MARIËT THEUNE
BORIS VAN SCHOOTEN
RIEKS OP DEN AKKER
WAUTER BOSMA
DENNIS HOFS
ANTON NIJHOLT
University of Twente
Human Media Interaction
Enschede, The Netherlands
{h.j.a.opdenakker | b.w.vanschooten | m.theune | w.e.bosma | d.h.w.hofs | a.nijholt}@utwente.nl

EMIEL KRAHMER
CHARLOTTE VAN HOOIJDONK
ERWIN MARSI
Tilburg University
Communication and Cognition
Tilburg, The Netherlands
{e.j.krahmer | c.j.m.vanhooijdonk | e.c.marsi}@uvt.nl

Questions, Pictures, Answers: Introducing Pictures in Question-Answering Systems

Abstract

We present the Dutch IMIX research programme on multimodal interaction, speech and language technology. We discuss our contributions to this programme in the form of two research projects, IMOGEN and VIDIAM, and the technical integration of the various modules developed by IMIX sub-projects to build a demonstrator that shows the project's aims and achievements. In this paper we pay special attention to the role that images could play in iterative question answering.

1 IMIX: multimodal interactive question answering

This paper describes research performed at the Human Media Interaction group of the University of Twente, in cooperation with the Communication and Cognition group at Tilburg University, in the framework of the IMIX programme. IMIX is a research programme in the field of Dutch speech and language technology funded by the Netherlands Organisation for Scientific Research (NWO). It started in 2004 and brings together research on question answering (QA), dialogue management, speech recognition and generation, and information presentation.

While early research in the field of QA concentrated on answering factoid questions[1] one by one in an isolated manner, recent research extends QA in several new directions. One of these directions is towards developing iterative QA or QA dialogue systems that allow users to ask follow-up questions and that can help users to clarify questions (Kato et al., 2006). Several projects that are part of the ARDA AQUAINT program concern scenario-based QA, the aim of which is to build systems that can handle non-factoid, explanatory, analytical questions posed by users with extensive background knowledge (Williams et al., 2004). Most of these systems interact with the user by means of text only. LIMSI's Ritel system is a spoken dialogue system for open domain QA, but it only deals with factoid questions (Galibert et al., 2005).

The IMIX programme further extends iterative QA in the direction of multimodal QA. The system is primarily text-based, but it allows users to use speech input as well as text input, and the answer is presented in the form of text, speech and pictures. The next step is to allow users to refer to the pictures in the answer presentation when asking follow up questions. How can QA, dialogue, and multimodality be combined? How can such a multimodal iterative QA system help users find the information they need?

A number of Dutch research groups participate in the IMIX programme, collaborating to build a common demonstrator. This demonstrator is meant as a vehicle to prove that the research carried out in the individual projects can be applied and integrated in the context of a QA system. The IMIX demonstrator is an iterative multimodal QA system aimed at providing general medical information. The system will be able to deal with more complex questions than simple factual ones, and it will be able to engage in a simple dialogue with the user aiming at obtaining a better understanding of the question or at a more equal level of communication between the user and the system.

Our Human Media Interaction group participates in two of the IMIX research projects: IMOGEN (Interactive Multimodal Output Generation), together with Tilburg University, and VIDIAM (Dialogue

[1] Factoid questions are those for which the answer is a single fact. For example "When was Mozart born?" "How tall is the Eiffel Tower?"

Management and the Visual Channel). A common interest of these two projects is the use of images in QA dialogues. Multimodal answer presentation, combining text, speech and pictures is the focus of IMOGEN, while VIDIAM researches the ways that users could interact with the system by pointing at pictures in their follow-up utterances, and how the dialogue manager could meaningfully react on these multimodal utterances.

This paper is structured as follows. First, we provide a general overview of the different projects within the IMIX programme by discussing the functionality and the architecture of the IMIX demonstrator in section 2. Then, in sections 3 and 4 we zoom in on the respective contributions of the IMOGEN and VIDIAM projects. In section 5 we discuss the role that images could play in iterative question answering, and the research that has been done so far on this topic within IMOGEN and VIDIAM. We outline our future work in this area in section 6, and end with some conclusions in section 7.

2 The IMIX demonstrator: functionality and architecture

The IMIX demonstrator is a system that helps users finding information they need in the medical domain. The system is presented as an Information Search Assistant, i.e., more like a librarian who helps users to find the medical information they need than a medical expert (see op den Akker et al., 2005). This Information Assistant is personified in the form of a Dutch-speaking version of the Rutgers University Talking Head (RUTH).[2]

Regarding the nature and difficulty level of the questions it is supposed to handle, particular choices have been made. We try to provide a coverage that is sufficient to answer the information needs of real, non-expert users. We call the type of questions that we cover "encyclopaedic questions". These are general medical questions that do not require expert medical knowledge (this excludes diagnostic questions). Nevertheless, correctness of answers is less well-defined and harder to measure than with factoid questions. Our system typically answers with a piece of text about 1-3 sentences long. The length and level of detail of a correct answer depends on the information need of the user, which cannot be precisely known beforehand. Enabling the user to give feedback on her information need is one of the purposes of our dialogue system. Although IMIX focuses on a closed medical domain, many of the techniques used are also applicable to open domain QA. This is especially true of the dialogue management module. It uses only little domain knowledge, which makes it easy to generalise to other domains.

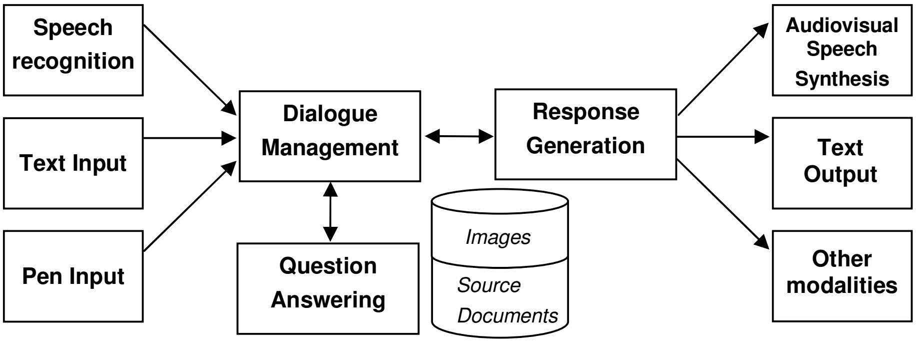


Figure 1: Global architecture of the IMIX demonstrator.

A global architecture of the IMIX demonstrator is shown in Figure 1. The user can enter a question to the system in Dutch, either by typing or speaking. Both input modes can be combined with pen input (pointing or drawing). The speech recognition module of the demonstrator is provided by the NORISC project (University of Nijmegen), which focuses on advanced methods for Dutch speech recognition (Hämäläinen et al., 2005; Han et al., 2005). Pen input is dealt with in the VIDIAM project (this paper and van Schooten, these proceedings).

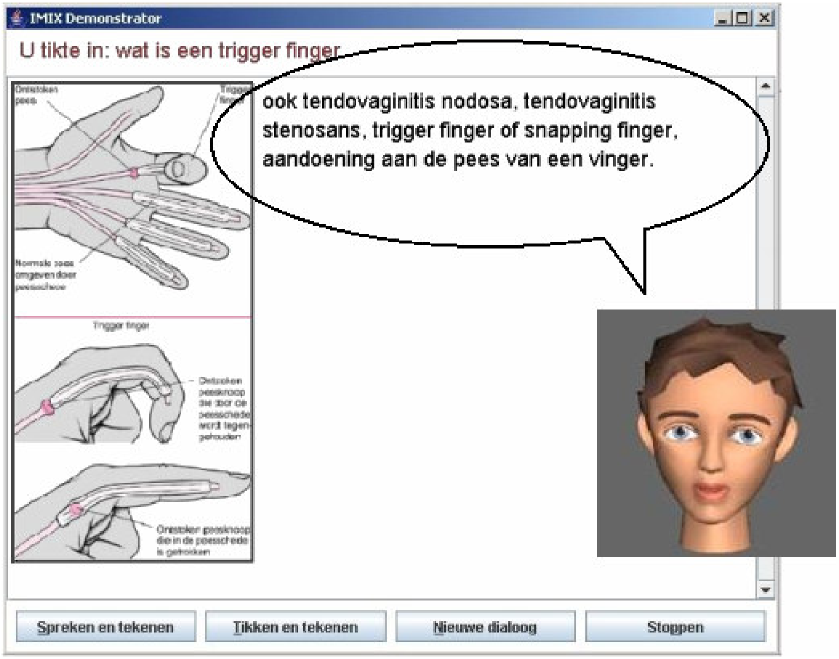
The dialogue manager analyses the user's input and determines how to deal with it. If the user has posed a domain question, this is sent to QA for answering. As the dialogue proceeds the user can also react with other utterances than questions, such as an acknowledgment of an answer presented by the system, or an indication that she doesn't understand the answer. This implies that not all follow-up utterances are domain questions that can be sent to QA. In addition, follow-up questions may contain anaphoric references that need to be rewritten to form a self-contained question that can be handled by QA. Within IMIX, two projects deal with dialogue management. PARADIME (Tilburg University) focuses

[2] http://www.cs.rutgers.edu/~village/ruth/

on the analysis and generation of dialogue acts (Keizer & Bunt, 2006), while VIDIAM focuses on the processing of follow-up utterances (van Schooten & op den Akker, submitted) and on interaction via the visual channel (van Schooten, these proceedings).

The task of the QA component of the demonstrator is to find possible answers to the user's question in a large set of medical documents, consisting of information from two different medical encyclopaedias, various websites and other sources. The IMIX demonstrator works with three different QA engines, provided by the research projects ROLAQUAD (Tilburg University), QA-DR (University of Groningen) and FactMine (University of Amsterdam). To find matches for the user's question, ROLAQUAD uses semantic document tags resulting from a pre-analysis of documents and questions (Lendvai, 2005). QA-DR determines the syntactic dependency structure of the question and matches it with the (pre-parsed) dependency structures of the source documents (Bouma et al., 2005). FactMine aims at developing unsupervised methods for the extraction of fact bases and ontological information from text (Tjong Kim Sang et al., 2005). These QA engines are considered as black boxes by the dialogue manager, which is independent of the strategy the QA engines follow in answer retrieval. We assume that a dialogue manager will only be able to give relatively simple and general hints to the QA back-end, such as pointers to the questions that should be considered context for the current question. The QA system is then responsible for the details of implementing this dialogue context into its QA query.

The QA engines return a ranked list of text snippets, usually consisting of one sentence, that should contain an answer to the question. This list is sent to the IMOGEN module, which generates the answer text that is presented to the user. It decides which of the potential answers to use (currently simply the first on the list) and how to extend this 'core' answer so that the output is a comprehensive and verifiable answer to the question (Bosma, 2005a). It also decides whether it is appropriate to include an image with the text (Bosma, 2005b). In addition, the text of the answer is presented to the user by means of a Dutch-speaking talking head. See Figure 2 for an illustration.



IMIX Demonstrator

U tikte in: wat is een trigger finger

ook tendovaginitis nodosa, tendovaginitis stenosans, trigger finger of snapping finger, aandoening aan de pees van een vinger.

Ontstoken pees
Trigger finger
Normale pees omgeven door peesschede
Trigger finger
Ontstoken peesknoop die door de peesschede wordt tegengehouden
Ontstoken peesknoop die in de peesschede is getrokken
Spreken en tekenen
Tikken en tekenen
Nieuwe dialoog
Stoppen

Figure 2: Example output of the IMIX demonstrator. Question (transl.): "What is a trigger finger?" Answer (transl.):"Also tendovaginitis nodosa, tendovaginitis stenosans, trigger finger or snapping finger, infection of the tendon of a finger." (The text balloon has been added to indicate that the text shown on the screen is also presented in speech by the animated character.)

The modules in the demonstrator have been integrated using Multiplatform, an integration platform previously used for the SmartKom and COMIC projects (Herzog et al., 2003). It enables modules, possibly written in different programming languages, to exchange XML messages. This is done using a publish-and-subscribe system. A module does not send a message directly to another module, but rather sends it to a message pool. Other modules can subscribe to a pool to receive messages. The system includes a mechanism for asynchronous request and response messages. Multiplatform visualises the modules and pools in a configurable diagram while the demonstrator is running. The integration work, including the development of a graphical user interface in which the user can pose questions and see the answers, has been carried out at the University of Twente. The demonstrator as a complete system runs on Linux and requires a rather large amount of resources. In order to make the demonstrator more accessible, we have also developed a more lightweight Java-only client, together with a multi-client server. This client is available as a Java Webstart at this address: http://wwwhome.ewi.utwente.nl/~hofs/imix/webstart/.

3 IMOGEN: going beyond simple answers

As mentioned in the introduction, in IMIX we go beyond the answering of simple factoid questions. An important category of questions in the medical domain is that of definition questions such as “What is RSI?” or “What is a trigger finger?” (as in Figure 2). Definition questions, unlike factoid questions, require a more complex answer, often constructed from multiple source documents. Ideally, the answer should be a short paragraph which succinctly defines the definiendum: the thing, be it person, organization, object or event which the user wishes to know more about (see Greenwood, 2005).

Most QA engines typically retrieve answers consisting of a single sentence, which is a step forward compared to the single word or phrase traditionally used to answer a factoid question. However, a longer answer is generally even more helpful. As shown by Lin et al. (2003), increasing the amount of text returned to users significantly reduces the number of queries that they pose to a QA system. This suggests that users also utilize information that is related to the ‘core’ answer to fulfil their information need. In addition, including additional text in the answer may allow the user to assess the accuracy of the answer extraction, and thus to verify whether the answer is correct. In IMOGEN we have developed various methods to extend the answer provided by QA in order to make it more informative and verifiable. We first discuss purely textual answer extensions (section 3.1), and then we describe how textual answers can be enriched with pictures to achieve multimodal presentations (section 3.2).

3.1 Extending answers based on discourse structure

The simplest method to extend a single-sentence answer as provided by QA is by simply taking the sentences that directly surround it in the source document and adding these to the answer presentation. However, these sentences are not necessarily those that are most relevant given the user’s question. Therefore we have developed a more sophisticated method of answer extension that makes use of the discourse structure of the source document. In brief, the discourse structure of the document is converted to a graph representation in which nodes represent sentences, and edges represent the relations between them. The strength of the relation between sentences (i.e., between nodes in the graph) is calculated as the length of the shortest path that connects them in the discourse graph. Extending the answer is done by selecting those sentences that bear the closest relation to the ‘core answer’ provided by QA. These are not necessarily the sentences that are linearly adjacent to it in the source text. We performed a comparative evaluation of answer presentations created using this method (Bosma 2005a), where the discourse graphs of the source texts were based on a handcrafted Rhetorical Structure Theory analysis (RST, Mann & Thompson 1988). See Figure 3 for an example RST structure and the corresponding discourse graph, where letters represent sentences (‘A’ being the core answer sentence) and numbers the strength of the relation between nodes, computed based on the distance between them, combined with sentence lengths (in number of words) and size of the subordinate text span (in number of sentences), which both provide an indication of topic importance. The evaluation results showed that the discourse-based answers were considered significantly more verifiable and contained less irrelevant information than those where simply the directly surrounding context had been included in the answer.

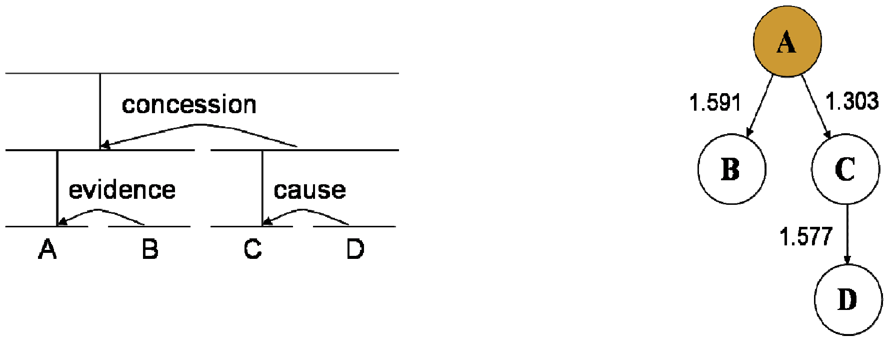


Figure 3: RST structure and a corresponding discourse graph.

With some adaptations, this method can also be used to create extended answers combining sentences from different source documents, e.g., as retrieved by different QA engines. Bosma (2006) describes a method for doing this using a knowledge-poor method, where discourse structure is automatically determined based on document lay-out, and where redundancy between documents is detected using the dependency tree alignment algorithm of Marsi et al. (2006). Another application of this alignment algorithm is sentence fusion, where partially overlapping sentences are merged to form a new sentence (Marsi & Krahmer, 2005). When applied in the IMIX demonstrator, this would mean that the system would be no longer limited to producing extractive answers (using the original sentences from the source texts) but could also produce abstractive answers, where the answers returned by QA are combined and rewritten for optimal information density. This is a topic for future work, however.

3.2 Extending answers with images

Cognitive research into multimedia has shown that good use of extra modalities in addition to text may substantially contribute to the user's understanding, recall and processing efficiency of the presented material (Levin, 1981; Mayer, 2001). Therefore we have developed a method for extending the answers returned by QA not only with additional text, as described above, but also with appropriate images (Bosma, 2005b). Using the text of the answer as input query, from a corpus of pictures we retrieve the picture that best matches this query. We call this 'answer-based image retrieval'. In terms of Inoue (2004), this is a form of 'query-by-text' image retrieval, which is best carried out based on textual annotations of the images to be retrieved.[3]

A problem with the use of annotations for image retrieval is that manually annotating an image corpus is very time-consuming, whereas automatic image annotation is difficult to achieve. For this reason many approaches to text-based image retrieval make use of existing captions and/or file names for retrieval. However, these provide only limited information. Therefore we opt for using another information source: the 'scope' of the image in its original document; i.e., the span of text in the document that is illustrated by the image. Annotating the scope of an image involves selecting a text fragment from the original document containing the image. Although this is not a trivial task, it is less time-consuming than full annotation. Possibly, image scopes could even be detected automatically based on textual references to the image in the document ("see picture", "Figure X shows", etc.).[4]

Our method is based on the assumption that when two texts (or text fragments) are similar, it is appropriate to reuse a picture associated with one of the texts in the other. In short, what we do is compare the text of the answer we want to illustrate (the query) with the scopes of the images in the corpus. The picture of which the scope is most similar to the query, is included in the answer presentation. This process is illustrated in Figure 4. The numbers in the figure indicate the similarity between the query (on the left) and the scopes of the images in the corpus (on the right), computed using Latent Semantic Analysis (see below). In this example, the image in the middle will be selected because it represents the strongest match with the query.

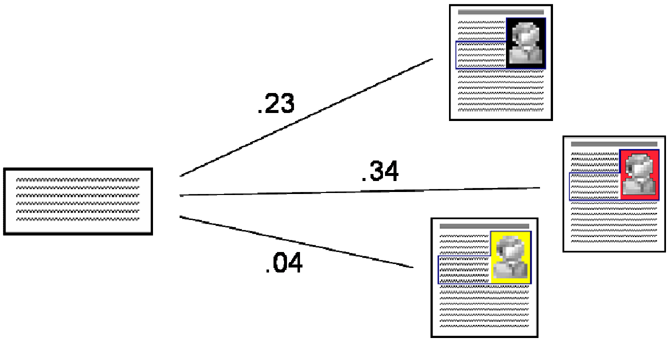


Figure 4. Answer-based image retrieval based on scope matching.

To compute the similarity between the query and the image scopes, we use Latent Semantic Analysis (LSA, see Landauer et al. 1998), an algorithm for measuring text similarity that is widely adopted in information retrieval. It uses word occurrence matrices of the texts in a corpus, which are generalized using deep statistical analysis to construct a 'semantic space' representing the text collection. Within this space, two texts can be compared for similarity. (These texts do not need to come from the corpus used to construct the semantic space.) LSA is a knowledge-poor method, as it takes raw text as input and makes no use of human-constructed knowledge bases, syntactic or morphological information, etc. This makes it attractive for general use. However, it would be interesting to see how the use of LSA for measuring similarity compares with the use of other methods based on the linguistic information that is

[3] This stands in contrast with 'query-by-image', where the goal is to find an image that is similar to some input image, and which may be carried out based on image content, i.e., low-level visual features of the images (Inoue, 2004).
[4] This was suggested to us by Eduard Hovy (p.c.)

already available in IMIX. For instance, Bosma & Callison-Burgh exploit the dependency analyses from QA-DR (see section 2) to measure text similarity using the dependency tree alignment algorithm of Marsi et al. (2006).

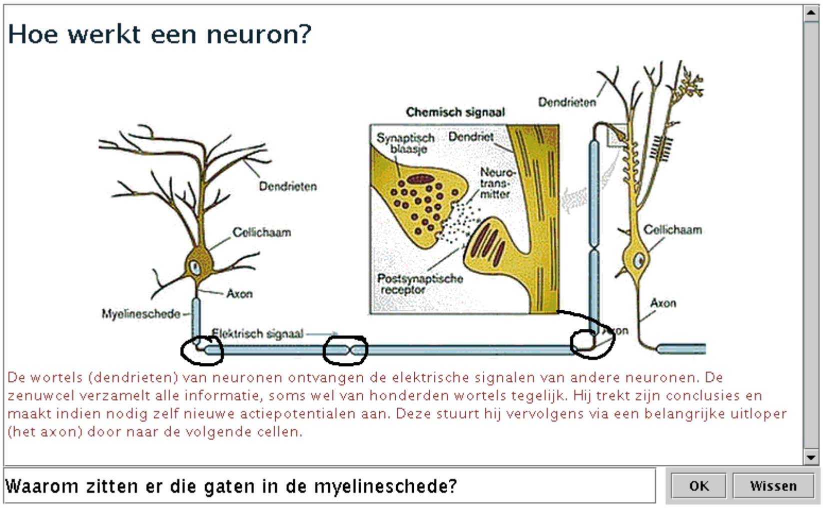
The image retrieval method described here has been applied in the IMIX demonstrator, using a very small corpus of medical images. We are planning to perform formal evaluations on a larger corpus, see section 6.

4 VIDIAM: what do users do in a QA dialogue?

The research in interactive QA (or QA dialogue, or information seeking dialogue) as an extension to QA has just begun. There is little experience with real users and real applications that shows us how people interact with a system as is envisaged in IMIX. Even less is known about the possible contribution of multimodality in this interaction (that is, with the system answering with pictures, and the user able to point at visual elements on the screen).

The VIDIAM project explores these new areas. In particular, it goes some way towards handling domain questions in context (i.e., follow-up questions), multimodal questions (questions where the user points at a visual element on the screen), and reactions of users that are not domain questions. Within the IMIX demonstrator, VIDIAM develops a dialogue manager module that serves as a mediator between the user and the QA engine which has only limited capabilities of handling dialogue. Although VIDIAM does concern itself with system-initiative dialogue (such as clarification questions posed by the system), our emphasis here will be on follow-up utterances from the user in reaction to the answers presented by the system.

In order to find out how users may follow up in a QA dialogue, we collected two corpora of follow-up utterances. One corpus is about free-form follow-up reactions, which may or may not be follow-up questions (the second-utterance corpus, containing 575 follow-up utterances from 40 users on 120 question-answer pairs), and the other is specifically about multimodal follow-up questions to a multimodal answer (the multimodal second question corpus: 202 follow-up questions from 20 users on 20 question-answer pairs).


Hoe werkt een neuron?

De wortels (dendrieten) van neuronen ontvangen de elektrische signalen van andere neuronen. De zenuwcel verzamelt alle informatie, soms wel van honderden wortels tegelijk. Hij trekt zijn conclusies en maakt indien nodig zelf nieuwe actiepotentialen aan. Deze stuurt hij vervolgens via een belangrijke uitloper (het axon) door naar de volgende cellen.

Waarom zitten er die gaten in de myelineschede?

Figure 5. An example of a multimodal follow-up question. At the top is the original question (translated: "How does a neuron work?") and in the middle is the answer. In the box at the bottom is the user's follow-up question (translated: 'Why are there these holes in the myelin sheath?'). The encirclings in black are the mouse strokes made by the user.

We used a special method for collecting these corpora, which is low-cost and specifically tailored for follow-up utterances in QA systems. Instead of having an actual dialogue with the user (using either a dialogue system or a Wizard of Oz setup), we have the user react to a set of canned question/answer pairs. The first user turn does not consist of posing a free-form question, but of selecting a question out of an appropriately large set of interesting questions. The second turn of the user consists of posing a follow-up utterance to the answer then presented (see Figure 5). The dialogue simply ends when the user posed his/her follow-up utterances, and there is no need for the system to reply to the follow-up utterances, hence there is no dependency of the corpus on any dialogue strategies used. Obvious weaknesses of the method are that the dialogues are only two turns long, and the second turn is never a question that starts a new topic. However, we found that our corpora contain many of the most important phenomena that a dialogue system will need to support. We argue this is a good low-cost corpus collection method for bootstrapping a QA dialogue system.

In the corpora we found a number of interesting things. In the second-utterance corpus, we found that 44% of the follow-up utterances were not follow-up questions. Besides follow-up questions, the other major classes we found were:

Negative questions and statements (20%). Negative feedback was most often given to incorrect answers (note that 18% of the given answers were deliberately incorrect). Linguistically they took different forms: questions repeating the original question, statements such as “No, I didn’t mean that”, and a variety of other forms. Only a minority of these utterances seemed to contain additional information beside the negative feedback. It is not clear how such information might be used, but we found that negative feedback can effectively be used to determine if the answer is correct.

Q: Wat zijn hartkloppingen?	*Q: What are heart palpitations?*
A: De patiënt neemt dit waar als hartkloppingen.	*A: The patient experiences this as heart palpitations.*
Q: Maar wat zijn hartkloppingen dan?	*Q: But what are heart palpitations?*

Negative question example

Verify questions (3.6%). These indicate that the user is not sure about the meaningfulness or correctness of the answer. There were fewer of these than of negative feedback, indicating that users are usually able to determine whether an answer is correct or not. A dialogue system might react by (further) extending the answer so that the correctness can be verified appropriately, see section 3.1.

Q: Hoe merk je dat je hoge bloeddruk hebt?	*Q: What do you notice when you have high blood pressure?*
A: Hoge bloeddruk (hypertensie) is meestal een aandoening die over het algemeen geen symptomen veroorzaakt.	*A: High blood pressure (hypertension) is usually an affliction that generally does not cause any symptoms.*
Q: Dus je merkt niet dat je een hoge bloeddruk hebt?	*Q: So you don’t notice anything when you have high blood pressure?*

Verify question example

Reformulations (4.4%). These are rephrasings of the original question, usually without any linguistic cues as to the fact that the question is a reformulation of a previous question, rather than a question without any dialogue context. These occurred most often in reaction to cases where the system stated that no answer could be given to the initial question (which comprised 6% of the cases). Reformulations also occurred when the answer was incorrect or inappropriate.

Acknowledgments (13%). Most of them are of the form “Okay” or “Thanks”. They indicate that the user is satisfied with the answer.

Only 2.4% of the follow-up utterances remained as not classifiable in any of these classes. About half of these were meta-questions, like questions about the source of the document.

The remaining 56% of follow-up utterances were follow-up questions, that is, the questions that can meaningfully be interpreted as domain questions, but which are to be understood in the context of the dialogue.

4.1 Handling follow-up questions

The two corpora we collected contain both unimodal and multimodal follow-up questions, which we can use as a basis for a wide-coverage follow-up question handling method. In the corpora we encounter a range of phenomena which can be potentially handled by several different basic techniques. We consider the following techniques to be particularly useful and tractable:

- Rewriting a follow-up question to include the relevant context so that it becomes a self-contained question. This context may include previous questions and answers, and the content of pictures presented in the answers. This is a well known method in the literature (Fukumoto et al., 2004).
- Considering the document that the previous question was taken from as likely to contain the answer to the follow-up question. A generalisation of this can be made by also considering other candidate documents that the QA engine might have found (as typical QA systems retrieve a set of documents that contain potential answers). Searching within the previously retrieved documents was shown to be a strategy successfully employed by other projects (De Boni & Manandhar, 2005). It may be implemented by passing a signal to the underlying QA engine that the documents retrieved for the previous question(s) should be used.

- Answering a question about a picture directly, using available information about that picture. In order to be able to apply this technique, the system should have appropriate information about the picture; in particular, annotations specifying what it describes and what visual elements are in it.

Each of these techniques is appropriate for certain kinds of follow-up questions. To find out what the potential of these techniques is, we classified the follow-up questions according to their appropriateness w.r.t. specific instances of these techniques. For each class, we will discuss how it could be handled by the dialogue manager, and how often utterances of that class occurred in the corpus.

Self-contained: the answer does not need any rewriting before making it answerable by QA. While such a question may be submitted to a QA engine directly, the previous document search technique may still be applied here, as long as the question is still within the context of the current topics. Note that we do not have topic shifts (that is, questions starting a new topic in the middle of the dialogue) in our corpus; hence we will not look at the issue of topic shift detection here. Some 25% of the unimodal and 18% of the multimodal follow-up questions were self-contained.

Rewritable-simple: the answer is potentially rewritable to a self-contained question using a simple, specific, syntactical transformation. A dialogue manager may handle this kind of question by detecting which transformation is applicable and finding the appropriate referents and inserting them. The previous document search technique is of course still applicable as a complementary technique. The transformations that we looked at were anaphor substitution (22% of the unimodal follow-up questions were of this class, and 22% of the multimodal ones, see example below), prepositional phrase attachment (10% of unimodal, but only 1% of multimodal), and ellipsis completion (7% of unimodal, and none were found in the multimodal corpus).

Q: Wat produceren ribosomen?	*Q: What do ribosomes produce?*
A: Eiwit.	*A: Protein.*
Q: Wat voor eiwit produceren ze?	*Q: What kind of protein do they produce?*
→ Wat voor eiwit produceren ribosomen?	*→ What kind of protein do ribosomes produce?*

Anaphor substitution example

Rewritable-complex: the answer is potentially rewritable, providing that we may reformulate it in an arbitrary way to include the dialogue context. This class is not considered very suitable for automatic rewriting, since that will likely be too complex, but the fact that reformulation is possible at all does show that these questions are meaningful follow-up questions. The previous document search technique is particularly suitable in this case. We found that 23% of the unimodal and 14% of the multimodal follow-up questions were of this class.

Visual-element-identity (multimodal only): not rewritable, but answerable by giving the identity of a particular visual element of a picture in the last answer. For example, questions like “What is this?” or “Is this the …”? while indicating a visual element. This type of question is not a meaningful QA question since it deals with properties of a specific picture. This class of questions could be answered directly by simply annotating all visual elements in each picture with appropriate names. We found that 20% of the multimodal questions were of this type.

Visual-property (multimodal only): not answerable by a QA engine because it has specifically to do with the content of a picture in the last answer, while it is not just about the identity of a visual element. For example, a user may point at a part of a picture and say: “What exactly is happening here?” Figure 5 is an example of this class. This is a difficult type of question to handle properly, but might simply be answered by producing a suitable figure caption paraphrasing the entire figure. We found that some 26% of multimodal follow-up questions were of this type.

Discourse-property (unimodal only): not answerable by a QA engine because it has specifically to do with the text discourse of the last answer. We found that some 13% of the unimodal follow-up questions were of this type. Usually, these refer to the fact that the answer text fragment refers to something that is accidentally left out of the fragment; hence we formerly named this class “*missing-referent*”. This type of question can be appropriately answered by showing more of the text that the answer came from, e.g., using the method of van Langen (2005), who describes an approach to answer extension that ensures no referents are missing from the answer text. More complex discourse-related questions occasionally occur, however, such as questions about very specific assumptions or ambiguities in the answer.

Q: Waar duiden koude voeten op?	*Q: What do cold feet indicate?*
A: Hierdoor kan een verhoogde bloeddruk ontstaan of een vernauwing van de (…)	*A: This can cause elevated blood pressure or a constriction of the (…)*
Q: Waardoor?	*Q: What can?*

Discourse property question example

4.2 Dialogue manager architecture

We are developing a dialogue manager that uses the techniques suggested in the previous section to handle follow-up utterances. The dialogue manager has a processing pipeline consisting of several stages (see Figure 6).

First, a user utterance, consisting of text or speech output with mouse strokes, is pre-processed. The linguistic output is parsed by a broad-coverage dependency tree parser and a POS tagger, and appropriate words are tagged with their general domain types (disease, person, symptom, etc.). The gesture recogniser matches the mouse strokes with the visual elements that they may refer to. For identification of visual elements within pictures, these pictures have to be annotated with the location and name of the visual elements contained in them. This pre-processed input is then fed to the multimodal fusion and reference resolution modules, which produce a set of reference hypotheses using a referent database containing the essential information of all (linguistic and deictic) referents encountered so far.

This processed data is fed into the utterance type recogniser. It classifies user utterances according to the classes described above: negative feedback, acknowledgements, verify-questions, and the various kinds of follow-up questions. According to the class found, the dialogue manager either reacts directly to the user, or passes the question on to the question rewriter, which decides if and how to rewrite the question. The question rewriter passes the rewritten question, along with other possible context hints, to the QA. Both the question rewriter and the QA may reply with specific errors and warnings, to which the action generator may react appropriately.

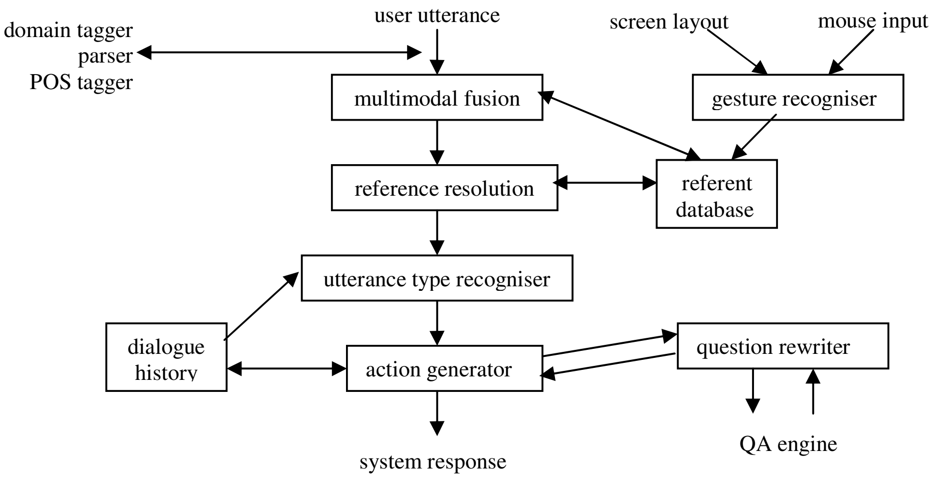


Figure 6. Dialogue manager architecture.

We shall summarise some of the results we have obtained. For more details, the reader is referred to Van Schooten (these proceedings) and Van Schooten & Op den Akker (2006). We first look at gesture recognition and reference resolution. Reference resolution of regular anaphors turned out to be relatively easy, since we found that, in the unimodal corpus, 75% of their antecedents were found in the initial question, so we could give priority to these antecedents. The algorithm could point out the correct antecedent for about 70% of the relevant anaphors. We do not have any anaphor resolution results on the multimodal corpus yet, but we do have preliminary results on the performance of relating gestures to visual elements, which is the first step in multimodal anaphor resolution. This performance was good, *providing* that the gestures were of the type that our gesture recogniser handles (namely, encircling gestures), and the visual elements were known to the system. After hand-annotating the visual elements found in the pictures in the corpus, our gesture recogniser identified 88% of the user's encircling gestures to the correct visual referents. However, 16% of these gestures pointed to referents that were not annotated or not meaningfully annotatable. Also, though the use of encircling gestures was encouraged, 32% of all gestures were not encircling gestures. Extending our recogniser with the other types of gestures we found is work in progress.

The next processing stage is utterance type classification, which we performed using a rather basic classification approach, namely by looking at the occurrence of specific words, phrases, and POS tags. Our system achieved this with mixed performance. Recognising the non-follow-up questions (acknowledge, negative, verify) was relatively easy, and we correctly identified about 70% of the non-follow-up questions in the corpus, with few false positives. Recognising the different classes of follow-up question was more difficult. For the unimodal corpus, we managed an overall performance of only 50%

over the total of 6 follow-up question classes (self-contained, the three kinds of rewritable-simple, rewritable-complex, and discourse-property). For the multimodal corpus, we did not manage better than 40% over the total of 6 classes (non-multimodal, self-contained, the two rewritable classes, and the two visual classes).

The next stage after utterance type recognition is rewriting those questions that were identified as rewritable. This proved to be difficult. Tackling anaphor substitution proved to be the most fruitful: the system managed to correctly rewrite 42% of the questions it labeled as rewritable anaphors. We did not get any useful results with the other types of syntactic transformation we identified. We do not yet have any rewriting results on the multimodal corpus.

We also looked at the potential of the previous document search strategy, by performing the QA manually for each follow-up question by searching the document that each answer was taken from. We found that we could answer 73% of the discourse-property questions, and 35% of the other follow-up questions, by manually selecting an appropriate text fragment from the document that the last answer came from. Since the documents were rather small, only a few paragraphs long on average, this may be considered a good performance for such a simple strategy. Even in its simplest form of only looking at one document, the strategy seems fruitful as a QA technique, as long as it is complemented with other QA techniques.

5 The role of images in QA Dialogues

In the previous sections we have described how QA answers can be enriched with images and how users interact with multimodal answer presentations. Questions we have not yet addressed, but which are quite relevant for our research include: What is the function of images in an answer presentation? What types of answers lend themselves to being extended (or possibly even replaced) with an image? What type of images is suitable for what type of answers? In section 5.1 we present some functions of pictures in relation to text that have been proposed in the literature, and we discuss why we should take these functions into account for multimodal answer presentation. In 5.2 we present a first sketch of how this could be done using an image annotation scheme that should be useful for both generating multimodal answer presentations and for the resolution of multimodal references to these presentations. Finally, in section 5.3 we look at other modalities and modality combinations than just text and pictures, presenting some experiments that we carried out to investigate which modalities are best suited for the presentation of physical exercises to prevent Repetitive Strain Injury.

5.1 The relation between text and pictures

Pictures can have different functions in relation to a text. Carney & Levin (2002) distinguish five functions: decorational, representational, organizational, interpretational and transformational. Decorational pictures make a text more attractive, but they do not enhance understanding or recall (and might even distract the reader). Representational pictures mirror the text content, and can help the user understanding a text by making it more concrete. According to Carney & Levin they are the most commonly used type of illustration. Organizational pictures provide a structural framework for the text content, for example maps. Interpretational pictures function as clarifiers of material that is difficult to understand, e.g., by depicting abstract processes. Finally, transformational pictures have a mnemonic function, e.g., showing a picture of a tailor in combination with a text about a Mr. Taylor to enhance memorization of the name. This is a very specific function, which seems irrelevant for the medical information domain and which we will ignore in this paper.

Picture functions such as those outlined above should be taken into account when generating multimodal information presentations. For instance, knowledge about picture functions may be useful for setting retrieval thresholds during answer-based image retrieval (see section 3.2), i.e., for deciding whether a certain picture is sufficiently relevant to be included in the answer. Interpretational pictures may contribute much to the user's understanding of the information that is presented, but this also means that choosing the 'wrong' interpretational picture in answer-based image retrieval will be quite detrimental, so the relevance threshold should be set quite high. In contrast, decorational pictures do not contribute to understanding, but for this function there is a much lower risk associated with including an irrelevant picture. (However, even for decorational pictures there must always be some relationship with text content. A picture of a penguin would not be suitable to decorate a medical text; whereas an Aesculapius symbol would.)

In addition, certain types of pictures match certain types of information. Interpretational pictures appear to be most helpful in combination with texts describing cause and effect systems or processes; an example from the medical domain would be the working of neurons, which is illustrated in Figure 5. Organizational pictures are appropriate when the answer contains a spatial element, for instance a description of the location of various muscles in the body. Representational pictures are useful in all answers referring to concrete body parts, possibly showing the symptoms of some disease. Of course questions that directly ask for visual information, e.g., "What does the spleen look like?" should always be answered with a representational picture.

This leads to the conclusion that we need to annotate picture functions in the image corpus used for retrieval. This is discussed in the section 5.3. In addition, we will need more information about the answer (and the question it answers) to determine which kind of picture is most appropriate. One information source that might be used for this are the semantic tags assigned to question and answer by the ROLAQUAD QA engine. This semantic coding includes tags for concepts such as <body-part> and <symptom> (which could be visualized representational pictures) and processes such as <causes>, <prevents> and <treats> (which could be clarified using an interpretational picture). Alternatively, it might be sufficient to detect general question patterns such as “How does X work?” or “What does Y look like” rather than having to rely on domain-specific semantic annotations. Investigating this is part of our future work, discussed in section 6.

5.2 Annotating pictures

What is relevant to annotate in a picture corpus depends on the application in which the pictures are used. Do we need a picture for producing an enhanced answer presentation? Or do we need to annotate pictures for reference by the user in posing follow-up questions, so that the dialogue system knows what part the user is referring to when she asks a follow-up question? To deal with the latter issue, a picture annotation scheme has been developed that is based on an analysis of the corpus of multimodal follow-up utterances described in section 4.1 (see also van Schooten, these proceedings). The annotation is needed to determine the visual referents of the users’ pen strokes and to resolve linguistic references to graphical properties, such as shape and colour.

We chose to annotate both the image as a whole, and its visual elements individually. We annotated each visual element with its location (a polygon describing its contour), a set of keywords, a noun phrase describing it unambiguously and, additionally, colour, shape, and function. We chose to annotate keywords, colour, and shape because users tend to use these to linguistically refer to visual elements. The function label distinguishes the overall function the element has in the picture. We found that, for anatomical images, only a few function types are needed to be able to describe almost all visual elements. These are: graphical (a physical object), annotation (text labels identifying physical objects), callout (a line connecting a text label to a visual element), flow (typically arrows indicating direction of transport or movement), and caption (since many images include captions and titles).

For the image as a whole we annotate a set of keywords, a caption, a noun phrase describing the image, and, optionally, a prepositional phrase describing the special context of an image. This last annotation requires some explanation. Many medical illustrations are anatomical images, and we found that these usually simply describe anatomy, but were often related to a specific disease or treatment. We consider this disease or treatment the special context. A prepositional phrase for such a disease may be for example “of an Alzheimer patient”. If we have an anatomical element in a picture, say, an element described as “a neuron”, then we can unambiguously specify the element by attaching its description to the special context, i.e., “a neuron of an Alzheimer patient”.

What if we also want to use these annotations for multimodal answer generation / answer-based image retrieval? Several parts of the proposed annotation scheme seem quite useful for both tasks: captions and descriptions of the picture and its elements provide information on the semantic content of the picture, which is obviously relevant for image retrieval. Some of the information annotated with an eye to reference resolution seems less relevant for image retrieval, however. Examples are the use of colours and the exact placement of elements in the image. This information could simply be ignored. Finally, for the purpose of multimodal answer generation, it would be good to incorporate some additional information in the annotation scheme. First, obviously we would like to stick to the use of scope information for image retrieval (see 3.2), as the scope provides more clues for matching pictures with answers than captions and noun phrase descriptions. In addition, as pointed out above, information about the function of the picture in relation to its scope should be annotated. Several options exist for the annotation of picture functions.

- We could use the broad functional categories of Carney & Levin (2002), discussed in section 5.2: decorational, representational, organizational, interpretational and transformational.
- A more detailed alternative is the function taxonomy of Mash & White (2003), which groups functions into three global categories that could be characterised as decorational, representational + organizational, and interpretational. In total, this taxonomy contains 49 different functions, where one image can have multiple functions.
- André (1995) and Bateman et al. (2002) have developed discourse structure representation schemes that cover both text and images. These have the advantage that they fit in with the RST-based discourse structure representations used for textual answer extension as described in section 3.1.
- An alternative theory for document description that covers graphics is the Text-Graphics Discourse Theory (TGDT). It builds on earlier approaches to text discourse but adds four novel components: 1. Representation of the semantics of graphics. 2. Non-sequential access by the reader (attention shifts). 3. Bi-modal, text-graphics lexicons, mental models and reasoning by

the reader. 4. The production of an integrated model of the discourse (see Futrelle and Rumshisky, 2006).

In section 6 we describe some planned empirical research that should help us determine which of these (or other) annotation schemes describing the relation between images and text is most suitable for multimodal presentation generation.

5.3 Modality choice in information presentation

In the context of the IMIX demonstrator, the basis for any multimodal information presentation is formed by the (possibly extended) text snippet returned by QA. This means that in the demonstrator, pictures or other visual media such as film clips or animations[5] will always have a secondary status with respect to the text: they are selected to support the text, and not vice versa. However, presentation as text + images may not be the best choice for every type of information. In some cases, other modalities or modality combinations may be more suitable. For example, a picture might be most suitable as the main modality to express information with a strong spatial component. Therefore, we carried out a number of experiments on the choice of modalities for presenting a specific kind of medical information: instructions on how to perform physical exercises for the prevention of RSI (repetitive strain injury).

In the first experiment, we focused on unimodal presentations (van Hooijdonk & Krahmer, submitted). The different modalities we used were text, pictures and film clips; see Figure 7 for an example of a text (given in translation) and a picture used in this experiment. Participants were presented with instructions on how to perform RSI prevention exercises with different degrees of difficulty, presented using one of these modalities. They could study the instructions as long as they pleased, after which they performed the exercise. The results of this experiment showed that participants in the picture condition had the shortest learning times, and participants in the text condition had the longest learning times. Performance, measured in terms of the number of correctly performed exercises, was highest for participants in the film clip condition. Even though the general trends are similar for easy and difficult exercises, the results for different media vary much more for the difficult than for the easy ones, which indicates that modality choice should also be sensitive to complexity.

In a follow-up experiment, the participants were presented with all three forms of instruction for a number of exercises, and had to indicate which presentation mode they preferred. Most participants preferred the film clips. Given that pictorial instructions were found to be more efficient, this subjective preference might be explained by the film clips' higher visual appeal.

Stretch out your right arm, with the back of your right hand facing you.

Put your left hand on the palm of your right hand.

With your left hand, push the fingers of your right hand toward you.[6]

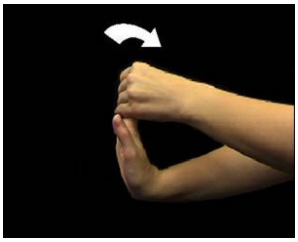


Figure 7. Presentation of exercise instructions by means of text and static image.

Finally, we performed an experiment in which we compared the results of the unimodal presentation conditions described above with those of two multimodal presentation conditions (van Hooijdonk & Krahmer, 2006). In the multimodal conditions, the RSI prevention exercises were presented to the participants by means of film clips + text or pictures + text. These multimodal presentations were created by combining the unimodal presentations from the previous experiment with each other, without any modifications. It turned out that the results for the multimodal conditions were consistently worse than those for the unimodal conditions, both in terms of learning times and performance. These results might be attributed to Mayer's Modality Principle (Mayer & Moreno, 2002), which tells us that it is better to accompany animations with speech than with text, since unlike text and animation, speech and animation are not competing for the same (visual) channel. Presumably, this principle also applies to text and pictures. Another possible explanation, however, is that the combined modalities in the multimodal condition were not optimally adapted to each other. Both text and picture or film clip were originally designed to be the main carrier of information, resulting in presentations where the modalities were competing with each other rather than complementing each other. It seems likely that, given the

[5] In principle, it should be possible to retrieve these using the same method currently used for pictures (section 3.2).

[6] Original text in Dutch: Strek uw rechterarm met de rug van uw rechterhand naar u toe. Leg uw linkerhand op de handpalm van uw rechterhand. Duw met uw linkerhand de vingers van uw rechterhand naar u toe .

spatial nature of the exercises being presented, a multimodal presentation where the image is central and the text only has a supportive role would have given better results.

In short, using knowledge about the functions of pictures in relation to text to generate proper text + picture combinations (as discussed in sections 5.1. and 5.2) may be a good starting point for multimodal information presentation, but the experiments discussed here have shown that for some types of information text is not suitable as the main modality. What is really needed to generate optimal multimodal presentations is additional knowledge about the strengths and weaknesses of each modality in presenting information, and rules for calculating and generating complementary modality combinations that do not necessarily have text as the main carrier of information. A basis for a sophisticated approach to modality choice is provided by work such as that of Arens et al. (1993) on media allocation, and that of Bernsen (1994) on the characteristics of representational modalities. A system that already follows such an approach is RIA (Zhou et al. 2005). This system dynamically creates multimodal presentations by matching the features of available output modalities with the features of the information to be expressed by the system. However, unlike IMIX, RIA is not a QA system but a dialogue system that operates in a limited domain, where the available data and media are known in advance. How this approach can be applied in a QA system, where information has to be retrieved from a large data collection, is still an open question.

6 Future work

The next step in IMOGEN's research into the automatic generation of multimodal answer presentations will be to carry out two large-scale production and evaluation experiments, loosely following the approach of Heiser et al. (2004) for the generation of instructions. The goal of these experiments is to gain insight into questions such as, which modalities or modality combinations do people use to express which types of information, which type of pictures do they select for their presentations, and which types of presentations are most appreciated by users. In our first experiment we will present a large number of human subjects (students) with a set of questions from the medical domain, and ask them to create a multimodal answer presentation for each of the questions. They can use the World Wide Web to find the required information and suitable pictures. (We will direct them to some of the same sources that are used in the IMIX demonstrator, which are also available on the web.) We will annotate and analyse the resulting answer presentations. In our second experiment we will ask another group of human subjects to rate (a selection of) these answer presentations, as well as a number of automatically generated presentations. This will provide us with an evaluation of our answer-based image retrieval method, and give us an idea of which general characteristics of multimodal presentations are more or less appreciated. Given that the production experiment results in a sufficient number of suitable presentations, we would also like to use the human-generated presentations as a standard against which different image retrieval techniques can be objectively evaluated, by measuring to which extent they pick out the same images from a corpus that were selected by human users (taking into account that there is never only one suitable image).

Other directions of future research include the use of audiovisual cues to express the system's degree of certainty about the correctness of the answer, and the use of sentence fusion techniques to combine multiple answer sentences into one, more informative answer.

The next step in VIDIAM's research will be developing the gesture recogniser and handling the newly identified types of multimodal follow-up questions, in particular the visual-element-identity and visual-property types. System performance can be readily evaluated using our existing corpora, described in section 4.1. We also wish to collect a new corpus of follow-up utterances, covering the complete range of non-follow-up questions, unimodal follow-up questions, and multimodal ones, using the question-answer pairs that will be made available by IMOGEN's production experiment.

7 Summing up

In this paper we have presented our ongoing research toward a new generation of QA systems that employ multiple modalities (rather than text only) to address the user's information need, and that allow the user to have a meaningful, multimodal dialogue with the system rather than posing a series of isolated questions. Research questions we try to answer include, what role do images play in interactions in which a user is looking for information? How can QA, dialogue, and multimodality be combined? And how can a multimodal iterative QA system help users to find the information they need? We have taken the first steps toward answering these questions by investigating the suitability of different modalities and combinations of modalities to present certain types of information, developing methods to retrieve images that are relevant to a given answer text, investigating how users react to answer presentations, and developing methods to handle unimodal and multimodal follow-up questions. The results of our research have been (and will be) implemented in a fully functional multimodal QA system, developed as a demonstrator within the Dutch research programme on Interactive Multimodal Information Extraction (IMIX).

Acknowledgement We thank Yulia Bachvarova for her comments on section 5 of this paper. The IMIX programme is funded by the Netherlands Organisation for Scientific Research (NWO).

References

Rieks op den Akker, Harry Bunt, Simon Keizer and Boris van Schooten. 2005. From question answering to spoken dialogue: towards an information search assistant for interactive multimodal information extraction. In: Proceedings of InterSpeech 2005, Lisbon, Portugal, pp. 2793-2796.

Elisabeth André. 1995. Ein planbasierter Ansatz zur Generierung multimedialer Präsentationen. PhD thesis, DISKI-108, INFIX Verlag, Sankt Augustin, Germany.

Yigal Arens, Eduard Hovy and Mira Vossers. 1993. On the knowledge underlying multimedia presentations. In Mark T. Maybury (editor), Intelligent Multimedia Interfaces. AAAI Press, Menlo Park, CA, pp. 280-306.

John Bateman, Judy Delin and Renate Henschel. 2002. XML and multimodal corpus design: experiences with multi-layered stand-off annotations in the GeM corpus. LREC'02 Workshop: Towards a Roadmap for Multimodal Language Resources and Evaluation, Canary Islands, Spain.

Niels Ole Bernsen. 1994. Foundations of multimodal representations: a taxonomy of representational modalities. Interacting with Computers 6(4):347-371.

Marco De Boni and Suresh Manandhar. 2005. Implementing clarification dialogues in open domain question answering. Journal of Natural Language Engineering 11(4):343-361.

Wauter Bosma and Chris Callison-Burch. 2006. Paraphrase substitution for recognizing textual entailment. To appear in: Proceedings of CLEF 2006, Alicante, Spain.

Wauter Bosma 2006. Query-based extracting: how to support the answer? In: Proceedings of DUC 2006, New York City, NY, USA, pp. 202-208.

Wauter Bosma. 2005a. Extending answers using discourse structure. In: Proceedings of the RANLP Workshop on Crossing Barriers in Text Summarization Research, Borovets, Bulgaria, pp. 2-9.

Wauter Bosma. 2005b. Image retrieval supports multimedia authoring. In: Proceedings of the ICMI Workshop on Multimodal Interaction for the Visualization and Exploration of Scientific Data, Trento, Italy, pp. 89-94.

Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas and Jörg Tiedemann. 2005. Question answering for Dutch using Dependency Relations. In: Proceedings of CLEF 2005, Vienna, Austria.

Russell N. Carney and Joel R. Levin. 2002. Pictorial illustrations *still* improve students' learning from text. Educational Psychology Review, 14(1):5-26.

Junichi Fukumoto, Tatsuhiro Niwa, Makoto Itoigawa and Megumi Matsuda. 2004. RitsQA: List answer detection and context task with ellipses handling. Working notes of the Fourth NTCIR Workshop Meeting, Okyo, Japan, pp. 310-314

Robert P. Futrelle and Anna Rumshisky. 2006. Discourse structure of text-graphics documents. In: Proceedings of the 1st symposium on Smart Graphics, Hawthorne, NY, USA, pp. 31-38.

Olivier Galibert, Gabriel Illouz and Sophie Rosset. 2005. Ritel: an open-domain, human-computer dialog system. In: Proceedings of InterSpeech 2005, Lisbon, Portugal, pp. 909-912.

Mark Andrew Greenwood. 2005. Open-Domain Question Answering. Ph.D. thesis, Department of Computer Science, University of Sheffield, UK.

Annika Hämäläinen, Johan de Veth and Lou Boves. 2005. Longer-length acoustic units for continuous speech recognition. In: Proceedings of the 13th European Signal Processing Conference (EUSIPCO), Antalya, Turkey.

Yan Han, Johan de Veth and Lou Boves. 2005. Trajectory clustering for automatic speech recognition. In: Proceedings of the 13th European Signal Processing Conference (EUSIPCO), Antalya, Turkey.

Julie Heiser, Doantam Phan, Maneesh Agrawala, Barbara Tversky and Pat Hanrahan. 2004. Identification and validation of cognitive design principles for automated generation of assembly instructions. In: Proceedings of the Working Conference on Advanced Visual Interfaces, Gallipoli, Italy, pp. 311-319.

Gerd Herzog, Heinz Kirchmann, Stefan Merten, Allasane Ndiaye, and Peter Poller. 2003. MULTI-PLATFORM testbed: an integration platform for multimodal dialog systems. In: Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS), Edmonton, Canada, pp. 75-82.

Charlotte van Hooijdonk and Emiel Krahmer. 2006. De invloed van unimodale en multimodale instructies op de effectiviteit van RSI-preventieoefeningen. To appear in Tijdschrift voor Taalbeheersing, Sept. 2006.

Charlotte van Hooijdonk and Emiel Krahmer. submitted. Information modalities for procedural instructions: the influence of text, static and dynamic visuals on learning and executing RSI exercises.

Masashi Inoue. 2004. On the need for annotation-based image retrieval. In: Proceedings of the Workshop on Information Retrieval in Context (IRiX), Sheffield, UK, pp. 44-46.

Tsuneaki Kato, Fumito Masui, Jun'ichi Fukumoto and Noriko Kando. 2006. WoZ simulation of interactive question answering. In: Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006, New York City, NY, USA, pp. 9-16.

Simon Keizer and Harry Bunt. 2006. Multidimensional dialogue management. In: Proceedings 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia, pp. 37-45.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to Latent Semantic Analysis. Discourse Processes 25:259-284.

Mieke van Langen. 2005. Question-Answering for General Practitioners. MSc. thesis, University of Twente.

Piroska Lendvai. 2005. Conceptual taxonomy identification in medical documents. In: Proceedings of the 2nd International Workshop on Knowledge Discovery and Ontologies (KDO-2005), held within ECML/PKDD, Porto, Portugal, pp. 31-38.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. Text 8, pp. 243-281.

Emily E. Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. Journal of Documentation 59(6):647-672.

Erwin Marsi, Emiel Krahmer, Wauter Bosma and Mariët Theune. 2006. Normalized alignment of dependency trees for detecting textual entailment. In: Proceedings of the 2nd PASCAL Recognizing Textual Entailment Challenge, Venice, Italy.

Erwin Marsi and Emiel Krahmer. 2005. Explorations in sentence fusion. Proceedings of the 10th European Workshop on Natural Language Generation, Aberdeen, Scotland, pp. 109-117.

Richard E. Mayer and Roxana Moreno. 2002. Animation as aid to multimedia learning. Educational Psychology Review, 14(1):87-99.

Boris van Schooten and Rieks op den Akker. 2006. Follow-up utterances in QA dialogue. Submitted to TAL (Traitement Automatique des Langues/Natural Language Processing), Special Issue on QA.

Boris van Schooten. 2006. Multimodal follow-up questions in QA dialogue. In these proceedings.

Erik Tjong Kim Sang, Gosse Bouma and Maarten de Rijke. 2005. Developing offline strategies for answering medical questions. In Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains, Pittsburgh, PA, USA, pp. 41-45.

John Williams, Andrew Hickl, John Lehmann and Sanda Harabagiu. 2004. Experiments with interactive question answering in complex scenarios. In Proceedings of the HLT-NAACL2004 Workshop on Pragmatics of Question Answering, Boston, MA, USA, pp. 60–69.

Steven Winikoff and Leila Kosseim. 2004. Is context actually helpful? Preliminary experiments in contextual question answering. In: Proceedings of the Workshop on Computational Linguistics in the North-East (CLiNE 2004), pp. 64–66.

Michelle X. Zhou, Zhen Wen and Vikram Aggarwal. 2005. A graph-matching approach to dynamic media allocation in intelligent multimedia interfaces. Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI 2005), pp. 114-121.