International Journal of Semantic Computing
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

THE ROLE OF GESTURE IN DOCUMENT EXPLANATION BY EMBODIED CONVERSATIONAL AGENTS

TIMOTHY BICKMORE

College of Computer and Information Science, Northeastern University, 360 Huntington Ave, WVH202
Boston, Massachusetts 02115, USA
bickmore@ccs.neu.edu
http://www.ccs.neu.edu/home/bickmore

LAURA PFEIFER

College of Computer and Information Science, Northeastern University, 360 Huntington Ave, WVH202
Boston, Massachusetts 02115, USA
laurap@ccs.neu.edu

LANGXUAN YIN

College of Computer and Information Science, Northeastern University, 360 Huntington Ave, WVH202
Boston, Massachusetts 02115, USA
yinlx@ccs.neu.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

We describe two empirical studies of how professionals explain documents to lay clients who have varying levels of knowledge about the domain under discussion. We find that hand gestures, and in particular deictic gestures by the professional at various parts of the document play a major role in explanations of documents with clients in face-to-face settings. We describe a preliminary computational model of document explanation by an embodied conversational agent, in which appropriate form and location of hand gestures are used by the agent in explaining a document to a user. Results from a pilot evaluation study indicate that individuals with low levels of domain knowledge prefer receiving explanations from such an agent rather than from a human. Examples are drawn from the healthcare domain, in which research consent forms and hospital discharge instruction forms are used as the documents being explained, and health literacy is used as the measure of client domain knowledge.

Keywords: Embodied conversational agent; document explanation; health literacy.

1. Introduction

When people describe objects to each other in face-to-face conversations, they not only use their voice, but very frequently use their hands to convey what they mean. They may point at the object in various ways, pick the object up and hold or manipulate it, or hold the object in one hand while using their other hand to depict complex shapes in space. We have been interested in studying these phenomena within the context of document explanation, in which a professional explains a paper document to a lay client in a face-

to-face interaction. Although there are many paralinguistic phenomena to study in such interactions, and the hands may not be appropriate or sufficient for conveying many kinds of information [26], we chose to study hand gestures because of their prominence and ubiquity in the many examples of document explanation we have observed.

Professionals often provide their clients with documents that are, to varying degrees, incomprehensible. Whether due to technical jargon, obscure concepts, or poor writing on the part of the professional, or low literacy, cultural barriers, or cognitive impairment on the part of the client, documents often fail to serve their intended communicative function. This may be why hand gestures are necessary to help explain the structure of these documents and the complex and unfamiliar concepts they contain.

1.1. Example Domain: Healthcare

Perhaps nowhere is the problem of clients failing to understand documents more important and pervasive than in healthcare. The consequences of a patient failing to understand prescriptions, hospital discharge instructions, or pre-surgery instructions can have serious, even fatal, consequences. The inappropriate complexity of documents has been discussed in the medical literature for over 50 years and in the past two decades this has been broadly recognized as a serious problem within the US medical community [4]. Indeed, a significant and growing body of research has emerged relating to the problem of "health literacy" (the ability to perform the basic reading and numerical tasks required to function in the health care environment [1]), which has brought attention to the ethical and health impact of overly complex documents in healthcare [5,35].

1.2. Example Domain: Law

The legal domain is another area where the explanation and understanding of documents is important. Legal documents are full of technical language and jargon that is difficult for the average person to understand [27]. As a result, people will often sign documents, agreeing to terms and conditions that they do not fully comprehend [45]. Informed consent documents are one such example. Institutional Review Boards are responsible for approving research only when adequate standards of informed consent are satisfied [17]. However, a longstanding problem with informed consent documents is that they are often written at a literacy level higher than that of the subjects who participate in the studies that consent documents are written for [34] [36]. As a result, subjects can have trouble understanding the voluntary aspects of the research, as well as the potential benefits and risks [42].

1.3. Embodied Conversational Agents for Document Explanation

Face-to-face explanation of a document by a professional remains a client's best means of understanding its contents. Within healthcare, evidence suggests that face-to-face encounters with a health professional—in conjunction with written instructions—remains one of the best methods for communicating health information to patients in general, but

especially those with low literacy levels [30]. Face-to-face consultation is effective because it requires that the provider focus on the most salient information to be conveyed [38] and that the information be delivered in a simple, conversational speaking style. Protocols for grounding in face-to-face conversation—the use of verbal and nonverbal cues such as head-nods, gaze and acknowledgement tokens ("uh-huh", "OK") to signal mutual understanding [14]—allows providers to dynamically assess a patient's level of understanding and repeat or elaborate information as necessary. Face-to-face conversation also allows providers to make their communication more explicitly interactive by asking patients to do, write, say, or show something that demonstrates their understanding [18].

Of course, one problem with in-person encounters with professionals is that their time available to explain documents to clients is typically scarce and expensive. Within healthcare, all professionals function in environments in which they can only spend a very limited amount of time with each patient [15]. Time pressures can result in patients feeling too intimidated to ask questions, or to ask that information be repeated.

Given the efficacy of face-to-face consultation, one technology that shows particular promise for describing professional documents to clients with limited domain knowledge is the use of embodied conversational agents that simulate face-to-face conversation with a professional. These systems can recognize and produce verbal and nonverbal conversational behaviors that signify understanding and mark significance, and convey information in redundant channels of information, to maximize message comprehension [12]. They can use the verbal and nonverbal communicative behaviors used by providers to establish trust and rapport with their patients in order to increase satisfaction and, in health care, adherence to treatment regimens [6]. They can adapt their messages to the particular needs of clients and to the immediate context of the conversation. Embodied conversational agents can provide information in a consistent manner and in a low-pressure environment in which clients are free to take as much time as they need to thoroughly understand it.

1.4. Overview of Paper

In this paper we describe our work towards the development of an embodied conversational agent that can explain complex professional documents to users, focusing on the use of hand gestures by the agent. We first review related work on the development of embodied conversational agents, intelligent tutoring systems and text generation of extended descriptions. We then describe two empirical studies we conducted to investigate the role of hand gesture in the explanation of documents by professionals to their lay clients, and present our preliminary work in developing an embodied conversational agent that can perform this function using a model of hand gesture derived from the empirical studies. Finally, we describe an evaluation study in which we assess the acceptance and efficacy of this agent, compared to document explanation by a human.

2. Related Work

Document explanation by conversational agents is related to a number of other technologies developed in recent years, such as intelligent tutoring systems. In this section we briefly review this work, but first review background research into the use of hand gesture by people in face-to-face conversation.

2.1. Hand Gesture by Humans during Document Explanation

Hand gestures can be used to perform a wide variety of functions within the complex behavioral milieu of face-to-face conversation. McNeill's seminal work on gesture defines the following typology [28]: **deictic gestures** are pointing gestures, either to an object in the shared physical space of the interlocutors, or to a conceptual entity the interlocutors are discussing; **iconic gestures** depict aspects of a real physical object; **metaphoric gestures** depict aspects of an abstract entity; **beat gestures** do not represent anything, but simply mark emphasis in what is being said; and **emblematic gestures** have standards of form within a speech community and can stand in for words. All gestures except beats are triphasic in that they involve a preparation phase (moving into location), a hold phase, and a retraction phase. Beats are biphasic, in that they simply involve a stroke and a retraction, and can be overlaid (co-articulated) on top of other gestures.

Marslen-Wilson et al., characterized the deictic gestures of a speaker who is explaining a document to a listener [25]. In this study, the speaker is re-telling the story of a comic book to another person. The speaker does not flip through the book to describe things frame-by-frame, but rather tells the story while simply holding the book in their lap. The cover of the book contained pictures of the two main characters and deictics were frequently made to these pictures throughout the duration of the story. A particularly interesting finding was that a deictic to the cover of the book was used 100% of the time whenever a lead character received first mention within a given episode (although the sample size was extremely small). By doing this, it is thought that the speaker is establishing a stronger referent to entities (i.e., the main characters) that are important to comprehending the story.

2.2. Embodied Conversational Agents

Embodied conversational agents are animated humanoid computer characters that simulate face-to-face conversation with users [12].

Deictic gestures represent perhaps the most common type of hand gesture implemented in embodied conversational agents. Jack, the virtual meteorologist agent, was one of the earliest, and could point at weather images that he stood in front of (in his virtual environment) while giving a weather report [33]. However, the interaction and gesture specifications were entirely scripted. The BEAT system incorporated a simple rule that generated deictics whenever a new object referenced in speech was "visible" to both the agent and user in the agent's virtual world [13]. The Cosmo agent used a

separate deictic planner that would determine the generation of deictics on the basis of speech act, gesture referent, speech referent, world model (including possible distractors) and discourse history [13]. Krandstedt and Wachsmuth developed a model for generating deictics in conjunction with definite descriptions in speech to refer to objects, which uses the concept of a pointing cone to decide between pointing at specific objects vs. pointing at a region of objects [23].

Most embodied conversational agents are screen-based, but several have been implemented using alternative modalities in which deictics play a special role. MACK could highlight a paper map that was "shared" with a user by means of an overhead projector [11]. Steve accompanied users into a virtual world where he could point out virtual objects that the user needed to manipulate [40], as do the agents of Krandstedt and Wachsmuth [23].

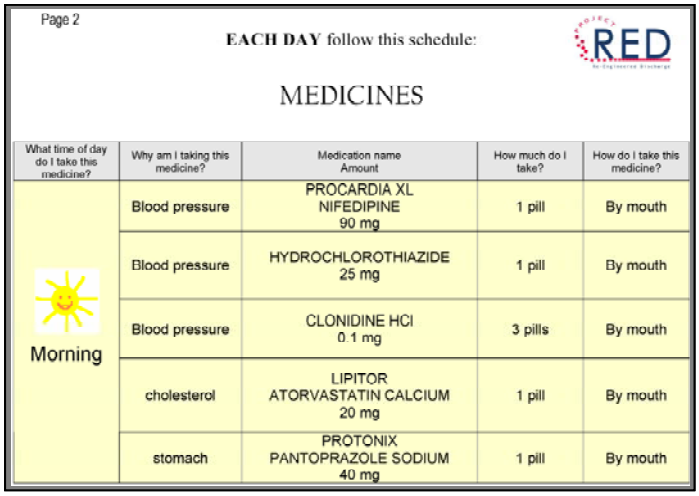
2.3. Embodied Pedagogical Agents

Document explanation can be seen as a kind of pedagogy, especially for users who are unfamiliar with the domain. Embodied agents have been used in a number of intelligent tutoring systems including Autotutor [19], Steve [40], Cosmo [24], Persona [3], Sam [9] and others. Most evaluations of these agents have shown weak instructional outcomes, but a few have shown promise. For example, in a series of studies involving the Cosmo agent, researchers found that students who interacted with an educational software system with a pedagogical agent produced more correct solutions and rated their motivation to continue learning and interest in the material significantly higher, compared to the same system without the agent [29]. In another study, students using the AutoTutor pedagogical agent in addition to their normal coursework outperformed both a control group (no additional intervention), and a group directed to re-read relevant material from their textbooks [37].

3. Empirical Studies of Document Explanation by Humans

We conducted two empirical studies to characterize how human experts explain documents to their clients in face-to-face interactions. The first study was conducted with human experts explaining two kinds of documents to a research confederate. The second study was conducted with one expert explaining one particular document to laypersons with different levels of knowledge about the domain. In both studies our primary focus is on the nonverbal behavior exhibited by the expert in order to inform the development of a computational model of document explanation. However, we also analyzed the verbal description strategies used, grounding behavior by both the expert and the client, and how all of this behavior changed with the client's knowledge level.

Two documents were created and used as stimuli in the empirical studies and subsequent evaluations. The first was an "After Hospital Care Plan" (AHCP) document, which is eleven pages long and consists of a mixture of text and images (Fig. 1 shows a sample page). The AHCP is designed to be given to patients before they are discharged from a hospital, and covers their diagnoses, medications, follow-up appointments, and



Page 2

EACH DAY follow this schedule:

MEDICINES

What time of day do I take this medicine?	Why am I taking this medicine?	Medication name Amount	How much do I take?	How do I take this medicine?
Morning	Blood pressure	PROCARDIA XL NIFEDIPINE 90 mg	1 pill	By mouth
	Blood pressure	HYDROCHLOROTHIAZIDE 25 mg	1 pill	By mouth
	Blood pressure	CLONIDINE HCl 0.1 mg	3 pills	By mouth
	cholesterol	LIPITOR ATORVASTATIN CALCIUM 20 mg	1 pill	By mouth
	stomach	PROTONIX PANTOPRAZOLE SODIUM 40 mg	1 pill	By mouth

Fig. 1. Sample AHCP Page: First page of medications

self-care procedures. While the AHCP is explicitly designed for patients with low health literacy, it is full of medical terminology, such as medication names and medical condition names. The second document was a research informed consent document (CONSENT), which was two pages long and consisted entirely of text, mostly in non-technical language. To minimize any carryover effects from the informed consent used for participation in our document explanation studies, the CONSENT document was from an entirely different area of medical research: acquisition of blood samples for genetic banking.

Conversations in both studies were videotaped using a synchronized 3-camera closed-circuit recording system. The videos were transcribed and broken into utterances, following [32], and speech acts were coded for each utterance using the DAMSL coding scheme [2]. Expert hand gestures, gaze (at the document, at the client, or elsewhere), and headnods, and client gaze and headnods were coded using ANVIL [21].

3.1. Empirical Study 1 – Expert/Confederate Role-Playing

The purpose of our first study was to gain an initial understanding of the nature of the document explanation process, and to characterize the frequency and form of nonverbal behavior used by an expert in explaining documents to their clients. We analyzed four example interactions in which four different experts explained documents to research confederates. Two of these conversations used the AHCP, and the experts were nurses who routinely explained AHCPs to patients. The other two conversations used the CONSENT document, and the experts were research assistants who routinely consented research study participants. All four interactions were "mock" conversations in that the listener was another research assistant. In the two AHCP examples, the nurse and "patient" are seated next to each other at a table with the document on the table between

Fig 2. Explanation of AHCP (left) and CONSENT (right) by experts.

them (Fig. 2). In the CONSENT examples, the research assistant and "client" are seated facing each other, and the research assistant holds the document up for the client.

Table 1 provides an overview of the four conversations.

Table 1. Empirical Study 1 Conversations Analyzed

			Utterances		
Conversation	**Document**	**Duration**	**Expert**	**Client**	**Total**
1-1	CONSENT	2:08	93	1	94
1-2	CONSENT	2:24	103	8	111
1-3	AHCP	6:46	282	32	314
1-4	AHCP	6:53	277	39	316

3.1.1. *Findings*

In all cases, the experts proceeded linearly through the documents from beginning to end, using the document structure to guide their explanation. As indicated in Table 1, the experts were responsible for the vast majority of utterances in each interaction. .

Of the 755 expert utterances in the four conversations, the expert was gesturing during 190 (25.2%) of them, and 98% of these utterances included a deictic gesture referencing the document. Deictic forms observed included: pointing at an image (4%); pointing at a word or phrase with the index finger (22%); pointing at a region or page (24%); pointing at something indeterminate (1%); underlining a word or phrase (18%); waving the whole hand over a region or page (6%); and whole hand touching on a region or entire page (25%). However, the distribution of gesture forms appeared to be fairly idiosyncratic, with one expert using whole hand gestures for 77% of his deictics, while the other experts only used them between 19% and 44% of their deictics.

The timing of gesture stroke relative to utterance was also coded as: before utterance, beginning of utterance (1st three words, following [31]), ending of utterance (last 3 words), middle of utterance, or continued from previous utterance. We found that 83% of the time, deictic gesture stroke occurred at the beginning of an utterance. We also found

that the expert would maintain his or her deictic through subsequent utterances that referred to the same document object 60% of the time.

While speaking but not gesturing, we found that the expert gazed at the document 65% of the time and at the client 30% of the time. However, when speaking and gesturing, the expert gazed at the document 83% of the time. This difference may be due to the expert's need to look at what he or she was pointing at, but may also be an additional form of deictic to draw the client's attention to the document (as observed in [25]).

3.1.2. *Modeling of Expert Hand Gesture*

We focused our initial modeling efforts on the occasioning and form of expert deictic gestures. Our goal was to develop a descriptive model that would predict when a deictic towards the document is typically used, and the form of this deictic, for example pointing at a specific word or image, or waving a hand over a page. To simplify our initial model, we collapsed the seven deictic categories above into POINT (pointing and underlining) and REGION (whole hand) gestures, to differentiate the specification of specific points vs. general regions on a page.

Following previous studies on reference and gesture occurrence, utterances that contain initial mentions of document items are more likely to be accompanied by gesture than utterances that either do not mention document items or only contain subsequent references [12,25]. Analyses indicated that not only was a new mention of a document object predictive of a gesture (43% of first mentions received deictics, compared to only 19% for subsequent mentions), but that the hierarchical part of the document referred to (entire document, page, section, etc.) seemed to be predictive of the form of the gesture used. For example, pointing with the index finger seemed to be used more frequently to refer to document items while waving over a page with a flat hand seemed to be used more frequently to refer to an entire page. Consequently, we coded the part of the document under discussion by the expert. The documents were broken up into topic level by identifying pages, regions and items within each document. Each topic was represented by an ID number in the format "<*page*>.<*section*>.<*item*>", eg. "1.4.2". We also created a code to indicate the topic level being introduced (PAGE, SECTION, or ITEM), as well as a code that indicated relative navigation in the document (IN, OUT, FORWARD, etc.), both based on changes in the topic ID.

Chi-squared tests for independence indicated that speech act, topic level, and document navigation were all strongly associated with the occurrence and form of deictic gesture performed during a given utterance (NONE, POINT or REGION, $p<.001$). We then used a commercial decision tree modeling tool (DTREG.com) to evaluate models based on various combinations of these coded predictors. The lowest error rate found (15.5%) was for a model that considered all available information (speech act, topic level, etc.). However, the model based on topic level alone was only slightly worse (15.6% error rate), so we based our initial computational model on topic level alone to simplify implementation.

Our first model predicts a deictic gesture according to the model in Table 2. However, this is based on combined statistics for four experts. In the next study we describe a model that is specific to one expert across three conversations, and may be more representative of what an individual expert may do.

Table 2. Document Deictic Generation Model for Conversations 1-1 through 1-4

	Gesture		
New Topic Level	**NONE**	**POINT**	**REGION**
No Change	92.8%	4.4%	2.8%
PAGE	57.7%	3.8%	38.5%
SECTION	23.7%	36.8%	39.5%
ITEM	23.7%	21.1%	55.3%

3.2. Empirical Study 2 – Variation of Client Knowledge Levels

In our second study, we analyzed three example interactions in which the same expert explained the same document to laypersons of differing domain knowledge levels. One goal of this study was to analyze more realistic interactions by using laypersons as clients, rather than knowledgeable research assistants. The second goal was to see if there were differences in the expert's verbal and nonverbal behavior for clients with different domain knowledge levels. We also conducted a much more detailed analysis of the function of expert hand gestures in these interactions.

In these studies, we used the standard AHCP document and a discharge nurse who routinely explains AHCPs to patients as our expert. A standard measure of health literacy was used to assess client domain knowledge. The REALM health literacy assessment categorizes individuals into 3rd grade and below, 4th-6th grade, 7th-8th grade, and high school [16]. We also created a comprehension test for the AHCP that was administered after the explanation session in an "open book" format, in which the client could refer to a copy of the AHCP while being asked questions by the research assistant. Study participant demographics, health literacy, and post-explanation test scores are shown in Table 3.

Table 3. Empirical Study 2 - Participant Characteristics

Conversation	**Age**	**Sex**	**REALM Literacy Category**	**AHCP Test Score**
2-1	71	F	2 (4th-6th Grade)	29%
2-2	61	F	3 (7th-8th Grade)	71%
2-3	25	M	4 (High School or Above)	100%

3.2.1. *Findings*

The durations of these interactions are significantly longer than those in conversations 1-3 and 1-4 from study 1, even though the same document was being described. Because the initial study consisted of mock interactions in which the "client" was knowledgeable

about the subject matter and was familiar with the AHCP, the expert glossed over many details in the document. In this second study, by comparison, all of the information in the

Table 4. Empirical Study 2 - Conversation Characteristics

		Utterances		
Conversation	**Duration**	**Expert**	**Client**	**Total**
2-1	9:30	391	37	428
2-2	10:30	396	101	497
2-3	12:05	452	82	534

document (as well as the appearance and structure of the document itself) was new to the clients. For example, in conversation 1-3, only the first of 19 medications were discussed, whereas every medication is discussed in detail in conversations 2-1 through 2-3.

3.2.2. *Findings on Expert Hand Gesture*

We observed a wide range of hand gestures used by the expert across the three conversations. As in our previous study, the vast majority of the 321 gestures observed were deictics at the document (85%), with the remainder comprised of beat gestures (5%), iconic gestures (4%, e.g., demonstrating how to take a nitroglycerine pill), metaphoric gestures (1%), emblematic gestures (2%, all holding some number of fingers up, e.g., "two pills") and deictics at things other than the document (2%, mostly at self and client). Document deictics were further sub-categorized as pointing at a word or phrase (64%), pointing at a region or page of the document (9%), waving the hand over the document (9%), underlining a word or phrase (5%), touching a region or page (2%), pointing at an image (1%), or indeterminate (10%).

Deictics at the document were used for a wide range of functions. The most common function was to simply reference a particular item in the document by pointing at it (Fig. 3). However, many other forms of reference were observed. The entire document was referenced in several ways, including holding the document up with both hands (Fig. 4), waving a hand in front of the document, and holding the document up with both hands and shaking it up and down. A page was often introduced by waving a hand over the page, pointing at the middle of the page, or holding the page up with both hands. Page regions or sections were introduced in a similar manner.

Fig. 3. Pointing reference to text block "…and that's hydrochlorithiazide…"

Fig. 4. Introducing the document "I'd like to go over this document with you.."

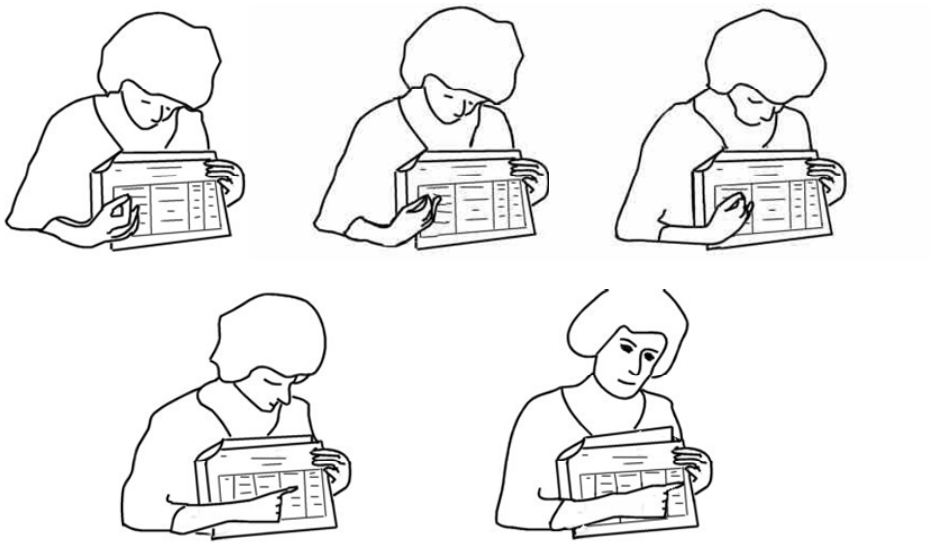


Fig. 5. Sequence of gestures describing table column headers
"…it tells you what the medication is for, the name of the medication, how many you take, and then how you take it."

Although other researchers have posited that hand shapes used in deictics carry additional meaning beyond the reference [20,44], we found that the hand shapes used in our interactions seem to be more a function of the size of the referent (partially determining whether a point, underline or whole hand is used) and the physical constraints the expert is under. For example, Fig. 5 shows a sequence of pointing gestures as the expert presents the cells in a table row, left-to-right. She starts with her index finger for the first cell, switches to her little finger for the next two cells, then completes the final two cells with her index finger again. We also observed her using the bent knuckle of her index finger or her thumb to point, in other contexts.

Deictics were also seen to be used as conversational place-holders. In one sequence, the expert introduces a medication by pointing at it, then keeps her finger on the medication name during a 14-utterance side-sequence regarding whether the client is currently taking that medication or not and their current dosage.

We observed a wide variety of beat gestures, many of them co-articulated with deictics. Beats were usually performed with the hand not holding the document, and typically while not pointing at anything. However, there were several occasions when the pointing hand would be moved in a bi-phasic beat gesture while pointing at a referent, occasions when both hands would move up and down (with the document) to effect a beat, and occasions when the expert would switch holding hands just to enable her to beat with the just-freed hand.

Perhaps the most interesting gestures were deictics whose referents were within tables in the document. Tables were introduced to the client in several ways. In conversation 2-1, the expert introduces the table by first describing the columns in the table, pointing to

Fig. 6. Reference to a table section by pointing down the first column "This page is a list of your appointments."

Fig. 7. Reference to a page using flat hand wave in front of document "…everything in yellow you're going to take in the morning."

each header cell, left-to-right, in turn (Fig. 5). In several instances, part of a table comprising several adjacent rows was introduced by pointing at the cells in the left-most column, top-down, as the section is introduced (Fig. 6). In conversation 2-3, an overview of a multi-page table is provided by repeating this section introduction behavior for each section of the table, turning the page as needed. Table sections were also introduced by hand wave, as shown in Fig. 7. On several occasions, the expert oriented the user to a particular row in the table, either by pointing at the left-most cell, or the cell with the primary identifying information for the row, then describe the row contents without further gesture, or by pointing at the relevant column header for each cell introduced.

The model for predicting gesture based on topic level change presented in Section 3.1.2 was updated for conversations 2-1 through 2-3, as shown in Table 5.

Table 5. Document Deictic Generation Model for Conversations 2-1 through 2-3

	Gesture		
New Topic Level	**NONE**	**POINT**	**REGION**
No Change	80.8%	13.1%	6.1%
PAGE	63.6%	13.6%	22.7%
SECTION	48.3%	32.8%	19.0%
ITEM	31.2%	65.9%	2.9%

3.2.3. *Comparative Analysis of Explanations for Clients with Low vs. High Knowledge*

To understand differences in an expert's behavior when explaining the same document to clients with different knowledge levels, we performed a detailed comparison of conversation 2-1 (client at 4^{th}-6^{th} grade literacy level) with conversation 2-3 (client at high school literacy level).

The first interesting difference is in the length of the explanations. The expert took significantly longer to explain the document to the high literacy participant, both in terms of time and number of expert utterances (Table 4). Discourse analysis showed that the

expert omitted many details in her presentation to the low literacy client while including more details and elaborating on more related information for the high literacy client. For example, of the 28 items on the first AHCP page of medications (Fig. 1), the expert explicitly mentioned 25 of these in her explanation to the high literacy client, but only 14 items in her explanation to the low literacy client.

On the other hand, the expert did provide more instructional scaffolding for the lower literacy client, such as overviews of the structure of the document before giving details, summaries of parts of the document just presented, and pointing out different kinds of similarities. For example, on the first AHCP page of medications, for the low literacy client, the expert first gives an overview of the medications table by describing the column headers before describing the first medication (Fig. 5), and also gives a summary of the medications on the first page after they are all presented, whereas this preview and summary is not provided for the high literacy client.

We also investigated how the expert might surreptitiously assess the client's literacy level during their conversation, since the expert did not have any prior knowledge of who the client was or what their REALM scores were. One obvious candidate is client grounding behavior, used to indicate their understanding of the information given by the expert. Grounding can be exhibited in a number of ways [31], but we looked primarily at acknowledgment tokens, consisting of client 'ack' speech acts (e.g., "OK") and client head nods while the expert is speaking. In conversation 2-1 (low literacy), the client only exhibited acknowledgments for 10% of the expert's utterances, while in conversation 2-3 (high literacy), the client exhibited acknowledgments for 49% of the expert's utterances.

3.3. Conclusions from Empirical Studies

In addition to the primary purpose of describing the contents of the document to the client, the expert usually has secondary goals as well. These may include: teaching the concepts and facts contained therein to the client as necessary; teaching the client the overall structure of the document and the location of specific facts, so that the client will be able to find these facts later; teaching the client the structure of specific document elements so that the client can interpret structure-specific meaning (e.g., information laid out in tables); and perhaps other information related to the document, such as how to fill out forms contained in the document or how to get additional interpretive help for the document. Thus, when an item in a document is described, it is usually for several purposes. In addition to the primary goal of describing (cataloguing) the item, the expert typically also wants to convey the spatial location of the item in the document (for later retrieval by the client) and the role of the item in the document structure (for later interpretation by the client). Deictic gestures support all of these goals by providing a (typically) unambiguous reference to the item and its location in the document.

There are a number of description strategies that an expert can use in describing a document, analogous to the metastrategies used in Sibun's work on generating description texts for apartments [41]. The default is page-by-page, then top-down, left-to-right within each page. However, particular document elements, such as tables that span

multiple pages, allow other strategies, such as providing a preview of the table structure across the pages, or pointing out table elements that re-occur on multiple pages.

If the expert and client mutually believe that the client knows the current description strategy being used, then the expert can incorporate the client's knowledge of the next item in sequence into the generation of his or her reference to the next item. For example, we observed cases in which the expert points at a text block and then reads the items in the block in order without further gesture (speech and sequence provide complete specification). We also observed cases in which the expert points at the beginning of a table row while explaining each of the cells in that row left-to-right in order (speech, partial spatial information in gesture, and sequence provide complete specification, Fig. 8).

Figure 8. Referencing table cell by pointing at column, row implied from sequence. "…the next one is nicotine…"

The decision by the expert to use a deictic gesture when mentioning an item in the document, then, is likely based on several factors, including: the ability of speech to uniquely identify the item in a concise manner, given distractors on the page and the prominence of the item on the page; the expert's perception of the client's ability to infer the next item in sequence according to the current description strategy in use; the degree to which the client's locating the item on the page satisfies the expert's multiple communicative goals; and the importance of the item in the expert's pedagogical strategy (we observed that less important information was less likely to be pointed at). However, since the expert is in a teaching role, he or she will likely err on the side of over-specifying when referring to document items, and thus will often point at an item even when speech and/or sequence uniquely identify an item on the page. We believe this is why we observed that the expert was using a document deictic gesture during 43% of her utterances.

4. Towards a Computational Model of Document Explanation

Our ultimate goal in this research is to develop an embodied conversational agent that can explain a document to a client as well as a human expert, given the document and a semantic representation of the document's contents. In this section we describe an

existing agent framework that we are building upon and our preliminary and planned work in extending this framework for document explanation.

4.1. Embodied Conversational Agent Framework

We are building upon an existing embodied conversational agent framework that was designed for health counseling applications [8]. The framework features a vector-graphics-based animated agent whose nonverbal behavior is synchronized with a text-to-speech engine (Fig. 9). User contributions to the conversation are made via a touch screen selection from a multiple choice menu of utterance options, updated at each turn of the conversation. For health counseling applications, dialogues are scripted using a custom hierarchical transition network-based scripting language. In addition to network branching operations, script actions can include saving values to a database or retrieving and testing values from the database, in order to support the ability to remember things about users and to be able to refer back to prior conversations. The system currently uses template-based text generation for agent utterances [39], so that the utterances can be tailored at runtime based on information in the database or other sources.

The agent has a range of nonverbal behaviors that it can use for co-verbal communicative and interactional functions, including: hand gestures [28], body posture shifts [10], gazing at and away from the user [43], raising and lowering eyebrows, affective facial displays, head nods, and walking on and off the screen. It also supports three different facial expressions, variable proximity (wide to close-up camera shots) and several idle-time behaviors (subtle shifts or self-adaptors).

Co-verbal agent behavior is determined for each utterance using the BEAT text-to-embodied-speech translation system [13], with several enhancements to support health dialogues. BEAT takes the text of an utterance as input (optionally tagged with semantic and pragmatic markers) and produces an animation script as output that can be used to drive an embodied agent's production of the utterance, including not only speech and intonation, but accompanying nonverbal behavior, such as hand gestures, gaze behavior, and eyebrow raises. BEAT was developed to be extensible so that new conversational functions and behaviors could be easily added. While we are aware of some of the limitations of BEAT [22], we find that it is adequate for our purposes.

4.2. Extensions for Document Explanation

We extended the embodied conversational agent framework in several ways to accommodate the verbal and nonverbal aspects of document description.

We added a set of animation system commands to allow document pages to be displayed by the character (Fig. 9), with page changes automatically accompanied by a page-turning sound. We also added a set of document deictic gestures so that the agent could be commanded to point anywhere in the document with either an index finger or a flat hand. While the document is displayed, the agent can continue using its full range of head and facial behavior, with gaze-aways modified so that the agent looks at the document when not looking at the user (Section 3.1.1). However, hand gestures are

currently limited to document deictics, and posture and proxemic shifts are disabled while the document is displayed.

4.2.1. *Generation of Nonverbal Behavior for Document Descriptions*

Our current document description system is utterance-centric, in that the description utterances are first generated, after which appropriate accompanying nonverbal behavior (such as document deictics) is selected. Utterances sent to BEAT for processing are first annotated with information about a document object's physical and logical locations in the document. These tags specify the document location ID described in Section 3.1.2, and the X,Y coordinates (normalized to 100%, 100%) of the page corresponding to the location ID, for example:

```
<DOC PART="2.1.1" LOC="25,40"> It is for your blood pressure. </DOC>
```

We created a BEAT behavior generator that tracks document context (current and previous document locations under discussion) and annotates the utterance parse tree with: page change specifications (whenever the document location ID indicates a change in page); document deictic gestures (per the rules described in Section 3.1.2); and additional gaze-aways (at the start of all utterances in which a document deictic gesture or page change is indicated).

5. Preliminary Evaluation Study

We conducted a pilot evaluation study to test the acceptance and efficacy of an agent-based document explanation system, compared with a standard of care control (explanation by a human) and a non-intervention control (self study of the document in

Fig 9. Document Explanation Agent Interface

question) [7]. The study had a 3 (AGENT vs. HUMAN vs. SELF) x 2 (AHCP vs. CONSENT) between-subjects experimental design, in which each participant evaluated two different conditions in a single session, always AHCP followed by CONSENT, with the presentation of the other conditions randomized. This document ordering was intended to minimize carryover effects from the informed consent procedure for the pilot study itself to the CONSENT treatment of the study.

5.1. Apparatus

The evaluation study was conducted early in our development effort, and used the deictic and gaze models derived from our first study, described in Section 3.1. Two interaction scripts were created for the agent, one for the AHCP and one for CONSENT, based on the interactions described in Section 3.1. In each script, users could simply advance linearly through the explanation (by selecting "OK"), ask for any utterance to be repeated ("Could you repeat that please?"), request major sections of the explanation to be repeated, or request that the entire explanation be repeated. Any number of repeats could be requested and, although the scripting language has the ability to encode re-phrasings when an utterance is repeated, for the current study the agent would repeat the exact same utterance when a repeat was requested for any state in the script. The agent was deployed on a mobile cart with a touch screen display. Study sessions were held in an observation room of our HCI laboratory, with the interactions videotaped using three closed-circuit video cameras.

5.2. Measures

In addition to basic demographics, we assessed health literacy using the REALM instrument (described in Section 3.2 [16]). We also used the AHCP knowledge test developed for the second empirical study (Section 3.2) and developed a new test for CONSENT based on the BICEP evaluation.[42] These tests were always administered in an "open book" fashion with the participant able to refer to a paper copy of the document during the test. We augmented the BICEP with scale measures of likelihood to sign the consent document and perceived pressure to sign the consent document.

Evaluation questionnaires were also developed for the HUMAN and AGENT study conditions, assessing satisfaction with the instructor and with the overall instructional experience, desire to continue working with the instructor, trust in the instructor, and how knowledgeable the instructor was, all evaluated on 7-point scales.

5.3. Participants

Eighteen subjects participated in the study, were recruited via fliers posted around the Northeastern University campus, and were compensated for their time. Participants had to be 18 years of age or older and able to speak English. Participants were 74% male, aged 19-33. Two were categorized as 4^{th}-6^{th} grade, three as 7^{th}-8^{th} grade, and the rest as high school level, according to the REALM health literacy instrument.

5.4. Procedure

Participants arrived at the HCI laboratory, were consented, filled out the demographic questionnaire and then had the REALM health literacy evaluation administered.

Following this they were exposed to one of the three experimental conditions for the AHCP document. For the AGENT condition, they were given a brief training session on how to interact with the agent, the experimenter then gave the participant a paper copy of the document, left the room and closed the door. At the end of the interaction the virtual agent informed the participant that they could take as much time as they liked to review the document before signaling to the experimenter that they were ready to continue. For the HUMAN condition, a second research assistant in our lab explained the document to the study participant, after which the participant could review the document on their own, as in the AGENT condition. This instructor did not have a health care background, but routinely administered informed consent for HCI studies and was allowed to watch the videotapes described in Section 3 to learn about the AHCP. The instructor was blind to the virtual agent interaction script content and evaluation instruments, and was simply asked to explain the document in question to the participant. For the CONTROL condition, the participant was simply handed the document and told to take as much time as they needed to read and understand it, and were then left alone in the observation room until they signaled they were ready to continue.

Following the first intervention, the research assistant verbally administered the AHCP knowledge test and instructor evaluations. The previous two steps were then repeated with the CONSENT document.

5.5. Results

We conducted full-factorial ANOVAs for all measures, with condition (AGENT, HUMAN, SELF), document (AHCP, CONSENT) and health literacy (four categories) as independent factors, and LSD post-hoc tests when applicable.

There was one main effect of document on test score ($F(1,18)=14.5, p<.001$) indicating that participants scored significantly higher on the AHCP test compared to the CONSENT test. There were no significant effects of condition or literacy on test score.

Instructor evaluations for the AGENT and HUMAN conditions indicated a number of significant effects. There was a significant interaction between condition, document and literacy on satisfaction with the overall experience ($F(1,14)=5.0$, $p<.05$) such that those in the highest literacy level were more satisfied with the agent compared to the human for CONSENT, but were more satisfied with the human for AHCP. However, lower literacy participants were more satisfied with the agent in all situations.

All participants rated the agent as more knowledgeable than the human for CONSENT, but the human more knowledgeable for AHCP ($F(1,140)=6.0$, $p<.05$).

There were also several main effects for literacy, with the lowest literacy participants scoring significantly lower on trust in the instructor (whether human or agent, $F(2,14)=4.4$, $p<.05$), how knowledgeable the instructor was ($F(2,14)=3.8$, $p<.05$), and desire to continue working with the instructor ($F(2,14)=4.2$, $p<.05$). For the CONSENT

document, there were significant effects of health literacy on likelihood to sign ($F(2,12)=6.4$, $p<.05$) and perceived pressure to sign ($F(2,12)=132.0$, $p<.001$), such that those with lowest literacy were significantly less likely to sign and felt significantly more pressure to sign, compared to those with higher levels of literacy.

Six participants interacted with both the human and the agent in a single session, so we also compared ratings from these participants using a matched-pair analysis for increased power. These participants rated the agent significantly higher on satisfaction with the instructor (paired $t(5)=2.7$, $p<.05$) and satisfaction with the overall experience (paired $t(5)=2.9$, $p<.05$), compared to the human.

5.6. Evaluation Study Conclusions

Although we did not see significant differences in test scores across intervention conditions, this was not too surprising given the relatively high literacy status of the participants and the fact that the tests were "open book". However, in a busy clinic, especially with low literacy patients, the agent may actually outperform a time constrained and impatient clinician. As some participants put it:

- "I'd rather have Elizabeth. I liked the interface. I liked the way the tone has been set to explain to people. It doesn't kind of exert too much pressure on the person who's listening, so I like that."
- "Elizabeth was cool, I would have taken that again. She was just so clear, she just went page by page so it wasn't missed. And then, I mean you can always just ask them [human] if you don't understand anyway, but it's different on a screen, I guess, because some people don't want to say that they don't understand. On a screen it's less embarrassing, no one's here so you can say 'Ok, let me hear that again.'"

While we did not see effects on test scores, we did see clear patterns emerge on satisfaction, with direct comparisons by participants who interacted with both the agent and human, as well as all evaluations by low literacy participants, indicating a preference for the agent.

6. Conclusions and Future Work

The empirical study demonstrated that professionals use a wide range of hand gestures when explaining documents to their clients, and that deictics play an especially important role in referring to the document, providing spatial previews and summaries of structured document elements, such as tables, and as conversational place-holders. The agent evaluation study demonstrated that much of this behavior can be performed by an automated agent, that people find the explanations by such agents at least as satisfying as those provided by professionals, and that those with low domain knowledge may actually prefer the agent because it has infinite patience, will not criticize them because they don't understand something, and appears to be less biased.

Our future work is focused on establishing the ecological validity of our results and extending and refining our computational model. Evaluation in actual time-limited medical settings with participants across the full range of health literacy categories and

using documents containing information of real importance to the participants (e.g., their own hospital discharge instructions) would be important for understanding the generalizability of our evaluation study findings. Our evaluation study also did not examine the use of agent hand gesture in isolation, for example, comparing our agent to an equivalent agent with hand gesture disabled, and such an evaluation would be important to more fully understand the impact of agent hand gesture on user comprehension. We are also developing a text generation module that can generate document descriptions from first principles given a description of a document's contents and structure, incorporating a range of verbal and nonverbal behavior, with a focus on pedagogical behavior, such as instructional scaffolding, for clients with low levels of domain knowledge. We also plan to study document explanation in other domains to determine the generality of our models.

The immediate application for this research is a "virtual nurse" that will explain discharge instructions (After Hospital Care Plans) to patients before they leave the hospital. An embodied agent displayed on a touch screen computer will be wheeled into a patient's hospital room and positioned over their bed. The agent will spend approximately an hour reviewing the patient's AHCP with them, while the patient is able to follow along with a paper copy of the document. This intervention will be evaluated in a randomized clinical trial involving 750 patients at Boston Medical Center beginning in early 2008.

In sum, the document explanation is a very important and compelling application domain for embodied conversational agents, and health document explanation to patients with low health literacy is particularly important for addressing the health needs of certain underserved populations.

Acknowledgments

Thanks to Francisco Crespo and Thomas Brown for their assistance in conducting the evaluation study, to Mary Goodwin for being our expert in the second empirical study, and to our collaborators at Boston Medical Center—Dr. Brian Jack and Anna Johnson—for providing the example After Hospital Care Plan and videotaping the mock hospital discharge consultations. Dr. Michael Paasche-Orlow provided a great deal of information on the topic of health literacy and many suggestions on the evaluation study. Jennifer Smith provided many helpful comments on the paper as well as the video frame sketches. This work was supported by a grant from the NIH National Heart Lung and Blood Institute.

References

[1] American Medical Association Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs Health literacy: report of the Council on Scientific Affairs. JAMA, 281, 6, (1999) 552-557.

[2] J. Allen and M. Core, Draft of DMSL: Dialogue Act Markup in Several Layers, 1997.

[3] E. Andre, T. Rist, and J. Muller Integrating reactive and scripted behaviors in a life-like presentation agent. Proceedings of AGENTS'98, 1998) 261-268.
[4] H. K. Beecher Ethics and clinical research. N.Engl.J.Med., 274, 24, (1966) 1354-1360.
[5] N. Berkman, M. Pignone, S. Sheridan, and K. Lohr, Literacy and Health Outcomes. Evidence Report/Technology Assessment No. 87 University of North Carolina Evidence-based Practice Center, 2004.
[6] T. Bickmore, Relational Agents: Effecting Change through Human-Computer Relationships, Media Arts & Sciences, Massachusetts Institute of Technology, Cambridge, MA, 2003.
[7] T. Bickmore, Pfeifer, L., and Paasche-Orlow, M., Health Document Explanation by Virtual Agents, Intelligent Virtual Agents, Paris, 2007.
[8] T. Bickmore and R. Picard Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer Human Interaction, 12, 2, (2005) 293-327.
[9] J. Cassell, M. Ananny, A. Basu, T. Bickmore, P. Chong, D. Mellis, K. Ryokai, J. Smith, H. Vilhjálmsson, and H. Yan, Shared Reality: Physical Collaboration with a Virtual Peer, Proceedings of CHI '00, 2000.
[10] J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich, Non-Verbal Cues for Discourse Structure, Association for Computational Linguistics, 2001, pp. 106-115.
[11] J. Cassell, T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle, and H. Vilhjálmsson, MACK: Media lab Autonomous Conversational Kiosk, Imagina '02, 2002.
[12] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds., Embodied Conversational Agents, The MIT Press, Cambridge, MA, 2000.
[13] J. Cassell, H. Vilhjálmsson, and T. Bickmore, BEAT: The Behavior Expression Animation Toolkit, SIGGRAPH '01, 2001, pp. 477-486.
[14] H. H. Clark and S. E. Brennan, Grounding in Communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds., Perspectives on Socially Shared Cognition, American Psychological Association, Washington, 1991, pp. 127-149.
[15] F. Davidoff Time. Ann Intern Med, 127, 1997) 483-485.
[16] T. C. Davis, S. W. Long, R. H. Jackson, E. J. Mayeaux, R. B. George, P. W. Murphy, and M. A. Crouch Rapid estimate of adult literacy in medicine: a shortened screening instrument. Fam Med, 25, 6, (1993) 391-5.
[17] DHEW, The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, DHEW 1978.
[18] C. Doak, L. Doak, and J. Root, Teaching patients with low literacy skills, 2nd ed., JB Lippincott, Philadelphia, PA, 1996.
[19] A. Graesser and e. al AutoTutor: A simulation of a human tutor. Cognitive Systems Research, 1, 1999).
[20] A. Kendon, Gesture: Visible Action as Utterance, Cambridge University Press, Cambridge, 2004.
[21] M. Kipp, Anvil - A Generic Annotation Tool for Multimodal Dialogue, 7th European Conference on Speech Communication and Technology (Eurospeech), 2001, pp. 1367-1370.
[22] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson, Towards a Common Framework for Multimodal Generation: The Behavior Markup Language, Intelligent Virtual Agents, Marina Del Rey, CA, 2006.
[23] A. Kranstedt and I. Wachsmuth, Incremental Generation of Multimodal Deixis Referring to Objects, 10th European Workshop on Natural Language Generation (ENLG 2005), 2005, pp. 75-82.
[24] J. C. Lester, J. L. Voerman, S. G. Towns, and C. B. Callaway, Cosmo: A Life-like Animated Pedagogical Agent with Deictic Believability, IJCAI 97, 1997.

[25] W. Marslen-Wilson, E. Levy, and L. Tyler, Producing interpretable discourse: The establishment and maintenance of reference. In R. Jarvella and W. Klein, Eds., Speech, place, and action, Wiley & Sons, Chichester, England, 1982, pp. 339-378.
[26] C. Martell, FORM: An Experiment in the Annotation of the Kinematics of Gesture, Computer and Information Science, University of Pennsylvania, Pittsburgh, 2005.
[27] M. E. J. Masson and M. A. Waldron Comprehension of Legal Contracts by Non-Experts: Effectiveness of Plain Language Redrafting. Applied Cognitive Psychology, 8, 1, (1994) 67-85.
[28] D. McNeill, Hand and Mind: What Gestures Reveal about Thought, Cambridge University Press, Cambridge, 1992.
[29] R. Moreno, J. C. Lester, and R. E. Mayer, Life-Like Pedagogical Agents in Constructivist Multimedia Environments: Cognitive Consequences of their Interaction, ED-MEDIA 2000, pp. 741-746.
[30] L. Morris and J. Halperin Effects of Written Drug Information on Patient Knowledge and Compliance: A Literature Review. Am J Public Health, 69, 1, (1979) 47-52.
[31] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell, Towards a Model of Face-to-Face Grounding, Annual Meeting of the Association for Computational Linguistics, 2003.
[32] C. Nakatani and D. Traum, Coding discourse structure in dialogue (version 1.0), University of Maryland, 1999.
[33] T. Noma, L. Zhao, and N. I. Badler Design of a virtual human presenter. Computer Graphics and Applications, IEEE, 20, 4, (2000) 79-85.
[34] J. R. P. Ogloff and R. K. Otto Are Research Participants Truly Informed? Readability of Informed Consent Forms Used in Research. Ethics & Behavior, 1, 4, (1991) 239.
[35] M. Paasche-Orlow, S. M. Greene, and E. H. Wagner How health care systems can begin to address the challenge of limited literacy. J Gen Intern Med., 21, 8, (2006) 884-887.
[36] M. K. Paasche-Orlow, H. A. Taylor, and F. L. Brancati, Readability Standards for Informed-Consent Forms as Compared with Actual Readability, Vol. 348, 2003, 721-726.
[37] N. K. Person, A. C. Graesser, L. Bautista, and E. C. Mathews, Evaluating Student Learning Gains in Two Versions of AutoTutor. In J. D. Moore, C. L. Redfield, and W. L. Johnson, Eds., Artificial intelligence in education: AI-ED in the wired and wireless future, IOS Press, Amsterdam, 2001, pp. 286-293.
[38] C. Qualls, J. Harris, and W. Rogers, Cognitive-Linguistic Aging: Considerations for Home Health Care Environments. In W. Rogers and A. Fisk, Eds., Human Factors Interventions for the Health Care of Older Adults, Lawrence Erlbaum, Mahwah, NJ, 2002, pp. 47-67.
[39] E. Reiter and R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, 2000.
[40] J. Rickel and W. L. Johnson Animated Agents for Procedural Traning in Virtual Reality: Perception, Cognition and Motor Control. Applied Artificial Intelligence, 1998).
[41] P. Sibun Generating Text Without Trees. Computational Intelligence: Special Issue on Natural Language Generation, 8, 1, (1992) 102-122.
[42] J. Sugarman, P. W. Lavori, M. Boeger, C. Cain, R. Edson, V. Morrison, and S. S. Yeh Evaluating the quality of informed consent. Clinical Trials, 2, 1, (2005) 34.
[43] O. E. Torres, J. Cassell, and S. Prevost, Modeling Gaze Behavior as a Function of Discourse Structure, First International Workshop on Human-Computer Conversation, 1997.
[44] D. Wilkins, Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). In S. Kita, Ed., Pointing: Where Language, Culture and Cognition Meet, Lawrence Elrlbaum, Hillsdale, NJ, 2003, pp. 171-215.
[45] M. S. Wogalter, J. E. Howe, A. H. Sifuentes, and J. Luginbuhl On the adequacy of legal documents: factors that influence informed consent. Ergonomics, 42, 4, (1999) 593-613.