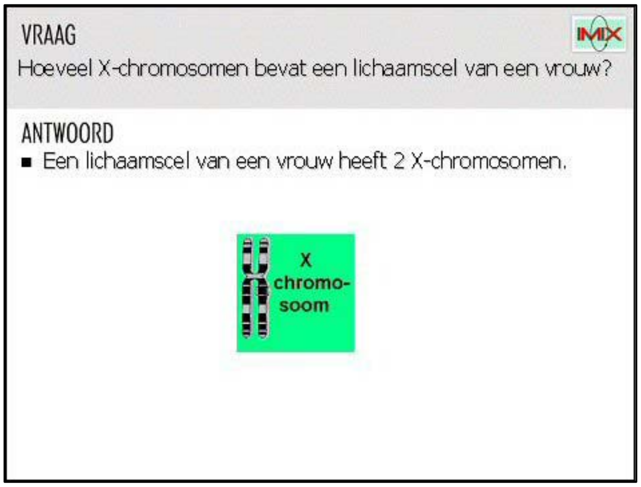
Wat zijn de bijwerkingen van een DKTP-prik?

ANTWOORD

Bijwerkingen van een DKTP-vaccinatie:

- Plaatselijke reacties
- Hangerigheid, onrustig slapen,koorts
- Langdurig, ontroostbaar huilen
- Flauwvallen
- Een verkleurd arm of been
- Koortsstuipingen

Bijwerkingen van een DTP-vaccinatie zijn milder dan van het DKTP-vaccin, aangezien kinderen ouder zijn als ze het DTP-vaccin krijgen. Bovendien heeft dit vaccin een andere samenstelling



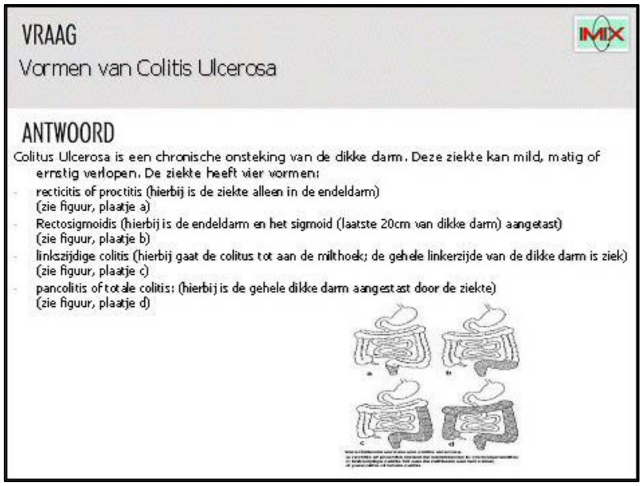
VRAAG

Hoeveel X-chromosomen bevat een lichaamscel van een vrouw?

ANTWOORD

■ Een lichaamscel van een vrouw heeft 2 X-chromosomen.

Figure 1. Examples of answer presentations with decorational visuals



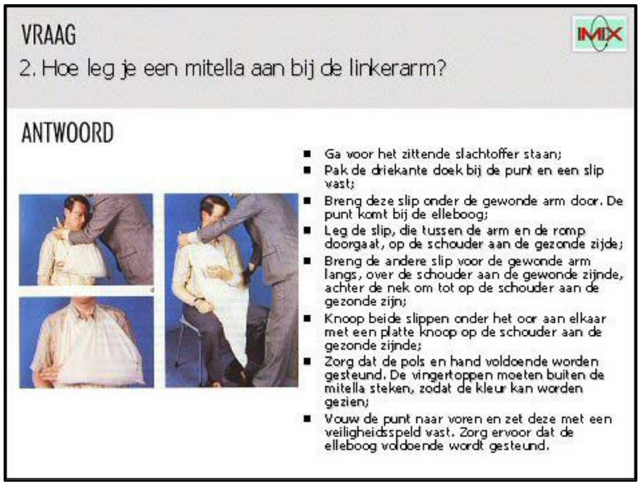
VRAAG

Vormen van Colitis Ulcerosa

ANTWOORD

Colitus Ulcerosa is een chronische onsteking van de dikke darm. Deze ziekte kan mild, matig of ernstig verlopen. De ziekte heeft vier vormen:

- recticitis of proctitis (hierbij is de ziekte alleen in de endeldarm) (zie figuur, plaatje a)
- Rectosigmoidis (hierbij is de endeldarm en het sigmoid (laatste 20cm van dikke darm) aangetast) (zie figuur, plaatje b)
- linkszijdige colitis (hierbij gaat de colitus tot aan de milthoek; de gehele linkerzijde van de dikke darm is ziek) (zie figuur, plaatje c)
- pancolitis of totale colitis: (hierbij is de gehele dikke darm aangestast door de ziekte) (zie figuur, plaatje d)



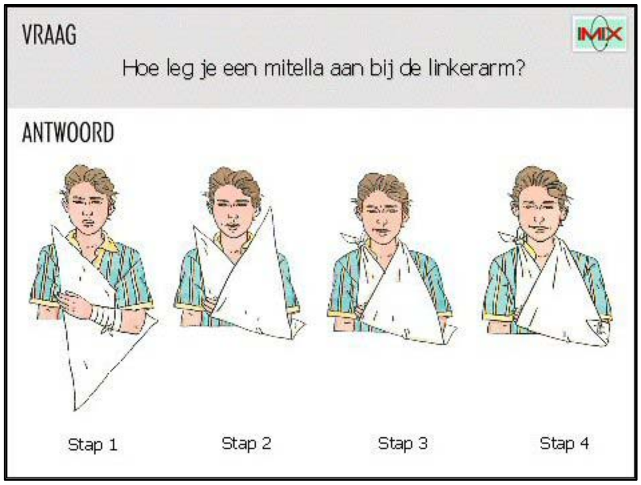
VRAAG

2. Hoe leg je een mitella aan bij de linkerarm?

ANTWOORD

- ■ Ga voor het zittende slachtoffer staan;
- ■ Pak de driekante doek bij de punt en een slip vast;
- ■ Breng deze slip onder de gewonde arm door. De punt komt bij de elleboog;
- ■ Leg de slip, die tussen de arm en de romp doorgaat, op de schouder aan de gezonde zijde;
- ■ Breng de andere slip voor de gewonde arm langs, over de schouder aan de gewonde zijnde, achter de nek om tot op de schouder aan de gezonde zijn;
- ■ Knoop beide slippen onder het oor aan elkaar met een platte knoop op de schouder aan de gezonde zijnde;
- ■ Zorg dat de pols en hand voldoende worden gesteund. De vingertoppen moeten buiten de mitella steken, zodat de kleur kan worden gezien;
- ■ Vouw de punt naar voren en zet deze met een veiligheidsspeld vast. Zorg ervoor dat de elleboog voldoende wordt gesteund.

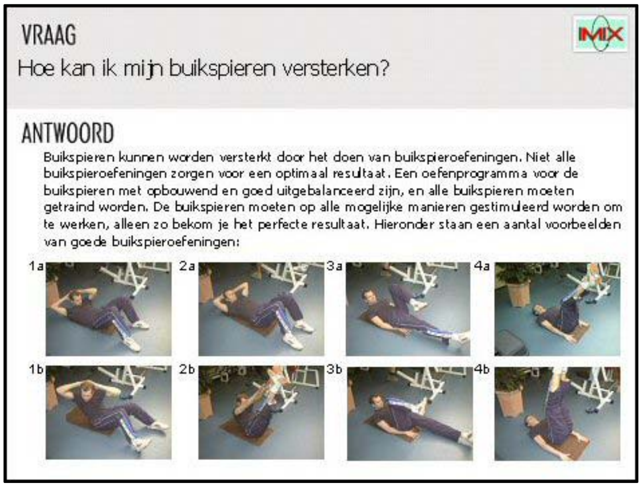
Figure 2. Examples of answer presentations with representational visuals



VRAAG

Hoe leg je een mitella aan bij de linkerarm?

ANTWOORD



VRAAG

Hoe kan ik mijn buikspieren versterken?

ANTWOORD

Buikspieren kunnen worden versterkt door het doen van buikspieroefeningen. Niet alle buikspieroefeningen zorgen voor een optimaal resultaat. Een oefenprogramma voor de buikspieren met opbouwend en goed uitgebalanceerd zijn, en alle buikspieren moeten getraind worden. De buikspieren moeten op alle mogelijke manieren gestimuleerd worden om te werken, alleen zo bekom je het perfecte resultaat. Hieronder staan een aantal voorbeelden van goede buikspieroefeningen:

Figure 3. Examples of answer presentations with informative visuals

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 authors). 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 [22]. 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).

Results

Descriptive statistics. Table 1 shows the frequencies of visual media (overall), photos, graphics, and animations in the complete corpus of coded answer presentations. Inspection of Table 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 frequencies of photos, graphics, and animations in the corpus exceeded the frequency of the variable visual media.

Table 2 shows the frequencies of photos, graphics, and animations related to their function. Note that the answer presentations in which photos, graphics, or animations co-occurred are not shown in the table. Table 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.

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 1. Frequencies of visual media in the complete corpus of coded answers (n = 1775).

Visual media		24.9
	Photos	8.6
	Graphics	14.9
	Animations	3.8

Table 2. Frequencies of photos, graphics, and animations related to their function (Scores are percentages of answers).

	Function of visual media			
	Decorational function	Representational function	Informative function	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

Table 3. Frequencies of types of photos and types of graphics related to their function (Scores are percentages of answers).

	Function of visual media			
	Decorational function	Representational function	Informative function	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

Table 4. Frequencies and χ^2 statistics of the presence of visual media (overall), photos, graphics, and animations related to the brief and extended answers (Scores are percentages of answers; n = 1775).

	Brief answers (n = 888)	Extended answers (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 5. Frequencies of the function of visual media related to brief and extended answers (Scores are percentages of answers; n = 444)

	Brief answers (n = 102)	Extended answers (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
Totals	100.0	100.0	

Table 3 shows the results. It turned out that photos without text occurred significantly more often than photos with text (χ^2 (1) = 303.77, p< .001). The reverse was found for graphics: graphics with text occurred significantly more often than graphics without text (χ^2 (1) = 1162,62, p< .001).

Moreover, 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). However, most photos with text were associated with a representational function or a decorational function (χ^2 (2) = 7.79 p< .025). An example of representational photo with text was a photo of a woman's chromosome pattern in which the particular sex chromosomes were indicated by text. Decorational photos with text did not add any information to the text of the answer presentation. For example, some answers discussed the side effects of ibuprofen. These answers were often illustrated with a photo of a box of medicines with the medicines' name on it. 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.

Brief and extended answers. Different types of answers were related to different answer presentations. Table 4 shows the frequencies 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.

Table 5 shows the frequencies 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 most often in brief answers, whereas representational visuals occurred most often in extended answers. Finally, informative visuals occurred most often in brief answers.

Table 6. Frequencies and χ^2 statistics of the presence, visual media (overall), photos, graphics, and animations within the different question types (Scores are percentages of answers).

	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 7. Frequencies and χ^2 statistics of the functions of visual media related to the different question types (Scores are percentages of answers; 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
Totals	100.0	100.0	100.0	100.0	

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 6 shows the frequencies 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 most 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 7 shows the frequencies 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 most often in definition questions *without* reference to body parts. Representational visuals occurred most often in definition questions *with* reference to body parts. Finally, informative visuals occurred most often in procedural questions *with* reference to body parts.

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, we discuss an evaluation experiment in which users were instructed to assess answer presentations on their informativity and attractiveness.

Experiment II: Evaluation

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 2 (length of the textual answer) × 3 (type of visual) factorial design with both the length of the textual answer (brief, extended) and the type of visual (no visual, illustrative visual, informative visual) as between participants variables. 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 more 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 were 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 + illustrative visual condition and in the text + informative visual condition.

In the two text + illustrative visual conditions, we presented the brief and the extended textual answers together with an illustrative visual. An illustrative visual had been given a decorational or a representational function in the production experiment. Figure 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 a concept (i.e., a workspace) mentioned in the textual answers. However, the answers would not be less informative if the photo was not present.

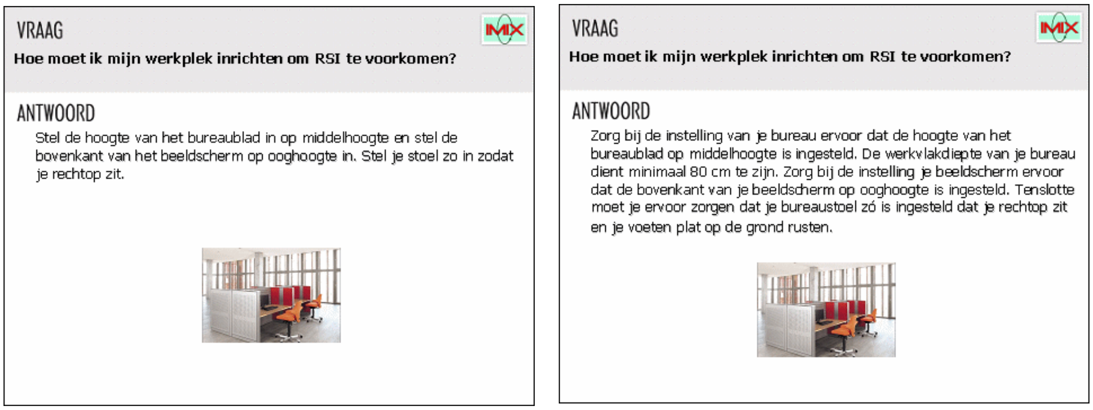
In the two text + informative visual conditions, we presented the brief and extended textual answers together with an informative visual. A visual was informative if it had been given an information function in the production experiment. Figure 5 illustrates a brief textual answer and an extended textual 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 be less informative 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 [23], 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 questions and answer presentations were presented in a 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 6 illustrates the procedure of the evaluation experiment. When the participants accessed the experiment, they first received instructions about the procedure of the experiment.



VRAAG

Hoe moet ik mijn werkplek inrichten om RSI te voorkomen?

ANTWOORD

Stel de hoogte van het bureaublad in op middelhoogte en stel de bovenkant van het beeldscherm op ooghoogte in. Stel je stoel zo in zodat je rechtop zit.

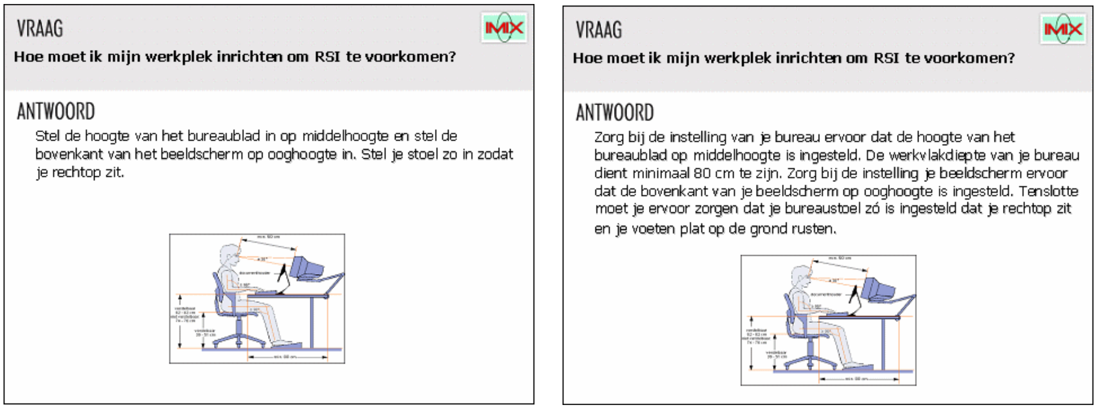
VRAAG

Hoe moet ik mijn werkplek inrichten om RSI te voorkomen?

ANTWOORD

Zorg bij de instelling van je bureau ervoor dat de hoogte van het bureaublad op middelhoogte is ingesteld. De werkvlakdiepte van je bureau dient minimaal 80 cm te zijn. Zorg bij de instelling je beeldscherm ervoor dat de bovenkant van je beeldscherm op ooghoogte is ingesteld. Tenslotte moet je ervoor zorgen dat je bureaustoel zó is ingesteld dat je rechtop zit en je voeten plat op de grond rusten.

Figure 4. Examples of a brief textual answer (left) and an extended textual answer (right) with an illustrative visual



VRAAG

Hoe moet ik mijn werkplek inrichten om RSI te voorkomen?

ANTWOORD

Stel de hoogte van het bureaublad in op middelhoogte en stel de bovenkant van het beeldscherm op ooghoogte in. Stel je stoel zo in zodat je rechtop zit.

VRAAG

Hoe moet ik mijn werkplek inrichten om RSI te voorkomen?

ANTWOORD

Zorg bij de instelling van je bureau ervoor dat de hoogte van het bureaublad op middelhoogte is ingesteld. De werkvlakdiepte van je bureau dient minimaal 80 cm te zijn. Zorg bij de instelling je beeldscherm ervoor dat de bovenkant van je beeldscherm op ooghoogte is ingesteld. Tenslotte moet je ervoor zorgen dat je bureaustoel zó is ingesteld dat je rechtop zit en je voeten plat op de grond rusten.

Figure 5. Examples of a brief textual answer (left) and an extended textual answer (right) with an informative visual

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 text + 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?”:

a. Testosterone is a sex hormone that is produced by males and females in the adrenal glands. Besides,

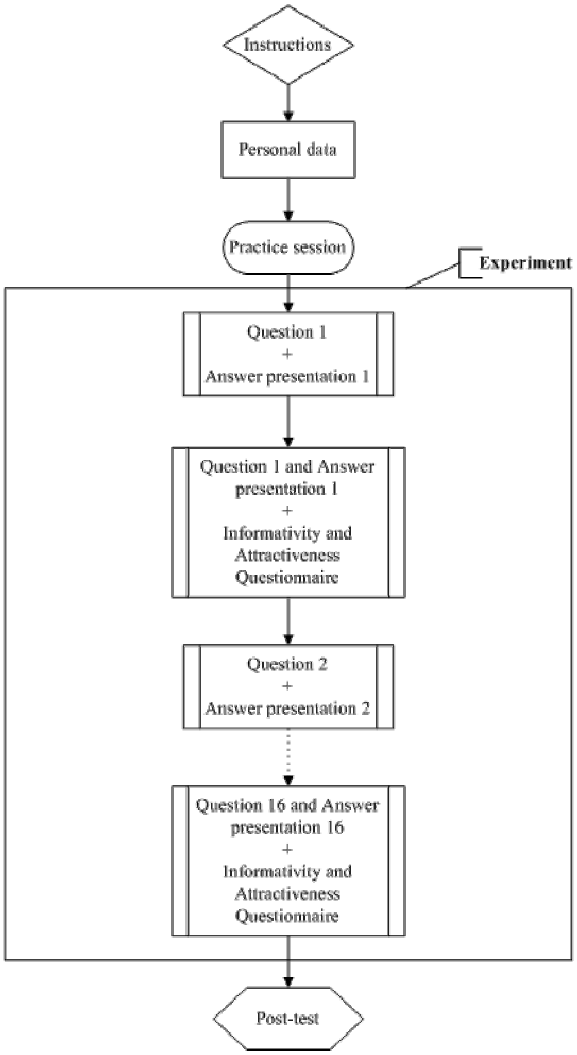

Instructions
Personal data
Practice session
Experiment
Question 1 + Answer presentation 1
Question 1 and Answer presentation 1 + Informativity and Attractiveness Questionnaire
Question 2 + Answer presentation 2
Question 16 and Answer presentation 16 + Informativity and Attractiveness Questionnaire
Post-test

Figure 6. Procedure of the evaluation experiment.

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 were 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 (brief answer + illustrative visual, extended answer + illustrative visual, brief answer + informative visual, extended answer + informative visual) × 2 (definition question, procedural question) repeated measures analysis of variance (ANOVA), with a significance threshold of .05. For post hoc tests, the Bonferroni method was used. Note that inconclusive results were found for answer presentations to questions with and without reference to body parts. Therefore, we report on the results found for definition and procedural questions.

Results

Informativity of the text and visual combinations. Table 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, η^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).

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 in the perceived informativity of the text and visual combination between the two question types. Procedural answer presentations with informative visuals were more informative than

Table 8. Mean results of the assessment on the informativity and the attractiveness of the 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)

Table 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)

definition answers presentations with informative visuals.

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 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).

Number of correct answers in the post-test. Table 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 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 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).

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.

Discussion

This paper 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. A total of 1775 answer presentations were collected of which almost one in four contained one or more visual media. 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 informative function whereas visual media with a representational function were more frequent within extended answers. A possible 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 will have a representational function (i.e., it visually represents the information already present in text).

The type of question was also associated with different answer presentations. Photos and animations occurred most often in answers to procedural questions with reference to body parts. These visual media may help to visualize the steps of a procedure. However, graphics occurred most often in answers to definition questions with reference to body parts. While photos represent reality, graphics schematize reality making them more suitable to illustrate the topics of definition questions.

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 more informative than *definition* answer presentations with informative visuals. An explanation for this result could be that procedures 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. Arguably, given that extended texts are inherently more informative than brief ones, it is conceivable that an informative picture adds less to an extended text, and as a result primarily enhances the attractiveness of the presentation

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 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 this paper, we conducted two exploratory studies to investigate when and how people produce multimodal information presentations and how other people evaluate such presentations. In both experiments, a consistent result was found: participants preferred informative visuals to illustrative visuals. 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 informative visuals in procedural answer presentations and illustrative (i.e., representational) visuals in definition answer presentations.

There are many opportunities for further work. 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. Finally, 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.

Acknowledgements

A preliminary report about the production experiment appeared as Van Hooijdonk et al. [24]. The current research is performed in the IMOGEN (Interactive Multimodal Output GENeration) project. IMOGEN is a project within the Netherlands Organisation for Scientific Research (NWO) research programme on Interactive Multimodal Information eXtraction (IMIX). The authors would like to thank Paul Flapper for his help in setting up the production experiment, and Jeroen Geertzen for his help with the statistics. The authors would also like to thank Lennard van de Laar for his help in setting up the evaluation experiment and Martin Brandt for processing the data of the evaluation experiment.

References

[1] M. Theune, B. van Schooten, R. op den Akker, W. Bosma, D. Hofs, A. Nijholt, E. Krahmer, C. van Hooijdonk, and E. Marsi. Questions, pictures, answers: introducing pictures in question-answering systems. In *Proceedings of the ACTAS-1 of X Symposio Internacional de Comunicacion Social*, 450-463, 2007.

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

[3] Carney, R., and J. Levin. Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*. 14: 5-26, 2002

[4] Mayer, R. *The Cambridge handbook of multimedia learning*, Cambridge University Press, Cambridge, 2005

[5] Tversky, B., J. Morrison, and M. Betrancourt. Animation: can it facilitate? *Int. Journal of Human Computer Studies*. 57: 247-262, 2002

[6] Mayer, R. Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology*. 81: 240-246, 1989

[7] Mayer, R., and J. Gallini. When is an illustration worth a thousand words? *Journal of Educational Psychology*. 82:715-726, 1990

[8] Mayer, R., and R. Moreno. Aids to computer-based multimedia learning. *Learning and Instruction*. 12: 107-119, 2002

[9] Marcus, N., M. Cooper, and J. Sweller. Understanding instructions. *Journal of Educational Psychology*. 88: 49-62, 1996

[10] Michas, I., and D. Berry. Learning a procedural task: effectiveness of multimedia presentations. *Applied Cognitive Psychology*. 14: 555-575, 2000

[11] Schwan, S., and R. Riempp. The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and Instruction*. 14: 293-305, 2004

[12] Hooijdonk, C.M.J., and E. Krahmer. Information modalities for procedural instructions: the influence of text, static and dynamic visuals on learning and executing RSI exercises. *IEEE Transactions on Professional Communication*, to appear

[13] Bernsen, N. Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers*. 6: 347-371, 1994

[14] Heller, R., C. Martin, N. Haneef and S. Gievka-Krliu. Using a theoretical multimedia taxonomy framework. *ACM Journal of Educational Resources in Computing*. 1: 1-22, 2001

[15] Sutcliffe, A. Task-related information analysis. *Int. Journal of Human Computer Studies*. 47: 223-257, 1997

[16] Arens, Y., E. Hovy, and M. Vossers. *On the knowledge underlying multimedia presentations*. In Intelligent Multimedia Interfaces. M. T. Maybury (Eds.) AAAI Press, Menlo Park, 1993, 280-306

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

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

[19] Oviatt, S., R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson, and L. Carmichael. Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, Vancouver, British Columbia ,Canada, 2003, 44-51

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

[21] Krippendorff, K. *Content analysis: an introduction to its methodology*. Sage Publications, Beverly Hills, 1980.

[22] Rietveld, T., and R. Van Hout. *Statistical techniques for the study of language and language behaviour*. Mouton de Gruyter, Berlin, 1993.

[23] Veenker, T. WWStim: A CGI script for presenting web-based questionnaires and experiments, 2005, Available: http://www.let.uu.nl/Theo.Veenker/personal/projects/wwstim/doc/en/

[24] Hooijdonk, C.M.J. van, E. Krahmer, A. Maes, M. Theune, and W. Bosma. Towards automatic generation of multimodal answers to medical questions: a cognitive engineering approach. In the Proceedings of the Workshop on Multimodal Output Generation (MOG 2007), 25-26 January 2007, Aberdeen, Scotland, pages 93-104.

About the Authors

Charlotte van Hooijdonk is a PhD student at Tilburg University, The Netherlands. Her PhD research focuses on how users process and evaluate multimodal information presentations using different evaluation methods. She has been a member of the IEEE-PCS since 2005.

Jurry de Vos is a Master student at Tilburg University, The Netherlands. His Master thesis focuses on how users evaluate multimodal information presentations in a medical question answering system. He is expected to graduate in August 2007.

Emiel Krahmer is a full professor at Tilburg University, The Netherlands. His research interests include multimodal and multisensory information processing, developing and evaluating language technology applications and the analysis of non-verbal communication of real and virtual humans.

Alfons Maes is professor in Business Communication and Digital Media at Tilburg University, the Netherlands, and is head of the Communication and Cognition research programme. His research interests include multimodal and digital communication and document design. He has published articles on discourse reference, document design, and instructive discourse. He is member of the IEEE.

Mariët Theune is an assistant professor at the University of Twente, The Netherlands, where she moved after having received her PhD at the Eindhoven University of Technology. A computational linguist by training, she has been involved in various research projects in the area of natural language generation and multimodal information presentation.

Wauter Bosma is a PhD student at the University of Twente, the Netherlands. His research aims to improve computer systems for automatic question answering and, to that end, generating multimedia answer presentations.