

# Experiments in Multimodal Information Presentation

Charlotte van Hooijdonk, Wauter Bosma, Emiel Krahmer, Alfons Maes, and Mariët Theune

## 1 Introduction

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, such as “Which presentation modes are most suitable given a particular communicative goal?” and “How should different presentation modes be combined?” The IMOGEN (Interactive Multimodal Output GENERation) project addressed these questions. This project was embedded in the Dutch national research programme IMIX (Interactive Multimodal Information eXtraction). Within IMIX a multimodal medical question answering (QA) system was developed. The purpose of this system is to answer encyclopedic medical questions from non-expert users. Questions can be typed or spoken (in Dutch), and answers are presented using speech, text and pictures. Questions can be asked in isolation, but the system is also capable of engaging in dialogs and answer follow-up questions.

---

Charlotte van Hooijdonk  
VU University, De Boelelaan 1105, NL-1081 HV Amsterdam, The Netherlands, e-mail:  
cmj.van.hooijdonk@let.vu.nl

Wauter Bosma  
VU University, De Boelelaan 1105, NL-1081 HV Amsterdam, The Netherlands, e-mail:  
w.bosma@let.vu.nl

Emiel Krahmer  
Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands, e-mail:  
e.j.krahmer@uvt.nl

Alfons Maes  
Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands, e-mail:  
maes@uvt.nl

Mariët Theune  
University of Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands, e-mail:  
m.theune@ewi.utwente.nl

In the IMOGEN project different aspects of multimodal information presentation were studied in order to improve the output quality of question answering (QA) systems. 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 presentation mode of the answers to these questions was typically text only. Nowadays, QA systems are also expected to give answers to more complex questions, which might be more informative and effective if they contained multiple presentation modes, such as text and a picture. Here, we focus on questions in the medical domain, since the QA system developed as a demonstrator in IMIX aimed at answering encyclopedic medical questions from non-expert users.

People can have different medical questions, including factoid definition questions, such as “What is RSI?” and procedural questions about how to take care of one’s health, such as “How to prevent RSI?” People may also have different information needs. In some situations, they are satisfied with a short answer in which, for example, the abbreviation RSI is clarified (*Repetitive Strain Injury*). In other cases, they want a longer answer in which more information is given about the causes and consequences of the disorder. (For example, *RSI stands for Repetitive Strain Injury. This disorder involves damage to muscles, tendons and nerves caused by overuse or misuse, and affect the hands, wrists, elbows, arms, shoulders, back, or neck.*) The answers to these medical questions can be presented through text or through a combination of presentation modes, such as text and a static or dynamic picture. For example, the most suitable answer presentation to the definition (*what*) question “What does RSI stand for?” would probably be a short textual answer, such as “RSI stands for Repetitive Strain Injury”. The answer to the procedural (*how*) question “How to organize a workspace in order to prevent RSI?” would probably be more informative if it contained a picture. This raises the questions how to determine for a given question, whether a short or a long answer would be preferable and which (combinations of) presentation modes are most suitable.

Multimodal information presentation has been studied in various research fields with various outcomes. Research in cognitive and educational psychology focused on how multimodal presentations affect the users’ understanding, recall and processing efficiency of the presented material (e.g., [7, 17, 22]). Guidelines resulting from this research often relate to specific types of information used in specific domains, for example cause and effect chains [16] or procedural information [18]. Yet, these guidelines do not tell us which modalities are most suited for which information types, as each learning domain has its own characteristics [11].

Research in user interfaces has tried to classify and characterize presentation modes. For example, Bernsen [3] proposed a taxonomy of generic unimodalities consisting of various features. Other scholars studied the so-called *media allocation problem* (i.e., how to determine which information to allocate to which medium) and tried to identify which factors play a role in media allocation [1]. They found out that many factors are relevant: the nature of the information, the communicative situation, goals of the producer, and features of the addressee.

In short, attempts have been made to generate optimal multimodal information presentations resulting in several presentation mode guidelines, frameworks, and taxonomies. Still needed is information about people's modality preferences in producing and evaluating presentations. Therefore, we carried out three experiments following the approach of Heiser, Phan, Agrawala, Tversky and Hanrahan [10]. The experiments investigated multimodal information presentation in the context of a medical QA system. In Experiment 1, people were asked to produce information presentations, which were then rated by others in Experiment 2. In Experiment 3, the answer presentations manually produced in Experiment 1 were compared to presentations with automatically retrieved pictures.

In this chapter we present the three experiments. In Experiment 1, we wanted to know how non-experts design (multimodal) answers to medical questions, distinguishing between *what* questions and *how* questions. In Experiment 2, we concentrated on how people evaluate multimodal (text+picture) answer presentations on their informativeness and attractiveness. In Experiment 3, we evaluated two versions of an automatic picture selection method, and compared answer presentations with automatically selected pictures to answer presentations with manually selected pictures.

## 2 Experiment 1: Production of Multimodal Answers

In this section we present an experiment that was carried out to determine which modalities people choose to answer different types of questions. In the experiment, participants had to create (multimodal) presentations of answers to general medical questions. More details on the experiment can be found in [12].

### 2.1 Participants

Participants were 111 students of Tilburg University, who participated for course credits (65 female and 46 male). Their average age was 22 (SD = 2.10, min = 19, max = 32). All participants were native speakers of Dutch. All were second-year undergraduate students who had received Internet search training in the first year of their studies. They were all familiar with PowerPoint and used it on a regular basis (daily: 3.6%, weekly: 22.5%, monthly: 51.4%, yearly: 18.0%, never: 4.5%). Finally, participants indicated on one 7-point semantic differential that their PowerPoint skills were above average (M = 5.01, SD = 1.10).

## 2.2 Stimuli

Participants were given one of four sets of eight general medical questions for which the answers could be found on the Internet. They had to provide two types of answers per question, a short and a long answer, using whatever combination of presentation modes they wanted. They did not get explicit instructions on the number of words or pictures to be used in their answers. Participants were specifically asked to present the answers as they themselves would prefer to find them in a QA system. Questions and answers had to be presented in a fixed format in PowerPoint™ with areas for the question ('vraag') and the answer ('antwoord'). Participants were given a short introduction about PowerPoint in which they were acquainted with inserting different types of objects in PowerPoint. Also, they received a PowerPoint manual. Of the eight questions in each set, four were randomly chosen from one hundred medical questions formulated to test the IMIX system. Of the remaining four questions, two were *what* questions (e.g., "What are thrombolytic drugs?") and two were *how* questions (e.g., "How to apply a sling to the left arm?").

## 2.3 Coding System and Procedure

Each answer was coded on the presence of visual media (i.e. photos, graphics, and animations) – pictures, in short – and on the function of these pictures in relation to the text, loosely based on Carney and Levin [7], i.e., decorative, representational, or informative.

**Decorative function** A picture has a decorative function if removing it from the answer presentation does not alter the informativeness of the answer in any way. Figure 1 shows an example of an answer with a decorative picture. The answer to the question "What are the side effects of a vaccination for diphtheria, whooping cough, tetanus, and polio?" consists of a combination of text and a graphic. The text describes the side effects of the vaccination, while the graphic shows a syringe. The answer would not be less informative if the graphic was absent.

**Representational function** A picture has a representational function if removing it from the answer presentation does not alter the informativeness of the answer, but its presence clarifies the text. Figure 2 shows an example of an answer presentation with a representational picture. The question "What types of colitis can be distinguished?" is answered through text and a graphic. The text describes the four types of colitis and where they are located in the intestines. This information is visualized in the graphic.

**Informative function** A picture has an informative function if removing it from the answer presentation decreases the informativeness of the answer. If an answer only consists of a picture, it automatically has an informative function. Figure 3 shows an example of an answer with an informative picture. The answer to the question: "How can I strengthen my abdominal muscles?" consists of text and

**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



**Fig. 1** Example of an answer with a decorative picture.

photos. The text describes some general information about abdominal exercises (i.e., an exercise program should be well balanced and train all abdominal muscles). The last sentence refers to four exercises that can be done do strengthen the abdominal muscles. These exercises are illustrated with eight photos. For each exercise two photos are given, indicating the first (a) and last (b) step of the exercise.

In total 1776 answers were collected (111 participants  $\times$  8 questions  $\times$  2 answers). One of the participants omitted one answer, so that the final data set consisted of 1775 answers. Six analysts independently coded the same set of 111 answers. Subsequently, every analyst independently coded a part of the total corpus (approximately 300 answers). Calculations of Cohen's  $\kappa$  showed that the analysts agreed almost perfectly in judging the occurrence of photos ( $\kappa = .81$ ), graphics ( $\kappa = .83$ ), and animations ( $\kappa = .92$ ). An almost perfect agreement was also reached in assigning the function of the picture media ( $\kappa = .83$ ).

## 2.4 Results

Analysis of the complete corpus of coded answer presentations showed that almost one in four answers contained one or more pictures ( $n = 442$ ), consisting of graphics ( $n = 232$ ), photographs ( $n = 124$ ), or animations ( $n = 49$ ). In 37 cases, a combination of these media was used.

**VRAAG**

Welke vormen van colitis worden onderscheiden?

**ANTWOORD**

Colitis ulcerosa als een chronische ontsteking van onbekende aard, die zich geheel beperkt tot de dikke darm. De naam colitis ulcerosa komt uit het Latijn; colit betekent een ontsteking van de dikke darm, ulcerosa betekent met wonden. De ziekte kan wettig, matig ernstig of ernstig verlopen. Bij de meeste mensen worden aan de meniplole van geringe, actieve (soms vele jaren) en remissie (soms vele jaren) afwisselend. Colitis ulcerosa kan ook zowel goede als slechte periodes hebben. Zoals reeds gezegd kan colitis ulcerosa enkel in de dikke darm (colitis in de dikke darm) of het buik- en maagvond, dit in tegenstelling tot de ziekte van Crohn die in het hele maag-darmkanaal kan voorkomen. Zoals op onderstaande afbeelding te zien is, zijn er 4 vormen van colitis ulcerosa afhankelijk van de mate waarin de dikke darm aangevat is (de ziekte begint steeds vanaf de endeldarm).



A) rectitis of proctitis: Het betreft de ziekte alleen aanwezig in de endeldarm.  
 B) rectosigmoiditis: Het betreft de endeldarm en het sigmoid (laatste 20 cm van de dikke darm) aangevat.  
 C) linkszijdige colitis: Het betreft gedeelte colitis tot aan de navelhoogte en in eigenlijke de gehele linkerzijde van de dikke darm dek.  
 D) pancolitis of totale colitis: Het betreft de gehele dikke darm aangevat door colitis ulcerosa.

Fig. 2 Example of an answer with a representational picture.

**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 moet 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:



1a 2a 3a 4a  
 1b 2b 3b 4b

Fig. 3 Example of an answer with an informative picture.

*Answer length*

Long answers ( $M = 86$ ,  $SD = 60$ ) contained significantly more words than short answers ( $M = 18$ ,  $SD = 25$ ),  $t(168.78) = -10.58$ ,  $p < .001$  (since Levene's test was significant, a correction on the degrees of freedom was made). Table 1 shows that long answers contained significantly more pictures than short answers ( $\chi^2(1) = 173.89$ ,  $p < .001$ ). Moreover, the distribution of the functions of visual media differed significantly over answer length ( $\chi^2(2) = 33.79$ ,  $p < .001$ ). Decorative pictures occurred most often in short answers ( $\chi^2(1) = 4.07$ ,  $p < .05$ ), whereas representational pictures occurred most often in long answers ( $\chi^2(1) = 125.78$ ,  $p < .001$ ). Informative pictures occurred most often in short answers ( $\chi^2(1) = 23.81$ ,  $p < .001$ ).

**Table 1** Percentages of function of visual media related to short and long answers ( $n = 442$ ).

	Short answers ( $n = 101$ )	Long answers ( $n = 341$ )
Decorative pictures ( $n = 70$ )	26.7	12.6
Representational pictures ( $n = 201$ )	20.8	52.8
Informative pictures ( $n = 171$ )	52.5	34.6

*Question type*

Analysis of the two *what* questions and the two *how* questions ( $n = 887$ , of which 271 contained pictures) showed that pictures occurred significantly more often in *how* questions ( $\chi^2(1) = 29.23$ ,  $p < .001$ ). Table 2 also shows that answers to *what* questions contained significantly more decorative and representational pictures, while answers to *how* questions contained more informative pictures ( $\chi^2(2) = 22.70$ ,  $p < .001$ ).

**Table 2** Percentages of functions of pictures related to *what* questions and *how* questions ( $n = 271$ ).

	<i>What</i> questions ( $n = 91$ )	<i>How</i> questions ( $n = 180$ )
Decorative pictures ( $n = 27$ )	19.8	5.0
Representational pictures ( $n = 129$ )	53.8	44.4
Informative pictures ( $n = 115$ )	26.4	50.6

## **2.5 Conclusion**

The results showed that people made use of multiple presentation modes in their answers and that the design of these presentations was affected by answer length and question type. What is not clear, is how people evaluate multimodal (text+picture) answer presentations. In the next section, an evaluation experiment is discussed in which this issue was investigated.

## **3 Experiment 2: Evaluation of Multimodal Answers**

In this section we present Experiment 2, which was conducted to investigate how users evaluate different types of multimodal answer presentations. In this experiment, participants had to assess the informativity and attractiveness of answer presentations for different types of medical questions. More details on the experiment can be found in [13].

### **3.1 Participants**

Participants were 108 native speakers of Dutch (66 female and 42 male). Their average age was 25 ( $SD = 8.24$ ,  $min = 18$ ,  $max = 64$ ). None had participated in Experiment 1.

### **3.2 Design**

The experiment had a  $16$  (question)  $\times$   $2$  (short or long answer)  $\times$   $3$  (decorative picture, informative picture, or no picture) mixed factorial design, with the question as a within participants variable and the answer length and picture type as between participants variable. The dependent variables were the participants' assessment of: (a) the clarity of the text, (b) the informativeness of the answer presentation, (c) the attractiveness of the answer presentation, (d) informativeness of the text-picture combination and (e) the attractiveness of the text-picture combination. The participants were randomly assigned to an experimental condition.

### **3.3 Stimuli**

We selected 16 medical questions for which the corpus collected in Experiment 1 contained: (i) an informative picture, which added new information to the answer



and (ii) a decorative picture, which did not. Although the results of Experiment 1 showed that participants used different picture types when producing short and long answers, only informative and decorative pictures were taken into account in Experiment 2. It was hypothesized that decorative pictures would be evaluated as most attractive but least informative, as they might make the text more vivid but do not add new information to the textual answer. Informative pictures on the other hand would be evaluated as least attractive but most informative, as they add new information to the textual answer but do not necessarily make the information more attractive. Representational pictures visually display the main topic of the textual answer, but do not add new information. In this respect, they are quite similar to decorative pictures, as they might make the textual answer more vivid but add no new information to the textual answer. Therefore, representational and decorative pictures were combined into decorative pictures.

The set of selected questions consisted of eight *what* questions and eight *how* questions. For each question a short and a long textual answer was formulated. The textual answers were chosen from the set of answers collected in Experiment 1. Small adjustments were made to these answers in order to make them more comparable. The short answer gave a direct answer to the question, while the long answer also provided some relevant background information. The average length of the short answers was 26 words and the average length of the long answers was 66 words. We made sure that the type of question did not affect the answer length for short answers ( $F[1, 14] = 3.59, p = .08$ ), nor for long answers ( $F < 1$ ).

Answers to the medical questions were presented in six different presentation formats: a short and a long textual answer, each used (i) on its own (unimodal), (ii) combined with an informative picture (multimodal) and (iii) combined with a decorative picture (multimodal). In the remainder of this section, we only discuss the multimodal answer presentations.

Two multimodal answer presentations, a short and a long answer, contained a decorative picture. Figure 4 shows the short and the long answer to the question “How to organize a workspace in order to prevent RSI?”, illustrated with a decorative photograph showing a workspace. The other two multimodal answer presentations contained an informative picture. Figure 5 shows the short and the long answer to the same question as in Figure 4, but this time illustrated with an informative graphic. The graphic depicts an ergonomic workspace in detail. It should be noted that all answer presentations were designed in such a way that the textual element by itself already contained enough information to answer the question; the informative pictures only added relevant background information.

All answer presentations were presented to the participants in a random order, which was the same for all participants.

**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.



**Fig. 4** Examples of a short textual answer (top) and a long textual answer (bottom) with a decorative picture.


**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.

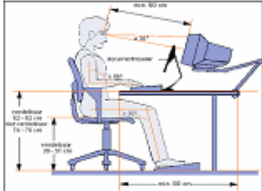


**VRAAG** 

**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 werkvladdiepte 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.



**Fig. 5** Examples of a short textual answer (top) and a long textual answer (bottom) with an informative picture.

### 3.4 Procedure

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 participants received an e-mail inviting them to take part in the experiment. This e-mail briefly stated the goal of the experiment, the amount of time it would take to participate, the possibility to win a gift certificate, and the URL of the experiment. When accessing the website of the experiment, participants received instructions about the procedure. Next, they entered their personal data (i.e. age, gender, level of education, and optionally their e-mail to win a gift certificate). After a short practice session, participants studied 16 question-answer combinations, one at a time. After each combination, they were shown the same combination with at the bottom five seven-point semantic differentials (implemented as radio buttons) which they had to use to rate the informativeness of the answer (the answer presentation is informative/ not informative), the attractiveness of the answer (the answer presentation is attractive/ not attractive), the informativeness of the text-picture combination (the text-picture combination is informative/ not informative), the attractiveness of the text-picture combination (the text-picture combination is attractive/ not attractive), and the clarity of the text (the text is formulated in a simple / complex way).

### 3.5 Results

Here we only report on the participants' assessment of the informativeness and the attractiveness of the text-picture combinations. For (partial) results on the other presentation aspects evaluated by the participants, see Section 5.6, where they are compared to the results of automatically illustrated presentations.

The results were tested for significance using a 4 (answer presentation)  $\times$  2 (question type) repeated measures analysis of variance (ANOVA). As shown in Table 3, short answers with an informative picture were evaluated as most informative, and short answers with a decorative picture as least informative ( $F[3, 68] = 9.32$ ,  $p < .001$ ,  $\eta_p^2 = .29$ ). Answers to *how* questions were rated as more informative than answers to *what* questions ( $F[1, 68] = 15.13$ ,  $p < .001$ ,  $\eta_p^2 = .18$ ). Finally, an interaction was found between answer presentation and question type ( $F[3, 68] = 4.27$ ,  $p < .01$ ,  $\eta_p^2 = .16$ ): for both short ( $F[1, 17] = 17.12$ ,  $p < .005$ ,  $\eta_p^2 = .50$ ) and long ( $F[1, 17] = 7.31$ ,  $p < .025$ ,  $\eta_p^2 = .30$ ) answers with an informative picture, answers to *how* questions were evaluated as more informative than answers to *what* questions. For answers with a decorative picture no significant differences were found between the two question types.

Long answers with an informative picture were evaluated as most attractive, long answers with a decorative picture were evaluated as least attractive ( $F[3, 68] = 4.64$ ,  $p < .01$ ,  $\eta_p^2 = .17$ ). Answers to *how* questions were evaluated as more attractive than

answers to *what* questions ( $F[1, 68] = 20.59, p < .001, \eta_p^2 = .23$ ). No interaction was found between answer presentation format and question type ( $F < 1$ ).

**Table 3** Mean results for the informativeness and attractiveness of answer presentation types (ratings range from 1 = “very negative” to 7 = “very positive”; standard deviations in parentheses).

Factor	Question type	Text with decorative picture		Text with informative picture	
		Short	Long	Short	Long
Informative	What	3.83 (.13)	4.01 (1.30)	4.91 (.81)	4.97 (1.20)
	How	3.70 (1.26)	4.27 (1.18)	5.53 (.70)	5.40 (.84)
	Total	3.76 (1.16)	4.14 (1.19)	5.22 (.69)	5.18 (1.00)
Attractive	What	3.93 (.87)	3.76 (1.14)	4.43 (.88)	4.69 (1.01)
	How	4.18 (1.12)	4.18 (1.10)	4.95 (.84)	5.08 (.76)
	Total	4.06 (.96)	3.97 (1.07)	4.69 (.75)	4.89 (.79)

### 3.6 Conclusion

The results show that answers with an informative picture were evaluated as more informative than answers with a decorative picture, especially for short answers, which is consistent with the production experiment (Experiment 1). The information load of the textual answers could explain these results. Short answers contain less information than long ones. Therefore, an informative picture adds more information to short answers than to long answers, and is thus perceived as more informative in combination with a short answer. Also, answers to *how* questions with an informative picture were evaluated as more informative than answers to *what* questions. Arguably, the medical procedures – as they occurred in this experiment – lend themselves better to be visualized than definitions, because they have a dynamic and spatial character. Interestingly however, long answers with informative pictures were evaluated as most attractive, suggesting that users like complete information together with highly informative pictures.

## 4 Automatic production of multimodal answers

In the previous sections, we discussed how humans produce and evaluate multimodal answers. However, most existing QA systems present their answers in one presentation mode, i.e. text snippets retrieved from a document corpus. Pictures that occur in the corpus documents are generally ignored, since the text-oriented retrieval methods used in QA systems cannot deal with them. A method for extending the answers returned by a QA-system with appropriate pictures has been proposed in [4].

In this section we describe the picture selection method, and in the next section we present a user evaluation (Experiment 3) in which the results of two variations of this method are compared with the manually created multimodal answer presentations used in Experiment 2.

### ***4.1 Multimedia Summarization***

Our approach to generating multimodal answers to questions is essentially automatic multimedia summarization, using established techniques from automatic text summarization. Most text summarization methods (used in the context of a QA system) are based on comparative analyses between the user’s query and parts of the source document(s). Multimedia summarization faces the difficulty that different media have different features and thus cannot be directly compared (e.g., the word “red” cannot be directly compared to the color red). Analyzing and converting media content to a semantic representation has been proposed as a solution for this problem [8, 15, 19, 20]. However, automatic analysis of media content is difficult and often unreliable. Manual annotation is an alternative which answers some of these objections, but this is very laborious. Another solution, which according to de Jong et al. [14] is often overlooked, is to use related linguistic content for analysis, instead of the media items themselves. If related text adequately describes a media item, text-based retrieval methods can be used to retrieve non-textual media.

We automatically generate multimedia presentations as answers to medical questions by using a query-based summarization framework ([5], this volume) in a multimedia setting. The query-based summarization framework relies on a combination of one or more feature graphs representing the source documents. A content unit can be a unit of any medium, such as a text snippet or a picture. The graphs express relations between the documents’ content units, and are constructed using content (e.g. cosine similarity, see the next section) or context (e.g. layout) to relate content units. This way, content can be presented for which there is only indirect evidence of relevance. For instance, a sentence that is adjacent – and thus contextually related – to a sentence that is similar to the query may be included in the answer, even though it is only linked to the query indirectly. This concept may also be applied to multimedia. A picture can be related to a piece of text by using layout information. A straightforward indication of relatedness of text and visual content is when the text is the picture’s caption, but the paragraph or section in which the picture is located may also be considered as related to the picture.

### ***4.2 Automatic Picture Selection***

In the IMIX system, the approach sketched above is used to select the best picture to illustrate a given textual answer to a medical question. To find this picture,

the illustration system compares the text of the answer with picture-associated text. The more similar the two text passages, the more likely the picture is relevant. The picture-associated text is interpreted as a textual representation of the picture. This may be either the picture’s caption or the paragraph (or section if no single paragraph could be related to the picture) in which the picture was found. The relevancy of a picture for the answer is calculated as:

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

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

Cosine similarity is a way of determining lexical similarity of text passages. The idea behind cosine similarity is that a text’s meaning is constituted by the meaning of its words. To measure cosine similarity between two passages, we represent both texts as a vector whose elements represent the contribution of a word to the meaning of the passage. Before measuring the cosine similarity, words are stemmed using Porter’s stemmer [21]. The cosine similarity is calculated as follows:

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

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

Determining how much a particular word contributes to the meaning of a passage is called *term weighting*. We use *tf · idf* term weighting, i.e. the contribution of a word to a passage is calculated as the word’s occurrence frequency in the passage (term frequency, TF) multiplied by the word’s inverse document frequency (IDF). IDF is a measure of how characteristic the word is for a passage. To measure the inverse document frequency, we require a large set of passages. For this we use the passage vectors of picture-associated text for all pictures in a medical corpus (see Section 5.3), plus the passage vector of the answer text. A word occurring in few of these passages receives a high IDF value, because the low occurrence rate makes it descriptive of the few passages it appears in. Conversely, a word occurring in many passages receives a low IDF value. The contribution of word  $k$  to passage  $a$  is measured as follows:

$$a_k = \text{tf}_{a,k} \cdot \text{idf}_k \quad (3)$$

Where  $\text{tf}_{a,k}$  is the number of occurrences of word  $k$  in passage  $a$ ; and  $\text{idf}_k$  is the IDF value of word  $k$ . The IDF value is calculated as follows:

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

Where  $|D|$  is the number of passages in the corpus (i.e. the number of pictures plus one); and the denominator is the number of documents which contain the word  $k$ . The final answer presentation consists of the textual answer and the most relevant picture and its caption.

Figure 6 shows an example of an answer presentation containing an automatically selected picture. In this figure and in Figure 7 the answer presentation is embedded in the web interface used for Experiments 2 and 3, which was designed to replicate the ‘look and feel’ of a medical QA system.



**Vraag 4/16**

*Bestudeer de hieronder afgebeelde medische vraag- en antwoordpresentatie zorgvuldig.*

**Wat zijn thrombolytica?**

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



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

Ga verder

**Fig. 6** Example of an answer presentation consisting of text and an automatically selected picture. The presentation answers the question “What are thrombolytics?” The text of the answer explains that thrombolytics are drugs used to dissolve blood clots. The picture depicts a schematic representation of clotted blood.



## 5 Experiment 3: Evaluating Automatically Produced Multimodal Answers

We carried out an experiment to evaluate two variants of the previously described approach for automatically adding pictures to textual answers. The study was largely identical to Experiment 2, except that we used automatically retrieved pictures instead of manually selected ones. More details on the experiment can be found in [6].

### 5.1 Participants

Seventy five people participated (44 female and 31 male). Their average age was 22 ( $SD = 7.11$ ,  $min = 18$ ,  $max = 55$ ). Fifty six of them (75%) were students recruited from Tilburg University. None had participated in the previous two experiments.

### 5.2 Design

The experiment had a  $16$  (question)  $\times$   $2$  (short or long answer)  $\times$   $2$  (retrieval method: using caption or section) mixed factorial design, with the question as a within participants variable and the answer length and retrieval method as between participants variables. The dependent variables were the same as in Experiment 2, i.e., the participants' assessment of: (a) the clarity of the text, (b) the informativeness of the answer presentation, (c) the attractiveness of the answer presentation, (d) informativeness of the text-picture combination and (e) the attractiveness of the text-picture combination. The participants were randomly assigned to an experimental condition.

One of the goals of Experiment 3 was to compare the automatically illustrated answer presentations to the manually created answer presentations used in Experiment 2; therefore we re-used the same design. Experiment 2 used manually selected pictures only, and relevance of the pictures was assumed. In contrast, some of the automatically selected pictures used in Experiment 3 were irrelevant, either because there was no appropriate picture in the database or simply because the algorithm failed to find one. However, choosing to use the same design for both evaluation experiments meant that in Experiment 3, the participants judged the informativeness of the text-picture combinations instead of directly assessing the relevance of the automatically selected pictures.

### 5.3 Stimuli

In our study, we used the same set of 16 general medical questions that had been used in Experiment 2, with the same short and long textual answers. The textual answers were illustrated with automatically retrieved pictures using the algorithm described in Section 4. The pictures were retrieved from a repository of medical pictures that had been automatically extracted from two medical sources. Each of the pictures in the repository had two corresponding textual annotations: the first annotation represented the caption of the picture in the original document, and the second represented the paragraph (or section) in which the picture was found.

The pictures and their annotations were extracted from two medical sources intended for a general audience and written in Dutch, providing information about anatomy, processes, diseases, treatment and diagnosis. The first source, *Merck Manual medisch handboek* [2], *Merck* in short, contains 188 schematic illustrations of anatomy and treatment, process schemas, plots and various types of diagrams. The other source, *Winkler Prins medische encyclopedie* [9], *WP* in short, contains a variety of 421 pictures, including photographic pictures, schema's and diagrams. These sources were selected because they cover the popular medical domain and they are relatively structured - paragraph boundaries are marked in the text and all 609 pictures have captions. The pictures have a high information density; only few pictures are decorative. Consequently, the pictures are relatively specific to their context, which complicates their reuse in a slightly different context.

For each of the textual answers, two answer presentations were generated. For one of the presentations, the picture was retrieved using its caption as associated text, and for the other the picture was retrieved based on the smallest unit of surrounding text (paragraph or section) from the original document of the picture. Regardless which text was used for selecting the picture (caption or surrounding text), the caption was always presented together with the picture in the answer presentation. However, in order to prevent excessive caption lengths, captions were truncated to their first sentence during presentation generation (the remaining sentences were used for retrieval but not in the presentation). If the surrounding text (section in short) was used for picture selection, this text was not included in the answer presentation. The corpus did not contain an appropriate picture for all answers, which forced the illustration system to select less appropriate pictures for some of the presentations. In some cases the selected picture was plain irrelevant, but in some other cases, the picture was related to the text but had a different perspective. For instance, the picture in Figure 7 addresses the deformation of red blood cells rather than their generation. In our estimation (not formally validated) around 30% of the automatically selected pictures used in Experiment 3 were irrelevant, in the sense that they had absolutely no connection with the answer text. For example, a picture of egg and sperm cells was selected to illustrate an answer about RSI. The other pictures were either fully relevant, such as the picture in Figure 6, or somewhat relevant, such as the picture in Figure 7.



**Fig. 7** Example of a picture which is related to, but not fully relevant for, the answer text. The presentation answers the question “Where are red blood cells generated?” The text explains that red blood cells are generated from stem cells in the bone marrow. Rather than illustrating this, however, the picture shows various deformations of red blood cells.

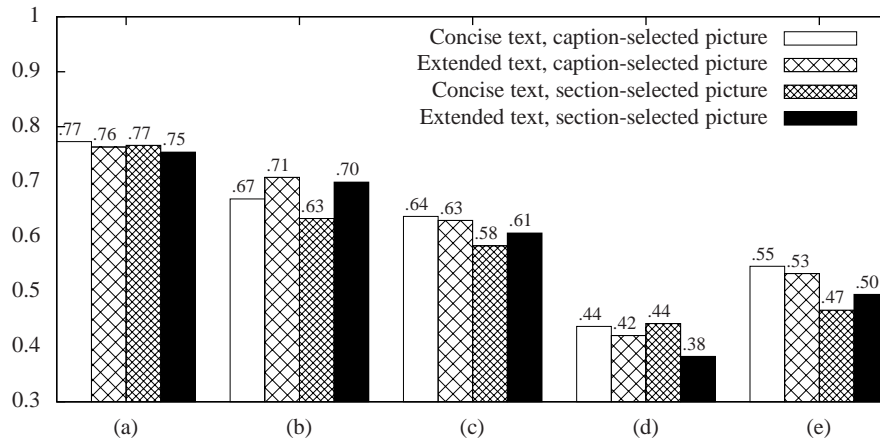
#### 5.4 Procedure

The procedure was identical to the procedure of Experiment 2; see Section 3.4.

#### 5.5 Data Processing

The results of the assessments were normalized to be in the range  $[0 \dots 1]$ . A rating  $n$  between one and seven (inclusive) was normalized as  $(n - 1)$ . For processing the results, the following non-standard method was used. For each condition and each medical question and assessment question, the average assessment was calculated. For pair-wise significance testing of differences between two experimental conditions for a particular assessment question, the percentage of answer presentations was measured for which the rating of one condition was higher than that of another. A condition that consistently received higher average ratings than the other for each medical question got a score of 100%; consequently, the other condition got a relative score of 0%. Significance was tested by means of 106-fold approximate randomization. A difference is considered significant if the null hypothesis (that the sets are not different) can be rejected at a certainty greater than 95% ( $p < .05$ ), unless stated otherwise.

The reasons for using the mutual rank instead of the average judgment were that the standard deviation of ratings of answers to some medical questions was higher than the standard deviation for answers to other medical questions. As a result, some



**Fig. 8** Average assessments of (a) textual clarity; (b) informativeness of the answer presentation; (c) attractiveness of the answer presentation; (d) informativeness of the text-picture combination, and (e) attractiveness of the text-picture combination.

medical questions affected the average rating more than others. This made it less likely to find significant differences in the average rating. Using the mutual rank avoided this problem.

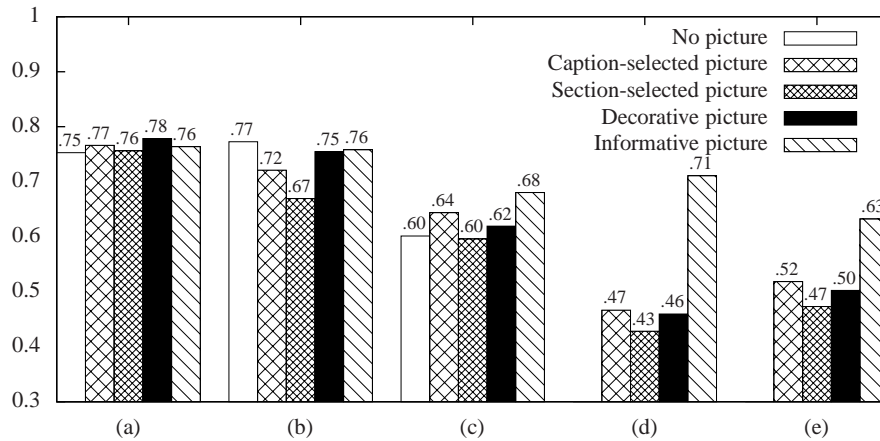
## 5.6 Results

### *Caption versus Section*

Figure 8(a) shows that the level of clarity of the textual component of the answer was judged similar. No significant differences between the conditions were found. Figure 8(b) indicates that for the informativeness of the presentation, long answers were rated significantly more informative than short answers. However, for long answers, the combination of picture and text (Figure 8(d)) was judged less informative. This difference was biggest for section-selected pictures, although not significant. Figure 8(c) and (e) show that the presentation as well as the picture-text combination were rated significantly more attractive if the pictures were selected by their captions than when the surrounding section was used for picture selection. No differences were found between short and long textual answers in the attractiveness of the presentation or the picture-text combination.

### *Automatically versus Manually Selected Pictures*

The results of two experiments are comparable only if the group of participants in one experiment is similar to the participants of the other experiment. In both Ex-



**Fig. 9** Average assessments of (a) textual clarity; (b) informativeness of the answer presentation; (c) attractiveness of the answer presentation; (d) informativeness of the text-picture combination, and (e) attractiveness of the text-picture combination. For comparability, these results include only registered students from Tilburg University. Therefore, the actual values may differ slightly from Figure 8.

periments 2 and 3, students and non-students took part and their answers to some of the assessment questions were significantly different. Therefore, to enable comparing the results of the two experiments, the group of non-students was filtered out in order to ensure that the experimental conditions were the only variables over both experiments. In total, 98 people (70 female, 28 male) who participated in either Experiment 2 or 3 were registered students. Forty-two of them contributed to Experiment 2 and 56 contributed to Experiment 3. The average assessments of the 98 participants are shown in Figure 9.

These results combine the 16 short and the 16 long answer presentations, comprising 32 data points for each condition and assessment question. They include the unimodal condition from Experiment 2, which was not discussed in Section 3.

For informativeness of the answer presentation, no significant differences were found between answer presentations with a caption-selected picture and answer presentations with a manually selected informative picture. However, answer presentations with a section-selected picture were rated as significantly less informative than answer presentations with a manually selected informative picture, a decorative picture, or no picture at all. For attractiveness of the answer presentation, no significant differences were found between answer presentations with an automatically selected picture (either caption- or section-based), a manually selected decorative picture, or no picture at all. We measured no significant effect of the presence of (different types of) images on the user's perception of the clarity of the text.

The informativeness as well as the attractiveness of the text-picture combination was not significantly different between answers with an automatically selected

picture (either caption- or section-based) or a manually selected decorative picture. However, the informativeness of the text-picture combination was rated significantly higher for answer presentations with a manually selected informative picture than for answer presentations with an automatically selected picture or a manually selected decorative picture. Participants also found manual informative pictures more attractive than any other category in combination with the text.

Average ratings of automatic presentations may have been negatively affected by inconsistent performance of the picture selection algorithm. If the relevance of automatic pictures is less consistent than that of manual pictures, this should be reflected in the variability of the results. Indeed we found that for automatic pictures, participants showed greater variability than for manual pictures in their assessments of textual clarity, informativeness and attractiveness of the answer presentation.

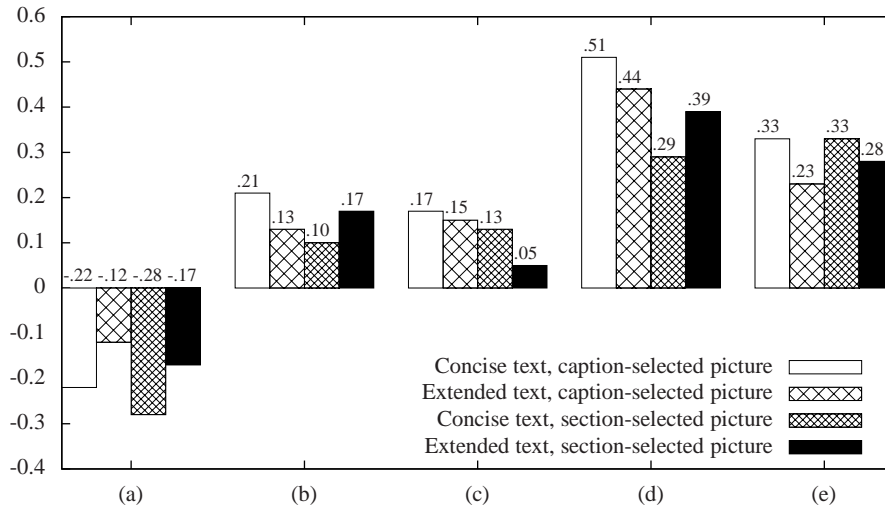
#### *Cosine Similarity as Indicator of Picture Relevance*

The selection criterion for automatic pictures was the cosine similarity of the textual component of the answer and the text associated with the picture (a caption or a section depending on the condition). The picture with the highest cosine similarity was selected. Because cosine similarity is used as a measure of relevance, this value can be interpreted as a confidence value, i.e. how confident the system is that the selected picture is actually relevant. In the IMIX system, in which this picture selection method is implemented, the answer is presented text-only if no picture has a confidence (cosine similarity) above a certain configurable threshold. Table 4 shows the averages of the cosine similarity values of the pictures selected for the answers in this experiment.

**Table 4** Statistics of the cosine similarity of the textual component of the answer and the text passage used for indexing the selected picture.

Condition	Average	Standard deviation	Range
Brief text; caption-selected picture	0.190	(0.00788)	[0.0687,0.347]
Extended text; caption-selected picture	0.188	(0.00631)	[0.0786,0.397]
Brief text; section-selected picture	0.133	(0.00501)	[0.0295,0.311]
Extended text; section-selected picture	0.162	(0.00654)	[0.0373,0.319]

But what is the meaning of cosine similarity as a confidence value? Cosine similarity can be used to predict the relevance of the picture if there is a correlation between the cosine similarity and the experimental participants' judgments of a presentation. Figure 10 shows the correlation of the confidence (cosine similarity) value and the participant judgments. A value of 1 (or -1) indicates a perfect increasing (or decreasing) linear correlation. This correlation was greatest for the participant judgments of the informativeness of the text-picture combination (.51 and .44 with short and long answer texts respectively). This is an encouraging result, given that this



**Fig. 10** Pearson correlation coefficient between the confidence (cosine similarity) of picture selection and the assessments of (a) textual clarity; (b) informativeness of the answer presentation; (c) attractiveness of the answer presentation; (d) informativeness of the text-picture combination, and (e) attractiveness of the text-picture combination.

aspect seems to correspond most closely to picture relevance. With respect to attractiveness, the correlation with confidence was significantly greater for short answers than for long answers. There was only a slight difference in correlation between attractiveness and confidence for different picture selection methods.

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

## 5.7 Conclusion

The results of the evaluation experiment indicate that the caption-based picture selection method results in more informative and attractive presentations than the section-based method, although the difference in informativeness was not significant. Furthermore, caption-based picture selection shows a greater correlation between confidence and informativeness, which indicates that the confidence value better predicts the informativeness of the picture. Compared to manually created

answer presentations, we found that answer presentations with an automatically selected picture were largely rated at the same level as presentations with a manually selected decorative picture or even no picture at all. This is not entirely surprising. In Experiment 3, the manually selected pictures used in Experiment 2 were used as a gold standard for decorative and informative pictures respectively. However, in practice, it is unlikely that this gold standard could be achieved with the set of 609 pictures from our medical corpus, because the picture sources used by the participants in Experiment 1 (which formed the basis for the answer presentations in Experiment 2) were unrestricted and thus offered far more opportunities to find a suitable illustration for a given answer text.

Finally, an investigation of the relation between system confidence and our experimental results revealed a negative correlation between textual clarity and the predicted relevance of the selected illustration.

## 6 General Discussion

This chapter described three experiments in which we investigated which (combinations of) presentation modes are most suitable for the answers of a medical QA system. In Experiment 1, we were interested in the spontaneous production of multimodal answers to medical questions. The results showed that people used pictures more frequently when producing long answers. Informative pictures were more frequently used in short answers, while representational pictures were most frequent in long answers. It is likely that when the answer does not contain much text, a picture will contain additional information with regard to the text. When the answer contains much text, it is likely that a picture adds less information to it (i.e. it visually represents the information already present in text). Short answers contained more decorative pictures than long answers, possibly because lack of room for discussing pictured information in short answers led the participants to add simple illustrations, requiring no textual explanation, more often than when creating presentations with long answers.

Also, people used decorative pictures more frequently in the answers to *what* questions. Informative pictures on the other hand occurred most often in the answers to *how* questions. Possibly, in textual answers to *what* questions the picture represented an element of the question. Pictures in the answers to *how* questions were often used to explain the steps within the procedure and therefore added information to the textual answer.

In Experiment 2, we concentrated on how people evaluate different multimodal (text and a picture) answer presentations on their informativeness and attractiveness. The results showed that answers with an informative picture were evaluated as more informative than those with a decorative picture. Moreover, *how* answers with informative pictures were evaluated as more informative than *what* answers with informative pictures. An explanation for this result could be that medical procedures – as they occurred in this experiment – lend themselves well to being visualized



as they have a temporal and spatial character. Definitions on the other hand often contain abstract concepts which are less easily visualized.

Another interesting result is that while short answers with an informative picture were evaluated as most informative, long answers with an informative picture were evaluated as most attractive. The information load of the textual answers might explain these results. Short and long textual answers differ in their information density, i.e. short answers contain less information than long ones. Therefore, an informative picture has more added value for short answers than for long answers, increasing the perceived informativeness of the short answer presentations. On the other hand, an informative picture adds relatively less information to a long textual answer and therefore primarily serves to enhance the attractiveness of the presentation.

In Experiment 3, we conducted a user evaluation in which two versions of the automatic picture retrieval method were compared: caption-selected illustrations versus section-selected illustrations. The caption-based picture selection method resulted in more informative and attractive answers than the section-based method, although the difference in informativeness was not significant. Furthermore, caption-based picture selection showed a greater correlation between confidence and informativeness, which indicates that the confidence value better predicts the informativeness of the picture. A system could use this to respond by not offering any picture if no relevant picture is available (as done in the IMIX system). All in all, the caption-based picture selection method offers more promising results than the section-based selection method.

When compared to manually created answer presentations, we found that answer presentations with an automatically selected picture were largely rated at the same level as presentations with a manually selected decorative picture (which did not add any information to the answer) or even no picture at all. This may be partially explained by the design of the experiment, where the visual element of the answer presentations was not needed to answer the question, since the textual element contained all the required information. Also, the results were undoubtedly influenced by the fact that our picture corpus did not contain appropriate pictures for all answers, in which case the algorithm had no choice but to select an irrelevant picture.

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

## References

1. Arens, Y., Hovy, E., Vossers, M.: On the knowledge underlying multimedia presentations. In: M. Maybury (ed.) *Intelligent Multimedia Interfaces*. AAAI Press (1993)

2. Berkow, R., Beers, M.H., Fletcher, A.J. (eds.): Merck manual medisch handboek, 2nd edn. Bohn Stafleu van Loghum, Houten, the Netherlands (2005)
3. Bernsen, N.: Foundations of multimodal representations. a taxonomy of representational presentation mode. *Interacting with Computers* **6**(4), 347–371 (1994)
4. Bosma, W.: Image retrieval supports multimedia authoring. In: E. Zudilova-Seinstra, T. Adriaansen (eds.) *Linguistic Engineering meets Cognitive Engineering in Multimodal Systems*, ICMI Workshop, pp. 89–94. ITC-irst, Trento, Italy (2005)
5. Bosma, W., Marsi, E., Krahmer, E., Theune, M.: Text-to-text generation for question answering. In: G. Bouma, A. van den Bosch (eds.) *Interactive Multi-modal Question Answering*. Springer Verlag (2010)
6. Bosma, W., Theune, M., van Hooijdonk, C., Krahmer, E., Maes, A.: Illustrating answers: an evaluation of automatically retrieved illustrations of answers to medical questions. In: *Proceedings of the AISB Symposium on Multimodal Output Generation (MOG 2008)*, pp. 34–41 (2008)
7. Carney, R., Levin, J.: Pictorial illustrations still improve students' learning from text. *Educational Psychology Review* **14**(1), 5–26 (2002)
8. van Deemter, K., Power, R.: High-level authoring of illustrated documents. *Natural Language Engineering* **2**(9), 101–126 (2003)
9. Fiedeldij Dop, P., Vermeent, S. (eds.): *Winkler Prins medische encyclopedie*, 3rd edn. Spectrum (1974)
10. Heiser, J., Phan, D., Agrawala, M., Tversky, B., Hanrahan, P.: Identification and validation of cognitive design principles for automated generation of assembly instructions. In: *Proceedings of Advanced Visual Interfaces*, pp. 311–319 (2004)
11. van Hooijdonk, C., Krahmer, E.: Information modalities for procedural instructions: the influence of text, static and dynamic visuals on learning and executing rsi exercises. *IEEE Transactions on Professional Communication* **51**(1), 50–62 (2008)
12. van Hooijdonk, C., Krahmer, E., Maes, A., Theune, M., Bosma, W.: Towards automatic generation of multimodal answers to medical questions: a cognitive engineering approach. In: *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*, pp. 93–104 (2007)
13. van Hooijdonk, C., de Vos, J., Krahmer, E., Maes, A., Theune, M., Bosma, W.: On the role of visuals in multimodal answers to medical questions. In: *Proceedings of the 2007 Conference of the IEEE Professional Communication Society*. IEEE (2007)
14. de Jong, F.M.G., Westerveld, T., de Vries, A.P.: Multimedia search without visual analysis: the value of linguistic and contextual information. *IEEE Transactions on Circuits and Systems for Video Technology* **17**(3), 365–371 (2007)
15. Maybury, M.T., Merlino, A.E.: Multimedia summaries of broadcast news. In: *1997 IASTED International Conference on Intelligent Information Systems*. IEEE (1997)
16. Mayer, R., Moreno, R.: Aids to computer-based multimedia learning. *Learning & Instruction* **12**(1), 107–119 (2002)
17. Mayer, R.E.: *The Cambridge handbook of multimedia learning*. Cambridge University Press, Cambridge (2005)
18. Michas, I., Berry, D.: Learning a procedural task: effectiveness of multimedia presentations. *Applied Cognitive Psychology* **14**(6), 555–575 (2000)
19. Nagao, K., Ohira, S., Yoneoka, M.: Annotation-based multimedia summarization and translation. In: *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7. Association for Computational Linguistics, Morristown, NJ, USA (2002)
20. Petrushin, V.A.: *Introduction into Multimedia Data Mining and Knowledge Discovery*, pp. 3–13. Springer London (2007)
21. Porter, M.: An algorithm for suffix stripping. In: K.S. Jones, P. Willet (eds.) *Readings in Information Retrieval*, pp. 313–316. Morgan Kaufmann (1997)
22. Tversky, B., Morrison, J., Bétrancourt, M.: Animation; can it facilitate? *Int. J. Human-Computer Studies* **57**(4), 247–262 (2002)
23. Veenker, T.: WWStim: A CGI script for presenting webbased questionnaires and experiments (2005). Website: <http://www.let.uu.nl/~Theo.Veenker/personal/projects/wwstim/doc/en/>