THE NEW REVIEW OF HYPERMEDIA AND MULTIMEDIA, Volume 4, 1998.
Special Issue on:
"Adaptivity and user modelling in hypermedia systems"
Peter Brusilovsky and Maria Milosavljevic (Eds)

The Dynamic Generation of Hypertext Presentations of Medical Guidelines

Berardina De Carolis and Fiorella de Rosis
Dipartimento di Informatica,
Università di Bari
nadja@aos2.uniba.it, derosis@di.uniba.it

Chiara Andreoli, Vincenzo Cavallo and M. Luisa De Cicco
Dipartimento di Medicina Sperimentale e Patologia,
Università "La Sapienza", Roma
cavallov@axrma.uniroma1.it

AbstractAbstract

We describe a project aimed at developing a tool to generate user and context-adapted textual descriptions of clinical guidelines on the World Wide Web. ARIANNA employs two knowledge sources (a decision tree and a taxonomy of concepts in the clinical domain) and schema and ATN-based NLG techniques, to dynamically generate the hypermedia. This appears to the user as a frameset with three main components: the guideline itself, the explanation of related concepts and the justification of individual steps. Each component is adapted to the user: i) the guideline is adapted to the user goal in consulting the system (tutoring vs decision support); ii) explanations of concepts are adapted to the user knowledge and to the interaction history; iii) justifications are reserved to the tutoring consultation mode.We describe a project aimed at developing a tool to generate user and context-adapted textual descriptions of clinical guidelines on the World Wide Web. ARIANNA employs two knowledge sources (a decision tree and a taxonomy of concepts in the clinical domain) and schema and ATN-based NLG techniques, to dynamically generate the hypermedia. This appears to the user as a frameset with three main components: the guideline itself, the explanation of related concepts and the justification of individual steps. Each component is adapted to the user: i) the guideline is adapted to the user goal in consulting the system (tutoring vs decision support); ii) explanations of concepts are adapted to the user knowledge and to the interaction history; iii) justifications are reserved to the tutoring consultation mode.

Keywords: dynamic generation of user-adapted hypermedia, clinical guidelines, concept explanation.

1. INTRODUCTION[1]

Clinical guidelines are "systematically developed statements to assist practitioner and patient decisions about appropriate health care in specific clinical circumstances" (12); their application fulfils different but confluent needs: (i) the patients' right to receive appropriate diagnostic procedures and treatments; (ii) a patient-directed and medically justified control of the health expenditure and (iii) the legal defence of physicians, who can demonstrate that they followed officially-established principles, in specific circumstances. To reach these goals, clinical guidelines must be made accessible to a broad population of users: the WWW is therefore the best candidate developing platform for such systems, whose potential users may be classified in the following groups:

- *students*, who may learn diagnostic and therapeutic procedures to follow in specific situations;
- *doctors* with several degrees of competence, who may apply correct diagnostic and therapeutic procedures;
- *patients*, who may get information about the scope and the efficacy of the health treatment they have to undergo.

Guidelines are not just a prospect of health care: in several clinical fields (such as cancer treatment) they are a broadly accepted reality; in Italy they recently became regulated by law. As they usually take the form of large documents, they are rigid and difficult to consult. In addition, to insure their introduction into health care, the medical staff needs to be convinced about the rationale behind suggestions and to understand all concepts involved: justification of suggestions and concept explanations are therefore essential constituents of any computer-based guideline consultation system. The clinical guidelines that may be found on the WWW are conceived as electronic versions of paper-based protocols; this approach is adequate for information but not for training and orientation. From these preliminary considerations, we may conclude that a computer-based clinical guidelines consultation system should include the following components: (i) the *guideline* itself; (ii) *explanations* of concepts mentioned in the guideline and (iii) *justifications* of individual steps. Each of these components should be adapted to the user: the guidelines should be adapted to their goal in consulting the system (tutoring vs decision support vs information) that is related, in its turn, to their category (student vs doctor or patient); explanations of concepts should be adapted to the user knowledge and to the interaction history; justifications should be reserved for the tutoring consultation mode.

We had a year's experience in building hypermedia guidelines (11): building a hyperdocument 'from scratch' required close cooperation of doctors and computer scientists; in addition, page building had to be repeated with different presentation forms as guidelines had to be adapted to the user characteristics; the risk was to end up this long work with a system in which inconsistencies in the page layout might occur. To overcome these limits, we began the new project that is the object of this paper: we defined a formalism to represent clinical knowledge behind guidelines and we designed a tool (ARIANNA) that translates this knowledge into dynamically generated and user-adapted hypermedia descriptions of guidelines. Two knowledge sources are employed in ARIANNA: a *decision tree,* that represents the decision process, and some *taxonomies of concepts* in the medical domain; schema and ATN-based NLG techniques are used to translate this knowledge into hypermedia.

The three components of a guideline that we mentioned above are seen by the users as parts of a frameset, as shown in **Figure 1**:

- a *scroll-shaped guideline* form (on the leftside of this figure) enables them to provide data about a particular clinical case they want to solve (in the example, these data are the direct/indirect bilirubin ratio and some relevant symptoms) and to suggests tests to perform, likely diagnoses and appropriate treatments. Anchors to explanations of concepts in the knowledge base are attached automatically to the generated sentences (again, in the example, a link to 'liver US scan'). The metaphor of 'scrolling' represents the form's ability to wrap round on itself to focus on the present step of the decision process, to show previous steps of the consultation when needed or to synthesise them when an intermediate

[1] **This Project was supported by the National Research Council, Committee on Information technology.**

conclusion is reached.[2] We preferred this presentation style to a sequence of individual pages because it reduces the risk of losing orientation in the hypermedia, which was one of the limits of our previous prototype;

- an *explanation frame* (rightside of the figure) enables displaying clarifications of concepts mentioned in the guideline when the user requests them; these explanations are adapted to the users' experience and to their knowledge of other concepts;
- a *justification frame* displays motivations of suggestions, to increase understanding of the rationale behind the guidelines; this frame is not visible in the figure because it comes at the same place as the explanation frame, the two frames never being shown together.

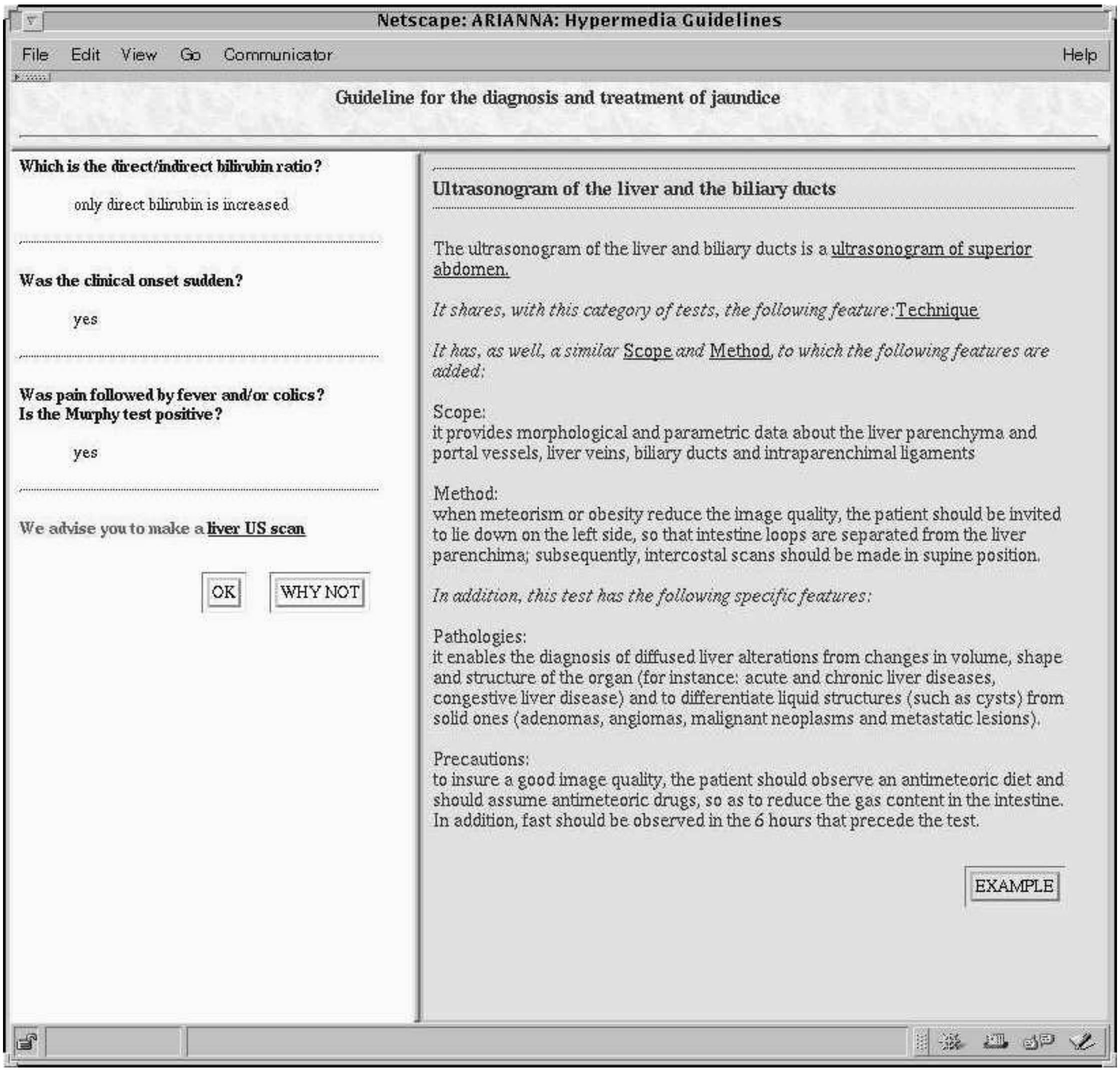

Figure 1: an example of ARIANNA's output (an english translation of the original, in italian)

[2] This metaphor reminds the idea of 'scrolling a roll', that is a "written document (as on parchment or paper) that is rolled up for carrying or storing" (Webster Third International Dictionary).

The decision to display concurrently the three mentioned components was driven by Black and others' studies on on-line dictionary information (1), who showed that making the defined words visually salient *within the text* and displaying the definitions in a *separate glossary window* at the side of the text increases the reader's willingness to access definitions. As the subjects involved in the cited study were not asked to learn the meanings of the defined words, it is not proved whether this method of displaying 'ancillary information' really affects learning: we then took the suggestion as a reasonable starting point, and plan to verify the validity of the method in our evaluation studies. In Section 2 of this paper, we illustrate the architecture of ARIANNA and describe its knowledge base. Section 3 outlines the generation method by describing, in particular, how the two main components of the frameset (guidelines and explanations) are produced dynamically, and provides an example of result of the generation procedure. Section 4 sets the implementation and evaluation state of this system and Section 5 compares it with similar projects in the field.

2. THE ARCHITECTURE OF ARIANNA

The architecture of ARIANNA is shown in **Figure 2**. Users envisaged for this first prototype are only students and doctors: we deferred to a future prototype the generation of patient-oriented guidelines, which requires an ad-hoc design. The user interacts with this tool by using a web browser; at the beginning of the consultation, she introduces a few personal data and selects a guideline name, among those known to the system. The consultation then starts, with the system displaying the selected guideline's first step. The user answers to questions or acknowledges the system's suggestions; she may check the justification of suggestions by clicking on the '*why*' button and ask for the explanation of a concept by clicking on the corresponding *anchor*. She may ask, in addition, why a different suggestion was not provided, by clicking on the '*why not*' button. This command opens a dynamically- generated menu, that shows the alternatives; once the user selected one of them, a 'contrast' explanation is generated, with the method that will be described later on[3].

All sections of the frameset are produced at run-time, from three interacting modules:

a. a **Guideline Generator** reads a decision tree in a top-down way and produces an 'interaction turn' in the scroll-shaped guideline: this can be a request of data, the suggestion of a test to be performed or a diagnostic or a treatment step; it interprets the user answer and proceeds with exploration of the decision tree by generating the next turn, until the decision process comes to a conclusion;

b. an **Explanation Generator** answers the user's request for explanation of a concept by employing knowledge embodied in several taxonomies of medical concepts and by taking into account what the user already knows or has seen during interaction;

c. a **User Modelling** component triggers a stereotype model from a few initial data and updates it during interaction, by building an individual User Profile that records the interaction history for every visited guideline, as well as the concepts that have been explained to that user.

[3] Explanation of why a particular decision was not made is a component of the question-answering ability of decision-support systems. MYCIN (3) responded to these questions by explaining *what prevented* the system from using rules that would have made that decision. The knowledge base of ARIANNA does not allow us to adopt a similar generation method: this is the reason why we exploit knowledge about concepts, to remark the salient differences between the suggested topic and the selected alternative.

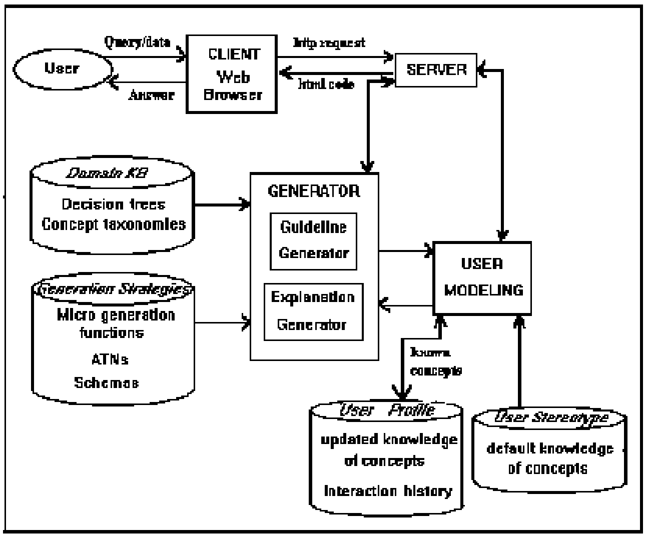

Figure 2: the architecture of ARIANNA

The Generator employes traditional NLG techniques (27): it decides what to say and the hypertext layout at each interaction step by employing three knowledge sources (about the user, the domain and the generation strategies). As emphasised in other similar systems (7), producing dynamically a hypertext to be consulted on the Web differentiates the generation process in at least two aspects. The first one is surface realisation, as the method has to encode how to write HTML tags and specify when and where to insert a link to another section of the hypermedia. The second one regards the dynamics of generation, that requires a client/server communication through CGI programs. Each link corresponds to a potential user command; each link selection calls a CGI Script that activates a Java program generating HTML pages "on the fly", according to the user category and the interaction history. We now describe in more detail the KB structure of ARIANNA.

2.1 Domain Knowledge Base : decision trees

The decision process is represented as a tree whose nodes are labelled with sets of data and whose arcs may be labelled with clinical test results; an example of decision tree is shown in **Figure 3**: this is the decision knowledge about jaundice that was employed in building the scroll-shaped guideline in figure 1**.** A node belongs to the class of Node-objects, that are identified by a name and are described by a few position-attributes: ***father*** = <father-node identifier>; ***child*** = <first-born-node identifier>; ***brother*** = <next-brother-node identifier>. Like in classical decision analysis (31), the ***Node*** class specialises in two subclasses: ***Chance-Nodes*** denote symptoms or test results; ***Decision-Nodes*** denote tests, diagnoses or a treatments[4]. These subclasses inherit the attributes of the ***Node class***, to which they add their own attributes:

Chance-Nodes: type = <C>; ***question*** = <sentence in interrogative form>; ***answers*** = <list of strings representing the possible answers>: items in this list are the labels attached to arcs connecting the node to its children, in their birth order (from left to right);

Decision-Nodes: type **=** <T/D/P>; ***suggestion*** = <test name>/<diagnosis name>/<treatment-plan name>; ***justification***=<sentence>; ***explanation***= "yes" if explanation exists, "no" otherwise.

4 " A choice node (also called *decision node*) denotes a point in time at which the decision maker can elect one of several alternative courses of action..... A *chance node* denotes a point in time at which one of several possible events beyond the control of the decision maker may take place." (31).

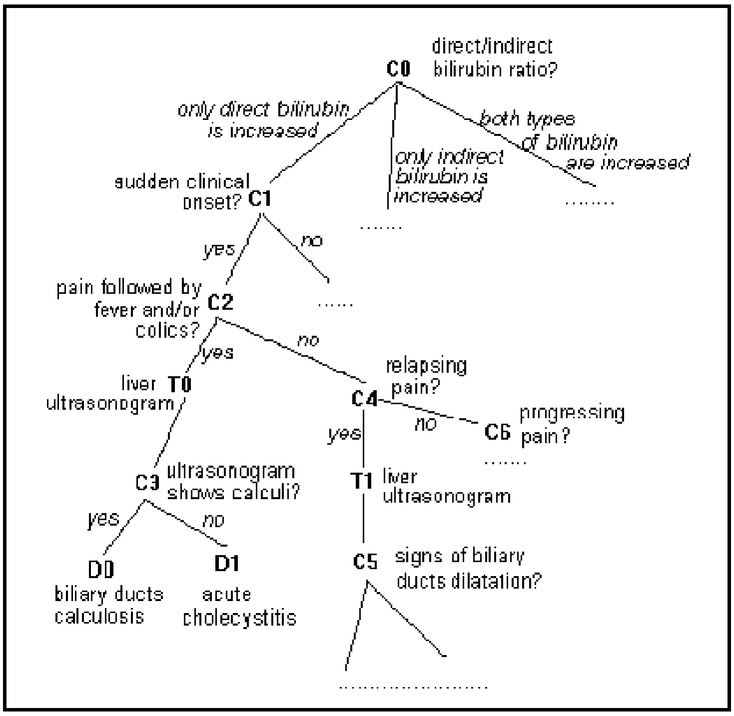

Figure 3: an example of decision tree

For example, in **Figure 3**, C0, C3 and C5 are chance nodes about test results while C1, C2, C4 and C6 are about symptoms; T0 and T1 concern the same test; D0, D1 are different diagnoses.

2.2 Domain Knowledge Base: taxonomy of concepts

Medical concepts mentioned in the three types of decision nodes are organised into separate hierarchies: this structure is suited to our purpose of generating explanations by employing the 'comparison with semantic similars' criterion, as "the assessment of similarity in semantic networks can in fact be thought of as involving only taxonomic (IS-A) links" (30). Inheritance in concept hierarchies is employed in describing the nodes, so that an intermediate node can add or specialise attributes to its father node; this representation enables our medical partners to describe concepts through specialisation (a notable advantage, given that this section of the knowledge base is the longest to build). As we will see later on, this representation is also a good starting point to generate context and user-customised explanations of concepts and to favour a 'structured' learning of concepts by the user. Our work has concentrated, so far, in building the taxonomic description of radiological and clinical tests: the ultrasound tests classification is shown, as an example, in **Figure 4**. The 'Ultrasound of the liver and biliary ducts', that was described in figure 1, is a low-level node in this taxonomy.

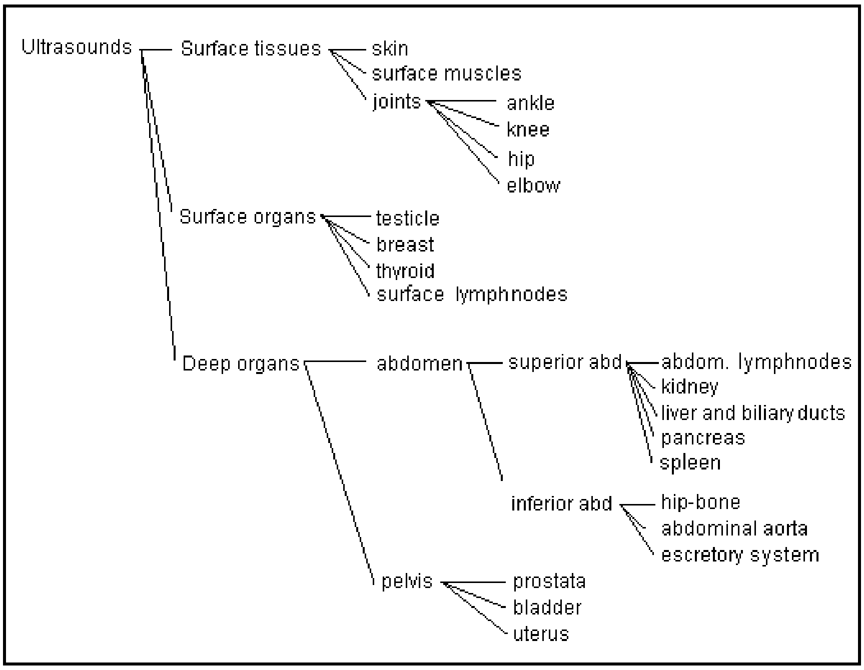

Figure 4: the taxonomy of ultrasound tests

Like in the decision trees, nodes in the concept taxonomies are subclasses of the general Node class, and are therefore described by a list of position-attributes, to which they add progressively, at every lower level of the hierarchy, new specific attributes. These new attributes are classified according to their information content: **Table 1** shows the categories of attributes for ultrasound tests. A canned text of variable length is attached to every attribute, so that explanation results of a dynamic and context-customised combination of these texts.

Class	**Name**	**Description**
category	type	test subcategory (Rx,CT etc)
	organ	organ on which the test is made
principle	technique	physical, biological and/or chemical principle behind the test
purpose	scope	what the test is applied for
	pathologies	anatomo/physiological changes the test is able to discover
selection	advantages	clinical, economical or other advantages of the test
	disadvantages	alike, in negative
execution	method	how the test should be performed
	precautions	conditions that must be insured for the test to be performed correctly
interpretation	results	how the test results may be interpreted

Table 1: attributes of test-objects

2.3 Generation strategies.

The Generator exploits a library of *micro-generation functions* which produce different, elementary outputs on the Browser: *write-sentence* and *write-anchor* write a text fragment, the second one with an anchor; *write-button* and *write-image* select and show a button or a picture, ...and so on. Other, more complex functions enable building parts of a frame (for instance: *write-title, write-frame*). These micro-generation functions are the alphabet of our generation language: as they are domain-independent, we employ them, as well, in other hypermedia generation systems like GeNet (10). They are associated as 'actions' to the arcs of the *Augmented Transition Networks* (ATNs) that represent more complex and application-dependent functions. The third and higher-level formalism that we employ in our Generator is a revised version of McKeown's *schemas* (20). Our schemas establish a connection among knowledge base components and generation strategies: they define the communication goal (in the Header) , when the schema may be applied (in the Constraints), the changes it produces in the User Model (in the Effects) and how it may be decomposed in a combination of lower abstraction schemas and/or ATNs. Two examples of schemas are shown in **Figure 5**: their meaning will be more clear when we will describe the generation method in more detail.

Name:	Explain-by-comparison-with-father	**Name:**	Describe-Peculiarities of-a-Concept
Header:	Describe (S H C)	**Header:**	Describe-Peculiarities (S H Ck C)
Constraints:	WantToKnow (H C) and Exists Ck (Father (Ck C) and KnowAbout (H Ck))	**Constraints:**	Exists ai (Attribute (ai C) and not Attribute (Ai Ck))
Decomposition:	Call ($Title (C)) Call ($Introduce-father (Ck C) Describe-Identities (S H Ck C) Describe-Differences (S H Ck C) Describe-Peculiarities (S H Ck C)	**Decomposition:**	Call ($Introduce-Peculiarities (Ck C)) Forall ai (Attribute (ai C) and not Attribute (ai Ck)) Call ($Illustrate-Attributes (ai))
Effects:	KnowAbout (H C)	**Effects:**	Forall ai (Attribute (ai C) and not Attribute (ai Ck)) KnowAbout (H ai)

Figure 5: two examples of explanation generation schemas

2.4. User Modeling

The decision of what to model and how to model it is driven by the decision of *what to adapt* and *how to adapt it*. The following features of ARIANNA are adapted to the user and the context:

- the system *goal* (decision-support vs training): in the first case, users are guided to follow the appropriate diagnostic-therapeutic procedure for a specific patient; as interaction may break off for several reasons, the interaction history is registered in the User Model, so that consultation of the guideline can start, later on, from the same point at which it was interrupted, by initially summarizing that situation. If the system goal is training, users explore the guideline in a 'what-if' mode, by not necessarily examining a specific clinical case: they can then go back to a previous chance node, to change their answer and check how the system suggestion changes accordingly;

- the *concepts explanation mode*: a concept may be explained either individually or by comparison or contrast with other concepts;
- the *individual concept description*: this may include a selected subset of the concept attributes.

The system goal, as well as the concept description, are adapted to the user level of education: the 'training mode' is reserved (by default) to students, the 'decision-support' mode to doctors; attributes introduced in the descriptions addressed to students and to doctors are different; the concept explanation mode is adapted to knowledge of other concepts in the taxonomy. We can now describe how users are modeled in ARIANNA:

- a *stereotype* is triggered from answers to 'educational level', 'years of experience' and 'type of specialisation' questions. The stereotype's body includes a list of default properties of the type: *KnowAbout (U Ci)* , where Ci is a concept in one of the taxonomies in the Domain KB. Stereotypes are such that: (i) undergraduate students are presumed to ignore all concepts, (ii) graduating students are presumed to know only the concepts at the highest level, in the taxonomies that correspond to their specialisation; (iii) general practitioners know highest level concepts in all taxonomies and, finally, (iv) specialist doctors are presumed to know all concepts, in the taxonomy that corresponds to their specialisation.[5]

  If we consider, for instance, the taxonomy of ultrasound tests that is shown in figure 4, graduating students and general practitioners are presumed (by default) to be familiar only with ultrasounds in general, and how they apply to surface tissues and organs and to deep organs. Radiologists, on the contrary, are presumed to know all the concepts in the taxonomy.
- immediately after activation, the stereotype is copied into a *User Profile* which is updated dynamically during interaction, by revising the list of *KnowAbout* properties and the history of interaction with the guideline the user is examining. This model is stored between sessions, so that the default stereotype is only employed at the first interaction of a user with ARIANNA, while it is substituted, from then on, with the updated individual profile (as shown in the system's architecture, in figure 2). Updating of the user knowledge is based, in particular, on the rule: *Explain (S U Ci) KnowAbout (U Ci).* We know that this assumption has long been discussed in the hypermedia generation environment, where it is claimed that one cannot be certain that a user really reads the displayed information and really learns it: to cope with the risk of such a wrong default assumption, we insert an anchor to explanation of attributes (or concepts) that the user was presumed to know, so as to give her the possibility to access this information.

3. THE GENERATION METHOD

As we showed in Figure 1, the Generator includes two main components, that are aimed at producing the guideline and the explanations. When interaction starts, the generator sets up the frame-page components and initializes its Title and Header; the decision tree is then examined starting from its root and by descending one level at each interaction turn.

[5] Stereotypes were defined, in this research, in a subjective way. We could profit of the teaching experience of medical co-authors and also of the background knowledge that we had acquired in the scope of the EU Project EPIAIM (8). In that Project, an experimental study showed a correlation between the level and type of qualification of doctors and the complexity of concepts with which they were familiar: although the domain of that study was different (epidemiology), we assumed that this general finding could be generalised to the domain of ARIANNA.

3.1 The Guideline Generator

The path followed in examining the decision tree depends on the user answers to questions generated from chance nodes. At every step of the tree-exploration process, the interaction history is updated. This history is employed to re-generate the guideline frame at each interaction turn, in two ways:

a. the *diagnostic reasoning mode* is employed in all steps that precede a diagnostic or therapeutic decision; the list of nodes in the interaction history is examined sequentially, and type-specific ATNs are applied to translate each of them into a sentence; buttons, anchors and lists of radio buttons are generated as well, to enable users to introduce their reply;

b. the *synthesis mode* is applied when the consultation comes to an intermediate or final decision; the interaction history is then simplified by keeping only the user answers: this enables the generator to produce a summary that reminds the reasons why the decision was taken (results of tests and other evidence). In the two cases, after generating a message, the system gives the control to the user, who performs one of the following actions: (i) click on an *acknowledge ('OK') button,* to make clear that she got the system suggestion; (ii) *answer to a question*, (iii) *ask to revise the consultation* by changing the answer given to a previous question (*'change' button*) or (iv) *ask for explanation* of a concept (click on an *anchor*) or a *justification* of a suggestion (*'Why' button*). After an 'acknowledge' or an answer, the system goes on by examining the next node in the decision tree. After a 'change' command, it goes back to the selected node. In this part of the guideline, the initiative is mainly left to the system, the user having only two opportunities to escape from the suggested line of reasoning: she can stop it temporarily to ask an explanation or can go back to a previous step to revise her answers. As we said in 2.4, we consider, though, that the second opportunity has a 'tutorial' scope, and should be enabled, on principle, only to students, whereas we assume that doctors will usually employ ARIANNA to take decisions about specific and well-defined clinical cases. Figure 1 shows, on the leftside frame, an example of part of the guideline responding to a 'decision' goal (therefore, with with no 'change' commands); this text was generated from nodes (C0, C1, C2, T0) of the decision tree in figure 3.

3.2 The Explanation Generator.

The *Explanation frame* enables the user to enter into a stand-alone hypermedia of medical concepts through a link from the guideline frame. For instance, again in figure 1, the rightside frame was displayed after the user clicked on the 'liver US scan' anchor in the scroll-shaped guideline. Explanation pages are generated on the fly by applying some criteria that were drawn from analysis of medical textbooks and from the teaching experience of doctors in our team:

a. explain a concept by specifying its position in the hierarchy and by illustrating its attributes in a logically predefined order: in the case of tests, this order is the sequence with which attributes are considered when a test is selected, performed and interpreted;
b. mention only the attributes that are adequate to the users' likely level of knowledge in the domain and to their interests;
c. whenever possible, describe the 'target' concept *by referring to other ('base') concepts* with which the user is familiar, so as to favour incremental and structured learning;
d. when describing a concept in reference to a known one, specify their relationship and describe *commonalities and alignable and nonalignable differences* between them (19). Treat: (i) *commonalities* (attributes with identical values)

by only mentioning their name and by inserting an anchor to their description, (ii) *alignable differences* (attributes taking different values in the two concepts) by mentioning the difference and then describing their value in the target and (iii) *non-alignable differences* (attributes that are peculiar of the target) by fully illustrating the attribute value[6]

e. when explaining a concept *in contrast with an alternative* candidate, emphasize only differences in 'salient' attributes.

When the user requires explanation of a concept, ARIANNA inspects the user model to assess whether it can find any other concept in the taxonomy that the user knows; if several concepts are found, the 'nearest' of them *in family relationship* terms is selected, the nearness being measured as a function of the length of the path which links the two concepts in the taxonomy. If several concepts with the same degree of family relationship are found (for instance, the grandfather and a brother, or two brothers), the most similar of them is selected, as the candidate which has more commonalities with the concept to be explained. This idea of similarity combines the two measures described in (30) to evaluate semantic relatedness in network representations of concepts: our degree of family relationship corresponds to the "distance between nodes in the IS-A hierarchy", whereas the number of commonalities is close to Resnik's "extent to which two concepts share information in common". The first index alone would not insure selecting the 'most similar' comparison concept, as links do not necessarily represent 'uniform distances' in our hierarchies: it only enables us to make a first selection of candidates to comparison. Considering commonalities leads us to privilege, in general, comparison with lower level nodes (a brother to a grandfather, a nephew to a grand-grand father or an uncle), that hold a more detailed reference knowledge.

Once similarity research has been completed, a generation strategy is applied which depends on the research results:

1. *stand-alone description* is applied when no other relative is presumed to be known to the user. In this case, the position of the concept in the hierarchy is specified and all the attributes which are suited to the user characteristics are illustrated, by inheriting values from ancestors.

   For space reasons, we only provide an example of how the *Method* attribute of liver USs is explained to an undergraduate student who is not familiar with this test:

   " *This test is performed by pressing a probe against the skin. Before doing that, a thin layer of conductive gel must be spread between the skin and the transducer, so as to avoid any air between them, that would disturb the test. The patient is usually invited to lie down: scans are performed, by moving the probe, in several plans, to get more information. Images are displayed in real-time on a monitor, and their resolution and contrast can be regulated. This test also enables transferring the image on paper or X-rays, for a permanent storage.*

   If the test is applied to deep organs, a low-frequency probe (3.5 to 5 MHz) must be employed. The probe is placed over the superior abdomen; patients are invited to breathe deeply and to keep their breath during the test, to enlarge the surface that can be scanned with the US: this improves examination of deeper regions".

   This text includes values inherited from USs in general (the first two paragraphs) and from the USs of deep organs (the third one).

[6] as concepts in the hierarchies are described 'by specialization', alignable or non-alignable differences may exist between two nodes along the same path; the first one occur when a lower level node 'specializes' the value of an attribute by adding details to values of higher level nodes (as for 'technique', in the diagnostic tests), the second one when it adds a new property (for instance: 'pathologies' are not defined for ultrasonograms of abdominal organs but are specified for ultrasonograms of specific organs). Of course, both types of difference may occur, as well, between nodes at the same level in the hierarchy: in this case, alignable differences correspond to 'different' rather than 'specialized' values of attributes.

2. *explanation by comparison* is applied every time a known concept in the same taxonomy is found; in this case, the family relationship between the two concepts is described and commonalities and differences are illustrated; alignable differences are given as *specialisations* of attribute values in comparisons with ancestors or as *differences* in these values, in comparisons with brothers.

   The text in figure 1 is an example of this type of explanation. The ultrasonogram of the liver and the biliary ducts is explained, in this case, by comparing it with ultrasonograms of superior abdomen (its 'father' node, in the hierarchy in figure 4), Commonalities include, in this case, only the Technique, while alignable differences are Scope and Method. Pathologies and Precautions are non-alignable differences. The example is a literal translation of the output in italian. This explanation is addressed to those who are familiar with USs of superior abdomen because they saw this explanation in a previous phase of interaction.

3. *explanation by contrast* is applied when the user is familiar only with an 'alternative' test, diagnosis or treatment or when he/she explicitly requests to contrast the two concepts, by clicking on a 'why not' button. An 'alternative' concept is defined by a set of conditions on the values of some attributes; for instance: an alternative to a diagnostic test has a different value of the 'type' attribute and the same value of the 'organ' attribute. In this case, as we said, explanation includes only the attributes that are considered as 'salient' for the contrast. Let us assume, for example, that, in exploring a guideline about thyroid nodules, the user asks to ARIANNA why it did not suggest performing a scintiscan rather than a US of the thyroid. In this case, the 'salient' attribute is the *Purpose*, and the system will generate the following explanation:

   " *USs and scintiscans of the thyroid are applied to different purposes.*

   USs reproduce exactly the anatomy of the thyroid, by evidentiating its two lobes and the isthmic region. They enable, as well, displaying the dimension of the thyroid, the structure of its parenchima and focal lesions, when they exist. This test is normally suggested as the first diagnostic approach to the thyroid pathology, as it is a simple, fast and (in the majority of cases) resolutive exam.

   Scintiscans provide, on the contrary, functional rather than anatomical information, as they enable evaluating the activity of this gland and of its nodules, when they exist. Thyroid nodules may be described, with this test, as 'hyperactive' or 'hot' areas, when they are made of highly cellular tissues with a degree of functionality higher than the normal parenchima; hypoactive or 'cold' areas correspond, on the contrary, to low functioning tissues such as cysts, haematomas or carcinomas."

In defining these generation criteria, we made reference to Markman and Gentner's work about similarity comparisons, in which it is claimed that any comparison process should promote 'structural alignment' (that is, analysis of similarities of structured representations of concepts, objects and words) by mentioning commonalities and alignable differences (19). According to Gentner and Holyoak, perceiving similarities enables the user to organize entities into 'familiar categories', thus favouring the learning process (15). Adding description of non-alignable differences favours, in our case, incremental learning of concept properties, by showing what is peculiar of the concept to be explained, if compared with another concept the user already knows. Explanation strategies are formalized into *explanation generation schemas*, of which we show two examples in Figure 5**.** The leftside schema describes how to explain a concept in comparison with its father: its Decomposition reminds to lower level schemas and activates two distinct ATN-functions, by a Call($x); the rightside schema describes how to illustrate non-alignable differences between two concepts. The classes of attributes to be included in an explanation are linked to the user characteristics: some of them (the category) are not mentioned in the explanation; others are defined as 'basic', and enter in any description (in the case of tests: principle and purpose); the remaining ones are not essential, and may be more or less interesting for different user categories. In prototyping ARIANNA, we established (i) which categories are relevant to doctors, (ii) which one to students and (iii) which one are 'salient' for contrast explanations. For example, for tests: (i) execution, (ii) selection and interpretation, (iii) purpose, respectively. At present, the user is not given the opportunity to modify the system-initiated attribute selection like it happens, for instance, in AVANTI (13). These are, however, decisions that we plan to verify in our ongoing evaluation study, to revise them if needed. After a concept has been explained, the User Profile is updated: control is then given again to the user, who can continue browsing in the concept hypermedia or can return to the main guideline consultation.

4. IMPLEMENTATION AND EVALUATION

A prototype of ARIANNA has been fully implemented and has been tested by some 'expert' doctors, who used it to build guidelines in different medical domains. The size of decision trees in these guidelines ranges from 30 to 80 nodes, while the concepts arranged in taxonomies are about 60: however, these figures provide only an instant picture of a daily changing situation. Doctors find easy to represent their knowledge with our formalism and, though building taxonomies of concepts takes some time, they do it incrementally as far as they define new guidelines, by adding concepts in pre-defined taxonomies. Some knowledge acquisition supporting tool would undoubtedly simplify their work: we are considering, at present, the possibility to bring up in a unique tool the capabilities of PRO-FORMA (14) in engineering decision criteria and of HealthDoc (17) in acquiring concept attribute values and in establishing the conditions under which each of them should enter into a user-tailored explanation. However, in our case these conditions should not be defined *for each concept*, but *for each taxonomy* of concepts, in order to be sure that concepts in a taxonomy are described consistently.

We are undertaking a systematic study in cooperation with the Department of Psychology, University of Reading, to validate the usability of the scroll-shaped guideline and of the concept explanation component by final users (something we always did in our previous experiences with explanation systems in the medical domain). Evaluation by final users is especially crucial in a system like ARIANNA: guidelines cannot be compulsory, and are successful only if they are applied by a large number of doctors, in different clinical situations. They must therefore be easily *accessible*, *convincing* and *flexible* so as to apply to structures with different resources, and must be *clear and easy to use*: a set of properties that we plan to assess in the mentioned evaluation study.

In the next release, we envisage to work at increasing the *flexibility* of guidelines, so as to adapt suggestions to the resources available and to the patients' preferences or intolerances; this requires attaching one or more alternatives to decision nodes, ranking them by order of preference and generating dynamic menus to enable the user to select an alternative when the system suggestion is not acceptable. Our second goal is to integrate concept explanations with context-customized image examples represented as semantic maps, on which the user may click to get details about the anatomy and the pathology of that particular case.

5. RELATED WORK

Before comparing ARIANNA with similar systems, let us justify why we considered that dynamic generation was advisable in this case. Flexible hypermedia may be seen as statically or dynamically-generated documents (2, 23); the decision on whether to adopt the first or the second approach in building a particular system relies on how much the authors want to vary the system's answers to the user's requests (18). This decision also depends, in our view, on the two following estimates: (i) the portion of the hypermedia that will likely be examined at every interaction, and (ii) the frequency with which the hypermedia will have to be updated. The first factor establishes whether it is convenient to prepare and store the whole set of electronic documents, or whether it is preferable to generate, at runtime, only that portion of the network that responds to the user's information needs. The second factor establishes whether it is convenient to prepare and store a set of documents that will become obsolete when knowledge behind them will be updated, or whether it is preferable to automatically generate the documents when the knowledge base is revised. Electronic encyclopedias, which are, by now, the most remarkable example of dynamic hypermedia, have been conceived to respond to the first need; the objects that can be viewed are potentially very many, but only a few of them are examined by each user: dynamic generation therefore appears to be convenient. ARIANNA is an example of a system that is designed to respond to both needs, as it is applied to two knowledge sources: the decision tree and the taxonomy of concepts. The knowledge of medical concepts is potentially large but does not change very frequently; decision knowledge is more limited but is updated more frequently, to make it in line with new medical research results. In addition, as every concept can be explained, in principle, in reference with any other concept in the same taxonomy, the number of explanations to be produced in a 'static' hypermedia grows with n(n-1)/2, that is in a quadratic way with the number n of concepts in the taxonomy. Let us consider the example of the taxonomy of ultrasound tests; this taxonomy includes 30 concepts; therefore, 435 different explanations should be generated in advance to compare every single concept in the taxonomy with every other concept in the same taxonomy; this number doubles if variations according to attributes are introduced; they still increase if variations due to 'contrast' comparisons with tests in other taxonomies are considered. These are the reasons why, though we started our experience with hypermedia guidelines as statically generated networks of pages, after building and validating some of them we moved towards a dynamic generation solution.

Let us now say what is new in our system. The application domain of ARIANNA (medical decision making) was the testfield of the majority of knowledge-based systems for several years: therefore, ARIANNA has several natural referents in experiences that are different in the purpose and in the methods applied. First, a guideline is not an expert system: in expert systems, doctors receive an advice about the most plausible diagnosis and the preferable treatment *once they have collected a set of data they need to interpret* (3). Typical forms of reasoning are, in these systems, frame-based activation of hypotheses, uncertain abduction (to select a hypothesis given a set of evidences) and rule-based refinement of hypotheses. In guidelines, *doctors are leaded to perform only the tests which are really needed*, as avoiding inessential tests and insuring that all the essential ones are performed in the correct sequence are the main goals of these protocols (12); this is the reason why knowledge representation in ARIANNA has more in common with decision analysis than with expert systems. As far as the explanation component is concerned, there are several notable experiences in the generation of medical explanation texts, all addressed to the patient and including a combination of general and personalized information, taken from hierarchies of medical 'problems' and from the medical record (24, 17, 4, 5, 6, 29, 21); we made a similar experience in the field of drug prescription, where patients were one of the categories to which explanations were addressed (9). Although the goal of these systems may appear to be similar to the goal of the explanation component of ARIANNA, this is not the case. Explanation is a very general term: in the mentioned studies, it is aimed at *informing and educating* the patient about some personal and specific problem, and general clarifications of terms are added only to favour understanding of this information; the idea is that increasing awareness of own problems will contribute to changing the patients' behaviour and their compliance with treatment or their self-care attitudes. ARIANNA's explanations, on the contrary, are addressed to health care staff and are a typical example of 'ancillary' information which completes the main system component (the guideline) and is aimed at *teaching* the meaning of concepts, in order to insure that guideline suggestions are throughfully understood and applied correctly (26).

As for the method adopted: the metaphor of 'scrolling dynamically generated roll' that we adopted for displaying the decision component in the frameset matches the conversational view of hypertexts that is discussed by Oberlander et al in this Volume (25). By scrolling the guideline frame, users can examine the whole interaction history, in the form of a dialogue in which the user and the system alternate. In the tutoring mode, they can go back to a previous turn of the dialogue, to change their answer and see how the conversation changes, consequently.
Explanation by comparison is a popular approach to computer-based concept description, after McKeown's *compare and contrast* schemas (20). In PEBA-II, which is probably the most notable experience in comparison generation (22), different types of comparison (illustrative, clarificatory and direct) are selected according to the user knowledge of the 'basic' concept, to its relationship with the 'target' and to the number of salient properties they have in common. Comparisons may include only one particular property, a list of similarities and differences or a point-by-point bi-focal description of all properties in the two concepts. The overall goal of the different types of comparison is 'to easy understanding of concepts by avoiding confusion', and the method is mainly based on a salience-ranked attribute taxonomy. In ARIANNA, concept description strategies are the consequence of the teaching goal of the explanation component: (i) to avoid too long explanations that would distract the user from the main system purpose (a goal similar to the 'avoid cognitive overload' in (32); (ii) to improve learning of concepts by clarifying their structural relations and (iii) to respond to implicit or explicit '*why-not'* questions about alternative options that the system did not choice in the decision process. The salience of attributes is employed only to establish what to include in the explanation; once the most similar concept has been selected, comparison only evokes the similarities and concentrates on a detailed description of differences, by mentioning only properties appropriate to the user's level of knowledge. As these properties have the same salience, their order of presentation is determined by the values they take in the two concepts: commonalities always come first, alignable differences second and non-alignable differences at the end. These strategies correspond to what we consider a systematic description of concepts, which is typical of *learning* tasks, as opposed to *information-seeking* ones (16).

ARIANNA has all the limits and the interest of a real-world application, in which pragmatics prevails on elegance of techniques: the main example of this choice is the use we made of schemas. We found schemas an efficient formalism to represent generation strategies non ambiguously and employed them as a 'working language' of our interdisciplinary group, to agree upon how to translate the knowledge base into hypermedia. We did not feel, however, as an obligation to represent schemas in a separate KB and to activate them by an inference engine: by employing Java as a declarative language, we could directly encode schemas as Java programs, thus avoiding to increase processing time. This was an advantage in ARIANNA, where the domain knowledge base is updated frequently while the generation strategies are not, and therefore processing time is more important than generation strategies' revisibility. As a consequence of this choice, the system is very flexible in the generation of new guidelines and new concept explanations but not much flexible in the change of generation criteria: if schemas or ATNs are changed, a new Java program has to be produced. However, new programs can be easily produced by modifying the old ones, due to the strict correspondence between the knowledge structures and the programs implementing them. The second simplifying choice about the method concerns the grain size of our KB about concepts; we use human-written canned texts rather than a semantic representation language for at least two reasons: first, we don't need to vary, at present, the presentation *style* of our messages, as they are all addressed to medical people; second, we lack of a surface generator in italian, of the type of PENMAN or its derivates: the real 'value added' of the explanation component of ARIANNA (to use Reiter's words) is therefore in ensuring content-adaptation rather than a refined style (28). Clearly, we cannot go, in our adaptation, below the sentence level and can do a limited 'textual repair' in assembling texts (nothing similar to what is done, for instance, in HealthDoc (17) or with other sentence-planning methods). This simple solution does not prevent us from obtaining a satisfying level of variability in the presently generated texts, but would probably be inadequate to producing, from the same KB, texts addressed to patients, as these should be very different in the content, in the style and in the general 'empathy' of the message.

Acknowledgements

We acknowledge Anna Altamura, Luigia De Marco and Michele Chierico for contributing to implementation of ARIANNA in the scope of their dissertations in computer science, and Francesco Giovagnorio and the anonymous reviewers for their helpful critiques and suggestions. Maria Milosaljevich and Peter Brusilovsky were so keen on helping us to improve this paper, that we were about to ask them to join the authors' team.

MAIN REFERENCES

1. A Black, P Wright, D Black and K Norman: Consulting on-line dictionary information whire reading. *Hypermedia,* 4, 3, 1992.
2. P Brusilovsky and E Schwarz: User as Student: Towards an Adaptive Interface for Advanced Web-Based Applications. *User Modeling 97*, Springer Wien, 1997.
3. B G Buchanan and E H Shortliffe: *Rule-Based expert systems: the MYCIN experiments at the Stanford Heuristic Programming Project.* Addison-Wesley Publishing Company, 1984.
4. B G Buchanan, J Moore, D E Forsythe, G Carenini, G Banks and S Ohlsson: An intelligent interactive system for delivering individualized information to patients. *Artificial Intelligence in Medicine,* 7, 2, 1997.
5. A J Cawsey, K Binsted and R Jones: Personalized explanations for patient education. *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, 1995.
6. A J Cawsey, B L Webber and R J Jones: Natural language generation in healthcare. *Journal of the American Medical Informatics Association,* 4, 6, 1997.
7. R Dale, J Oberlander, M Milosavljevic and A Knott: Integrating Natural Language Generation and Hypertext to Produce Dynamic Documents. *Interacting with Computers*, 10, in press.
8. F de Rosis, S Pizzutilo, A Russo, D C Berry and F J Nicolau Molina: Modeling the user knowledge by belief networks. *User Modeling and User-Adapted Interaction,* 2, 1992.
9. B De Carolis, F de Rosis and F Grasso: Generating recipient-centered explanations about drug prescription. *Artificial Intelligence in Medicine*, 8,2, 1996.
10. B De Carolis, F de Rosis, S Pizzutilo: Generating User-Adapted Hypermedia from Discourse Plans, *Springer LNAI 1321*, 1997.
11. B. De Carolis, G. Rumolo, V. Cavallo. User-adapted multimedia explanation in a clinical guidelines consultation system. *Springer LNAI. 1211*, 1997.
12. M Field and K Lohr (Eds): *Clinical practice guidelines. Directions for a new program.* National Academy Press, 1990.
13. J Fink, A Kobsa and A Nill: Adaptable and adaptive information provision for all users, including disabled and elderly people. *In this Volume.*
14. J Fox, N Johns, A Rahmanzadeh and R Thomson: PROforma: a method and language for specifying clinical guidelines and protocols. *Proceedings of Medical Informatics In Europe.* IOS Press, 1995.
15. D Gentner and K J Holyoak: Reasoning and learning by analogy. *American Psychologist*, 52, 1, 1997

16. N Hammond and L Allinson: Extending hypertext for learning: an investigation of access and guidance tools. *People and Computers V*, HCI 89, Cambridge University Press, 1989.
17. G Hirst, C Di Marco, E Hovy and K Parsons: Authoring and generating health-education documents that are tailored to the needs of the individual patient. In *Proceedings of User Modeling 96.* Springer CISM Courses and Lecture Series n 383, 1996.
18. A Knott, C Mellish, J Oberlander and M O'Donnel: Sources of flexibility in dynamic hypertext generation. *Proceedings of the 8th International Workshop on Natural Language Generation*, Herstmonceux Castle, 1996.
19. A B Markman and D Gentner: Commonalities and differences in similarity comparisons. *Memory and Cognition*, 24, 2, 1996.
20. K R McKeown: Discourse strategies for generating natural-language texts. *Artificial Intelligence*, 27,1985.
21. S W McRoy, A Liu-Perez and S S Ali: Interactive computerized health care education. *Journal of the American Medical Informatics Association*, 5,4, 1988
22. M Milosavljevic: Content selection in comparison generation. *Proceedings of the 6th European Workshop on Natural Language Generation,* Duisburg, 1997.
23. M Milosavljevic and J Oberlander: Dynamic hypertext catalogues: helping users to help themselves. in *Hypertext 98.* ACM Press, 1998.
24. J Moore: *Participating in explanatory dialogues.* AMC-MIT Press Series in Natural Language Processing, 1995.
25. J Oberlander, M O'Donnell, A Knott, C Mellish: Conversation in the museum: experiments in dynamic hypermedia with the Intelligent Labelling EXplorer. *In this Volume.*
26. R Pilkington and A Grierson: Generating explanations in a simulation-based learning environment. *International Journal of Human-Computer Studies*, 45, 1996.
27. E Reiter: Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? *Proceedings of the 7th International Workshop on Natural Language Generation*, 1994.
28. E Reiter: NLG vs templates. *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, 1995.
29. E Reiter and L Osman: Tailored patient information: some issues and questions. *Proceedings of the ACL-97 Workshop on "From research to Commercial Applications: making NLP technology work in practice"*, 1997.
30. P Resnik: Using information content to evaluate semantic similarity in taxonomy. *Proceedings of the 14th International Joint Conference of Artificial Intelligence,* 1995.
31. H C Weinstein and H V Fineberg: *Clinical Decision Analysis.* W B Sauders Company, 1980
32. I Zukerman and R McConachy: Generating explanations across several User Models: maximizing belief while avoiding boredom and overload. *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, 1995.