The automated generation of Web documents that are tailored to the individual reader[*]

Chrysanne DiMarco
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
E-mail: `cdimarco@logos.uwaterloo.ca`

Mary Ellen Foster
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
E-mail: `mefoster@logos.uwaterloo.ca`

Abstract

Many studies in communication have shown that presenting information in a manner that is tailored to the characteristics of a particular audience can have a significant effect on an individual. Incorporating this kind of tailoring facility in a system for the management and presentation of Web documents would be a very important enhancement of the Web's current state of development. Our long-term research goal has been to investigate and develop theories of text composition, in particular, computational models of rhetorical structure and lexical semantics, which are particular problems for incorporating style in natural language systems. Recently, we have begun to apply this research to developing Web-based natural language generation systems that can tailor documents to the individual user on demand.

Introduction

A recent and growing development in Web applications has been the advent of various tools that claim to "customize" access to information on the Web by allowing users to specify the kinds of information they want to receive without having to search for it or sift through masses of irrelevant material. But this kind of customization is really just a crude filtering of raw Web material in which the user simply selects the "channels" of information she wishes to receive; this selection of information sources is hardly more "customization" than someone deciding to tune their television to a certain station. True customization, or tailoring, of information would be done *for* the user by a system that had access to an actual *model of the user*, a profile of the user's interests and characteristics. And such tailoring would involve much more than just selecting streams of basic content: the content of the text, whether for an on-line Web page or a paper document, would be carefully selected, structured, and presented in the manner best calculated to appeal to a particular individual.

It is well known from studies in communication that presenting information in a manner that is tailored to the characteristics of a particular audience can have a significant effect on an individual. Incorporating this kind of tailoring facility in a system for the management and presentation of Web documents would be a very important enhancement of the Web's current state of development.

The current state of "customization" on the Web

Although there are quite a number of tools now available that claim to do "customization" or "personalization" of material on the Web, very few of these products do anything involving the tailoring of the actual text to a particular audience, and none use anything more than a rudimentary description of the user.

One of the first of these products, and typical of its kind, is the PointCast Network. The PointCast client can select from many "channels" of news to view continuously updated stock prices, sports scores, weather information, and the like. There are currently numerous other programs currently available or in development which provide the same sort of continuous, proprietary content. Among them are Marimba's Castanet, Intermind Connector, and BackWeb.[1] An-

[*] This paper can be found at the following Web site: http://logos.uwaterloo.ca/∼healthdo/Publications/aaai97-symposium.html. The official version of this paper has been published by the American Association for Artificial Intelligence (http://www.aaai.org).

[1] These products can be found at the following Web sites:
http://www.pointcast.com,
http://www.marimba.com/products/castanet.html,
http://www.intermind.com,
http://www.backweb.com.

other common approach to sifting through and selecting information on the Web uses, instead of proprietary content, customized subsets of publicly available Web pages which fit a user's preferences. For example, FreeLoader, Cognisoft, iFusion's ArrIve, and First Floor's SmartBookmarks use this idea.[2]

Although these products are often advertised as tools for customizing or personalizing Web access, the "customization" provided really only involves allowing the user to choose among streams of raw information content. Nothing is done to the language or layout of the document itself to make it more appealing or accessible to the individual user.

Two slightly more sophisticated products, Netscape's "Power Start Page", and The Microsoft Network's "Custom Start Page" use an approach similar to the systems mentioned above, but also allow users to set up a customized front page, with links to various predefined locations and a small number of personalized links. Both also allow some simple modification of presentation style—MSN permits a background sound, while Netscape allows changing of the layout and colour scheme. BroadVision claims that its "One-To-One" product can "build and manage visitor profiles and dynamically match profile data to Web content to personalize [a] site for each visitor, each time they visit", but the nature of the personalization is described in only vague terms.[3]

Despite the proliferation of the kinds of products described above, only a rare few attempt any formal user modelling as a basis for customizing a document. One of the most advanced is MicroMass's "tailoring engine" for Web documents, which uses the results of an on-line questionnaire to produce, in real-time, a health information newsletter tailored, down to the paragraph and sub-paragraph level, to a user's specified medical conditions and lifestyle.[4] MicroMass also has a product named "IntelliWeb" that dynamically creates Web pages from a content database. The selection and presentation of information is based on data either entered by the user, provided from a customer database, or obtained automatically by means of a World Wide Web profile administrator. In MicroMass's approach to customization, however, the language of the text selected for a particular user cannot be tailored in style or structure; the nature of the "customization" is really just selecting among prefabricated, simple blocks of text. IntelliWeb also seems more suited to tailoring a single Web-page template by changing a few simple pieces of text or a couple of illustrations. It does not appear to be intended for customization of textual materials of any length greater than a few sentences.

If the Web document designer wishes to write and present material in a way that will communicate well with the user, then just displaying the most relevant chunks of information will not be sufficient. For effective communication, both the form and content of the language used in a document should be tailored in rhetorically significant ways to best suit a user's particular personal characteristics and preferences. Ideally, we would have Web-based natural language generation systems that could produce fully customized and customizable documents on demand by individual users, according to a formal user model. As a first step in this direction, we have been investigating applications of our earlier work on pragmatics in natural language processing to building systems for the automated generation of Web documents tailored to the individual reader.

The HealthDoc approach to automated generation of tailored natural language documents

The HealthDoc project

Our long-term research goal has been to investigate and develop theories of text composition, in particular, computational models of rhetorical structure, lexical semantics, and fine-grained meaning, which are particular problems for incorporating style and rhetoric in natural language systems. In the past several years, we have addressed problems of syntactic style, to understand how particular syntactic structures can convey corresponding stylistic effects (DiMarco 1990, DiMarco and Hirst 1993a). We have applied our theory of style to various problems in language analysis and generation. We have produced prototype implementations of a stylistic analyzer and generator (Hoyt 1993, Hoyt and DiMarco 1994, Green 1992, Green and DiMarco 1996). We have also been working on the problem of style and lexical choice, with an emphasis on representing near-synonymy in generation systems (DiMarco and Hirst 1993b, DiMarco, Hirst, and Stede 1993, Hirst 1995).

This earlier work is now feeding in to our HealthDoc project (DiMarco, Hirst, Wanner, and Wilkinson 1995), which is developing natural language generation systems for producing health-information and patient-education documents that are customized to the personal and medical characteristics of the individual patient. The HealthDoc approach is applicable, we believe, to many kinds of situations in which the ability to target tailored documents to the characteristics of specific users would be desirable. This kind

[2] The respective Web sites are:
http://www.freeloader.com,
http://www.cognisoft.com/product.htm,
http://www.ifusion.com/company/arrive/what,
http://www.firstfloor.com/sb20data.html.

[3] The Web sites are:
http://personal.netscape.com/custom,
http://www.msn.com/csp/choices/first.asp.
http://www.broadvision.com/products/v2datasheet.html.

[4] The Web site is:
http://www.micromass.com/.

of customization would involve much more than just producing each document in half a dozen different versions for different audiences. Rather, the number of different combinations of factors might easily be in the tens or hundreds of thousands, and it would be impossible to produce, in advance of need, the large number of different editions of each publication that would be entailed by individual tailoring of information. This is exactly the kind of situation that we face in developing Web-based natural language generation systems that could produce tailored documents for the individual Web user.

Recently, we have been experimenting with the application of HealthDoc techniques to develop a related system, WebbeDoc, that can customize Web documents on demand, according to a profile of an individual user. In the sections below, we describe the design and implementation of the first WebbeDoc prototype, then outline the kinds of long-term research issues which will need to be addressed to develop a full working system. To start, we present an overview of the concepts and techniques that HealthDoc uses and that are being adapted for WebbeDoc.

The master document and generation by selection and repair

The key idea in the HealthDoc approach to producing tailored documents is that we start from an existing *master document* that is then customized for a particular audience. A master document is an encapsulation of all the variations on a given topic that might be needed for any potential reader; it is represented in an abstract, albeit language-dependent, text specification language that expresses not only the content of the document but also information that will assist any subsequent process of revision; this language will be described below. Selections from this document are made for both content and form, but are automatically post-edited—"*repaired*"— for form, style, and coherence.[5]

We regard this use of a master document as a new approach to natural language generation, in which generation from scratch is avoided. *Generation by selection and repair* uses a partially specified, pre-existing document as the starting point. In this way, we can finesse many of the intractable problems of generation, as we start from a document in which many of the decisions have already been predetermined: overall text organization, division of propositional content into sentences, choice of words, and lexical cohesive structure.

How repairs are made

The core of HealthDoc's tailoring facility is its *sentence planner*, which is presently under development for the main project. Because the bits and pieces of text selected from a master document might not necessarily be coherent or cohesive, the sentence planner performs complex linguistic repairs to restore coherence and cohesion to the "broken" selected document.

The sentence planner takes as input a set of sentence plans, written in Text Specification Language (described below), and performs the necessary repairs, with each type of repair performed by an independent repair module. The sentence planner is based on a blackboard architecture in which individual repair modules communicate and resolve their conflicts with one another. The architecture is described in greater detail by Hovy and Wanner (1996) and Wanner and Hovy (1996). Four repair modules are being built in the first phase of the main HealthDoc project: for discourse structuring, aggregation to remove redundancies, reference restoration using pronouns, and constituent re-ordering.

WebbeDoc: An application of the HealthDoc approach to the automated customization of Web documents

The idea behind WebbeDoc

We have now begun to apply the HealthDoc approach to designing Web-based document management systems that would produce textual materials tailored to the individual reader. We have developed a prototype of such a system, called WebbeDoc, that customizes a Web document describing the HealthDoc project.[6] WebbeDoc first displays only the most basic information about the project, using a bland style of presentation, and then allows the user to set various personal parameters and stylistic preferences. This

[5] It might be argued that a master document could just be a large set of simple blocks of text (or templates) to be included or excluded as appropriate for both content and form; the customized patient-information leaflets produced by Strecher and colleagues were done this way (Strecher *et al* 1994; Campbell *et al* 1994; Skinner, Strecher, and Hospers 1994). However, this approach requires that an extremely large number of bits and pieces of text be available: each fact expressed in each possible way. And the assembly of such bits and pieces suffers from the obvious problem that the resulting document might not be coherent or cohesive, or at the very least, not stylistically polished. It might be objected that the pieces of text could be carefully constructed so that all possible selections resulted in a well-formed document. Indeed, Strecher *et al* (1994) tried essentially this. However, they found it difficult to do even for their fairly simple document (Victor Strecher and Sarah Kobrin, p.c.); it would surely be very hard to achieve for complex documents unless the granularity were extremely coarse, thereby increasing the number of distinct elements required. In the limit, one would simply store a distinct document pre-written for every single combination of possibilities, a situation that we have already assumed to be impractical.

[6] A demonstration of WebbeDoc can be found at: http://logos.uwaterloo.ca/~healthdo/About/webbedoc.html.

causes WebbeDoc to "rewrite" its text and presentation style in accordance with the selected reader profile. Users can specify their role (e.g., computational linguist, funder, physician, layperson), age, and reading level, as well as stylistic preferences about the formality or "coolness" of the document to be generated. Examples of two different tailored versions of the opening section of the WebbeDoc page are shown in figure 1.

A document can be customized on all levels of linguistic structure: paragraph, sentence, and lexical choice, with each type of structure chosen for the appropriate pragmatic effect. WebbeDoc is doing more than blindly concatenating blocks of information content; it is selecting the most relevant pieces of text, with respect to both semantic content and pragmatic effect, so that they fit together in a coherent and cohesive manner. Our approach differs from that used by Strecher *et al* (1994) because, for WebbeDoc, the process of fitting together the bits and pieces of selected text is dependent on an explicit representation of the textual structure of the document. It is the existence of explicit rhetorical and other linguistic relations between the individual pieces of text that gives WebbeDoc the ability to produce coherent and polished tailored documents from the master document. (The nature of the document representation is described in detail in the section below entitled "Representing a master document".)

The document's structural representation contains not only linguistic information, but formatting specifications for each type of reader. So, in addition to textual customization, WebbeDoc can tailor the document's style and form of presentation; it can select, according to the user profile, the most appropriate artwork, font, colour, and general layout.

The present approach: "Generation by selection only"

WebbeDoc is a direct application of the ideas and mechanisms of HealthDoc; the project Web page that it customizes is itself a master document. In the first implementation of WebbeDoc, we have implemented a form of "generation by selection only": the structure of the master document is tightly constrained so that, after selection, no repairs will be needed to produce a coherent and stylistically adequate text.

An example of tailoring by WebbeDoc

In one section of WebbeDoc's HealthDoc master document, we use a sequence of three sentences to explain how an *authoring facility* must be provided to allow a writer to enter all the textual variations that make up a master document, together with the linguistic information that makes generation-by-selection-only work well.

This section begins with an introduction aimed at all audiences, differing only in the term used to described the authoring tool:

(1) An *authoring tool* allows a writer to enter all the different variations that make up a master document. This authoring facility *(for computational linguists)* / writer's workbench *(for laypersons who want a highly technical text)* / writer's assistant *(for laypersons who want a non-technical style)* has knowledge about language and document structure, so that it can guide the author into organizing a large number of snippets of text, all the different variations, into a single master document.

From there, the writer develops the theme through two subsequent sub-topics, on the need for a selection specification facility and on the kind of linguistic knowledge that the authoring tool must have. But each of these sub-topics has eight variations, made up from all the combinations of role (computational linguist, layperson), degree of technical detail (high, low), and degree of formality (formal, informal).[7] All the variations for the first sub-topic are shown in table 1.

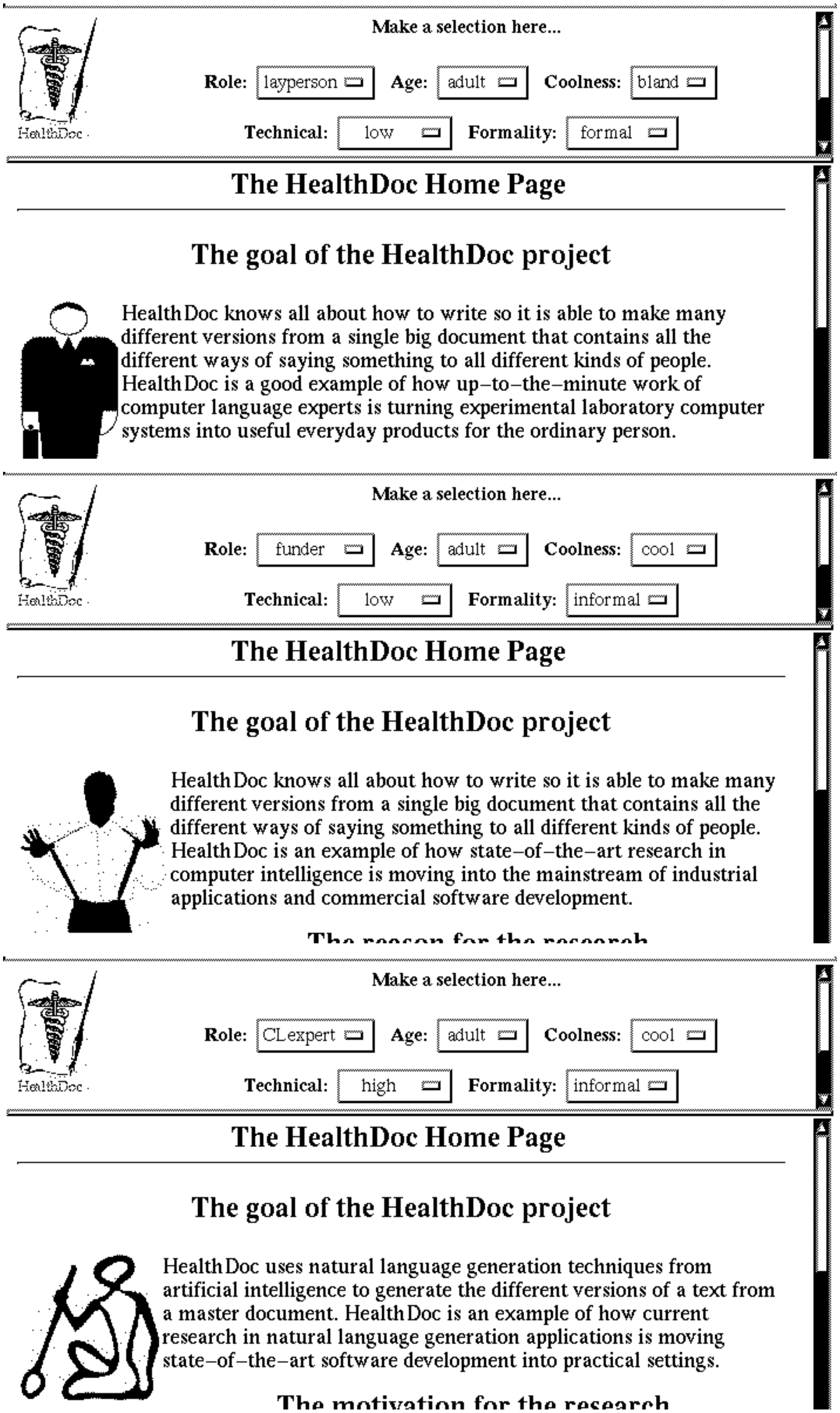
For the second sub-topic, which describes the kind of linguistic knowledge that WebbeDoc requires, there are similarly eight different variations, differing in features such as specialized language, number and complexity of sentences, and impersonal or deictic style.

From this sequence of three sentences, and their variations, WebbeDoc can produce eight different texts—no matter how the selections are made, the resulting paragraph will be coherent, with the same rhetorical structure, cohesive, with the appropriate discourse connectives, and stylistically appropriate, with the right level and choice of vocabulary.

For example, for a non-technical, informal layperson, the selected text would be:

> An *authoring tool* allows a writer to enter all the different variations that make up a master document. This writer's assistant has knowledge about language and document structure, so that it can guide the author into organizing a large number of snippets of text, all the different variations, into a single master document. Also, the writer's assistant helps you tag each snippet of text that you write with the patient features that you choose. And the assistant has expert knowledge about language. This knowledge gives it the ability to convert your original sentences into the system's special internal representation. This is called "Text Specification Language".

[7] Currently, WebbeDoc uses a total of five reader parameters: role (computational linguist, physician, funder, layperson); degree of technical detail (high, low); degree of formality (formal, informal); age (child, adult, senior), and degree of "coolness" (bland, cool). This gives a possible 96 ($4 \times 2 \times 2 \times 3 \times 2$) distinct combinations and different texts.



Make a selection here...

Role: layperson Age: adult Coolness: bland

Technical: low Formality: formal

The HealthDoc Home Page

The goal of the HealthDoc project

HealthDoc knows all about how to write so it is able to make many different versions from a single big document that contains all the different ways of saying something to all different kinds of people. HealthDoc is a good example of how up–to–the–minute work of computer language experts is turning experimental laboratory computer systems into useful everyday products for the ordinary person.

Make a selection here...

Role: funder Age: adult Coolness: cool

Technical: low Formality: informal

The HealthDoc Home Page

The goal of the HealthDoc project

HealthDoc knows all about how to write so it is able to make many different versions from a single big document that contains all the different ways of saying something to all different kinds of people. HealthDoc is an example of how state–of–the–art research in computer intelligence is moving into the mainstream of industrial applications and commercial software development.

The reason for the research

Make a selection here...

Role: CLexpert Age: adult Coolness: cool

Technical: high Formality: informal

The HealthDoc Home Page

The goal of the HealthDoc project

HealthDoc uses natural language generation techniques from artificial intelligence to generate the different versions of a text from a master document. HealthDoc is an example of how current research in natural language generation applications is moving state–of–the–art software development into practical settings.

The motivation for the research

Figure 1: Examples of different tailored versions from WebbeDoc

Highly technical text

	Formal	Informal
Linguist	In addition, the authoring facility will assist the writer to specify the selection criteria associated with each textual fragment.	In addition, the authoring facility assists you to specify the selection criteria that you wish to associate with each textual fragment that you write.
Layperson	In addition, the writer's workbench will assist the writer to specify the patient selection features associated with each text segment.	In addition, the writer's workbench assists you to specify the patient selection features that you wish to associate with each text segment that you write.

Non-technical text

	Formal	Informal
Linguist	Also, the authoring tool will help the writer to tag each fragment of text with a particular set of selection features.	Also, the authoring tool helps you tag each fragment of text that you write with the selection features that you choose.
Layperson	Also, the writer's assistant will help the writer to tag each snippet of text with the particular patient features.	Also, the writer's assistant helps you tag each snippet of text that you write with the patient features that you choose.

Table 1: The conceptual structure of a fragment of a master document

The key to WebbeDoc's ability to produce tailored documents by selection from a single master document is the manner of representation of the master document: a WebbeDoc master document has a well-defined structure of ordering relations, rhetorical relations, and other linguistic information, such as coreference links. In the first implementation, the master document was built manually according to our model of a master document, with additional structural constraints imposed so that piecewise selection and recombination would not create any infelicities such as abrupt changes of topic, unnecessary duplications of noun phrases, or unresolvable pronouns.

But to compose a master document of this style and internal complexity required the efforts of computational linguists, rhetoricians, and Web document designers; obviously this is not realistic for the average Web user! In a realistic and usable implementation, WebbeDoc would need an authoring tool and a sentence planner that could work in real-time to repair and polish the selected text—we can't expect the average Web document author to pre-compile all the possible combinations in advance. Therefore, to develop such a system, a number of research issues must be addressed, including representation of the master document; authoring and knowledge-based document management; and sentence planning for automated post-editing.

The next step: Generation of Web pages by selection and repair

Representing a master document

Text Specification Language, or TSL, is the language used to represent master documents in the parent HealthDoc system. We anticipate that WebbeDoc master documents will have a hybrid representation: part TSL (for the portions that will be subject to syntactic or stylistic repair), part "frozen" English text (for the portions that need never be revised). We have defined TSL to be an extension of the Sentence Plan Language (SPL) that is used by the Penman text generation system (Penman Natural Language Group 1989), whose KPML derivation (Bateman 1995) is used in HealthDoc. An SPL expression is an abstract specification of a sentence, which Penman can convert to the corresponding surface form. This permits expression of the content of the document. The basic SPL structures are annotated with selection and repair information to produce the corresponding TSL representation.

The format of the annotations for selection follows the structure of a user model, with annotations organized by personal and demographic category; for example:

```
:reader-role (layperson)
:reader-age (adult)
```

Other kinds of annotation for selection, such as reading level and preferred style of presentation, will, for the moment, be represented in a similar manner:

```
:technical-level (low)
:formality (informal)
```

The annotations can be included at any level in the SPL so that the system can make selections at any level of linguistic granularity. As stylistic and pragmatic customization becomes more complex, additional representations will probably be needed.

But this information isn't enough. We also require the internal discourse structure to be represented explicitly, to guide repairs to the structure of the text. Therefore, TSL contains several kinds of additional annotations, including *topic ordering information*, *coreference links*, and *rhetorical relations* between sentences. In addition to these current kinds of annotations, WebbeDoc's TSL will contain information on formatting and document presentation that would be marked up for inclusion according to specific user preferences.[8]

The model of a master document A master document is constructed according to a formal model; the model that we describe here is the most general, intended for the overall HealthDoc system, which does selection and repair of a master document. (The current version of WebbeDoc, which does generation by selection only, with no repairs involved, uses a more constrained model of a master document.)

We define the general model of a master document (MD) as follows:

- An MD has a coherent high-level communicative goal, such as to inform, to command, to persuade, to impress. For example, the purpose of the current WebbeDoc MD is to inform (and impress) the reader about the goals and technical achievements of the HealthDoc project.
- An MD has a coherent topic structure, with a division into topics, sub-topics, and so on. The smallest topic unit of an MD at the moment is a sub-sub-topic; however, we believe the form of the "smallest topic unit" will vary with the particular document.
- Each sub-topic corresponds to a section of the document that satisfies a more specific communicative goal, such as to justify or elaborate upon. For example, the first sub-topic of the sample WebbeDoc text given above elaborates on the selection specification facility in the authoring tool; the second sub-topic justifies the kind of specialized linguistic knowledge needed by the authoring tool. Essentially, a sub-topic is a semantically coherent piece of the document.
- Each sub-topic is a collection of *version sets* that are connected by ordering relations, rhetorical relations, coreference links, and formatting relations. A version set is a set of textual variations such that each variation fulfills the same communicative goal, but has a semantic content and pragmatic form tailored to a particular audience. Each variation in a version set is characterized by a logical condition and a semantically coherent piece of text. The logical condition uses terms that range over sets of mutually exclusive features.

  We interpret "mutual exclusion" to mean that the conditions assigned to the variations in a version set define a clean partition of the set, so that exactly one of the variations must be chosen.

  In the example given earlier, the first sub-topic is a singleton version set, sentence (1), while the second version set is made up of the eight sentences shown in table 1, and the third set also contains eight different sentence variations.
- *Ordering relations* may exist between the version sets that make up a sub-topic. These relations indicate the preferred order of the sequence of variations that have been selected to form the working document, and thereby specify the ordering of sub-topics prior to the invocation of the sentence planner.

  Preferred order can vary by reader. For example, the author of the WebbeDoc MD might decide that for computational linguists, the sub-topic about the authoring tool's linguistic intelligence should precede the sub-topic on the selection criteria, but for laypersons, the reverse order would be preferable.
- *Rhetorical relations* may exist between the version sets that make up a sub-topic. The rhetorical relations that we are currently using are taken from Rhetorical Structure Theory (RST) (Mann and Thompson 1988). In the current version of WebbeDoc, the *same* rhetorical relation must exist between any two members of adjacent version sets.

  In the example we have been using, the rhetorical relations are as follows:

  Any choice from the second version set (shown in table 1) *elaborates upon* sentence (1) (the first version set).

  Any choice from the third version set *justifies* any earlier choice from the second version set.

[8] Indeed, we anticipate that there will be a distinct "repair" module for document formatting in the sentence planner used with WebbeDoc.

- *Coreference links* may be defined between any two version sets. In our example, the following terms, used in the first and second version sets, are coreferential: *authoring tool*, *authoring facility*, *writer's workbench*, *writer's assistant*, and *it*. (The first two terms are also *near-synonyms*.)

- *Formatting information* may be defined at each topic and sub-topic level. Formatting information may also be defined between and within version sets, including illustrations, choice of colour, design of layout, and so on.

Authoring a master document

WebbeDoc master documents may be based on the natural-language text of pre-existing material, or they may be created from scratch (or some combination of the two). Either alternative requires the involvement of a human.

The author of a WebbeDoc master document would normally be a professional technical writer or Web-document designer, who will need to understand the nature of customized and customizable texts, but who should not be assumed to have any special knowledge or understanding of TSL or the innards of WebbeDoc. The authoring tool, therefore, should be no more difficult for the author to use than, say, the more-sophisticated features of a typical word processor. The text is therefore written in English, and will be translated to TSL by the authoring tool. (The English source text is retained in the TSL for use in subsequent authoring sessions—for example, if the document is updated or amended.)

It is the writer's job to decide upon the basic elements of the text, the formatting, ordering, rhetorical, and coreferential links between them, and the conditions under which each element should be included in the output. The elements of the text are then typed into the authoring tool in English, and are marked up by the writer with conditions for inclusion, links for cohesion and coreference, and annotations for ordering and formatting of the document layout. An example of the authoring tool's main interface (depicting part of the sample WebbeDoc master document described earlier) is given in figure 2.

The tool then translates the text into TSL. This is essentially a process of semi-automated parsing, so that whenever an ambiguity cannot be resolved, the writer is queried in an easy-to-understand form. The design and development of the authoring tool and its user interface is part of the current phase of the overall HealthDoc project (fall 1996 to spring 1997). The user interface is being developed by Parsons (1997), while Banks (1997) is implementing the English-to-TSL conversion (for more details on the underlying model of conversion, see DiMarco and Banks (1997)).

Functions of sentence planning and automated post-editing

In general, selecting material from pre-existing text and then editing it to recover coherence and cohesion can involve a wide range of problems in various aspects of sentence planning. For example, both syntactic and semantic aggregation may be needed, as well as chunking of whole and partial propositions. Pronouns and other forms of reference need to be chosen. And, of course, aggregation and sentence restructuring will affect the rhetorical relations between the elements of the text.

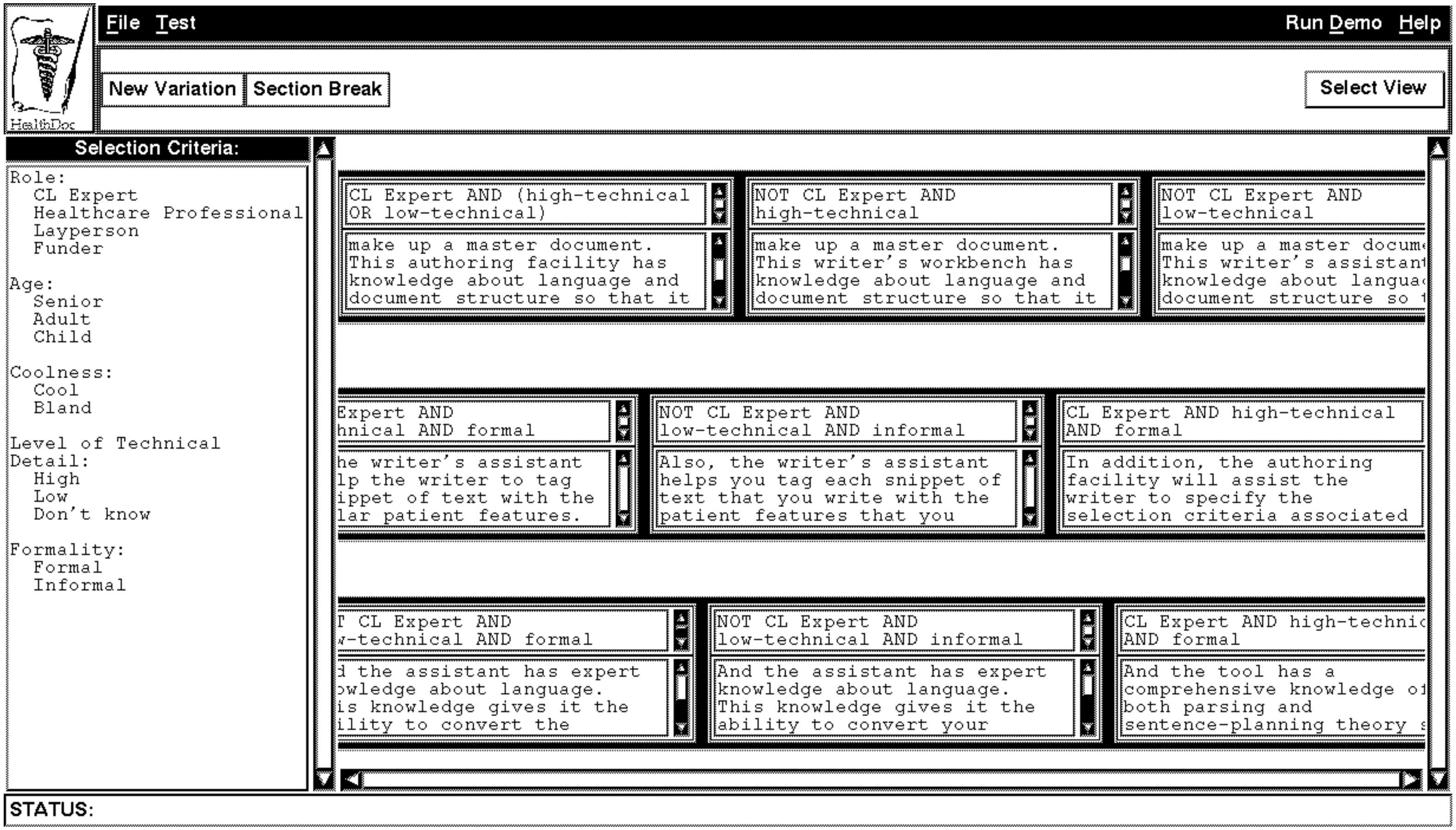
Our current work is focusing on the development of two key modules of the sentence planner: for discourse structuring and for aggregation. It is unlikely that every ordering of the blocks of text that are organized into a master document will produced a coherent sequence of selected pieces of text. To ensure that any resulting document makes sense, the discourse module uses the rhetorical relations that hold among the textual units to produce a sequence that is most likely to be coherent. In later work, an additional module will be built to determine the linguistic phrasing of the discourse relation.

The aggregation module eliminates redundancy in TSL expressions by grouping together entities that are arguments of the same rhetorical relation, verbal process, etc. Each aggregation rule recognizes an exact match of some portions of two input TSL expressions and returns a single, fused, expression. The actions of the aggregation module will generally affect the resulting syntactic structure.

A critical problem is the distribution of repair tasks among the planning modules, as there are often strong interactions. The responsibilities of each module and the overlaps between them are an area of on-going research for our sentence-planning group.

Conclusion

The HealthDoc project and its WebbeDoc offspring aim to provide a comprehensive approach to the automated tailoring of both paper documents and Web-based materials. We incorporate explicit user modeling as a basis for the document tailoring, and we take into account user information ranging from simple demographic data to complex pragmatic preferences. We have developed a model of language generation, "generation by selection and repair", that relies on a "master-document" representation that predetermines the basic form and content of a text and yet is amenable to editing and revision for customization. The WebbeDoc project aims to provide useful techniques for natural language applications on the Web and to address a number of important issues for research in more-general systems for language generation.



Role:
  CL Expert
  Healthcare Professional
  Layperson
  Funder

Age:
  Senior
  Adult
  Child

Coolness:
  Cool
  Bland

Level of Technical
Detail:
  High
  Low
  Don't know

Formality:
  Formal
  Informal

CL Expert AND (high-technical
OR low-technical)

make up a master document.
This authoring facility has
knowledge about language and
document structure so that it

NOT CL Expert AND
high-technical

make up a master document.
This writer's workbench has
knowledge about language and
document structure so that it

NOT CL Expert AND
low-technical

make up a master docum
This writer's assistant
knowledge about langua
document structure so t

Expert AND
hnical AND formal

he writer's assistant
lp the writer to tag
ippet of text with the
lar patient features.

NOT CL Expert AND
low-technical AND informal

Also, the writer's assistant
helps you tag each snippet of
text that you write with the
patient features that you

CL Expert AND high-technical
AND formal

In addition, the authoring
facility will assist the
writer to specify the
selection criteria associated

T CL Expert AND
v-technical AND formal

d the assistant has expert
owledge about language.
is knowledge gives it the
ility to convert the

NOT CL Expert AND
low-technical AND informal

And the assistant has expert
knowledge about language.
This knowledge gives it the
ability to convert your

CL Expert AND high-technic
AND formal

And the tool has a
comprehensive knowledge of
both parsing and
sentence-planning theory s

Figure 2: The main interface of the authoring tool

Acknowledgements

The HealthDoc Project is supported by a grant from Technology Ontario, administered by the Information Technology Research Centre. Vic DiCiccio was instrumental in helping us to obtain the grant, and has been invaluable in subsequent administration. Substantial portions of the sections of this paper that described the HealthDoc project and authoring of master documents were written by Graeme Hirst; they are used here with his permission. Some material in the section on the functions of sentence planning was written by Eduard Hovy; it is used here with his permission. We are grateful to Graeme Hirst and Eduard Hovy for many helpful comments on this research and this paper. The other members of the HealthDoc Project have also contributed to the work described here, especially Daniel Marcu, Kim Parsons, and Phil Edmonds. Jonathan Dursi kindly provided help with some of the LaTeXdevelopment.

References

Banks, Steven (1997). Master's thesis. Department of Computer Science, University of Waterloo, expected Spring 1997.

Bateman, John Arnold (1995). "KPML: The KOMET–Penman multilingual linguistic resource development environment." *Proceedings, 5th European Workshop in Natural Language Generation*, Leiden, May 1995, 219–222.

Campbell, Marci Kramish; DeVellis, Brenda M.; Strecher, Victor J.; Ammerman, Alice S.; DeVellis, Robert F.; and Sandler, Robert S. (1994). "Improving dietary behavior: The effectiveness of tailored messages in primary care settings." *American Journal of Public Health*, **84**(5), May 1994, 783–787.

DiMarco, Chrysanne (1990). *Computational stylistics for natural language translation.* PhD thesis, Department of Computer Science, University of Toronto, 1990. Published as technical report CSRI-239.

DiMarco, Chrysanne and Banks, Steven (1997). "Using subsumption classification on a stylistic hierarchy as the basis of a multi-stage conversion of natural language text to sentence plans." In preparation.

DiMarco, Chrysanne; Hirst, Graeme; and Stede, Manfred (1993). "The semantic and stylistic differentiation of synonyms and near-synonyms." *Proceedings, AAAI Spring Symposium on Building Lexicons for Machine Translation*, Stanford, March 1993, 114–121.

DiMarco, Chrysanne and Hirst, Graeme (1993a). "A computational theory of goal-directed style in syntax." *Computational Linguistics*, **19**(3), September 1993, 451–499.

DiMarco, Chrysanne and Hirst, Graeme (1993b). "Usage notes as the basis for a representation of near-synonymy for lexical choice." *Proceedings, Ninth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research*, Oxford, September 1993, 33–43.

DiMarco, Chrysanne; Hirst, Graeme; Wanner, Leo; and Wilkinson, John (1995). "HealthDoc: Customizing patient information and health education by medical condition and personal characteristics." *Workshop on Artificial Intelligence in Patient Education*, Glasgow, August 1995.

Green, Stephen (1992). "A functional theory of style for natural language generation." Master's thesis, Department of Computer Science, University of Waterloo, 1993.

Green, Stephen J. and DiMarco, Chrysanne (1996). "Stylistic decision-making in natural language generation." In *Trends in natural language generation: An artificial intelligence perspective.* Giovanni Adorni and Michael Zock (eds.). Springer-Verlag Lecture Notes in Artificial Intelligence (a subseries of Lecture Notes in Computer Science) number 1036, 1996.

Hirst, Graeme (1995). "Near-synonymy and the structure of lexical knowledge." *Working notes, AAAI Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, Stanford University, March 1995, 51–56.

Hovy, Eduard and Wanner, Leo (1996). "Managing sentence planning requirements." *Proceedings, ECAI-96 Workshop on Gaps and Bridges: New Directions in Planning and Natural Language Generation*, Budapest, August 1996.

Hoyt, Pat (1993). *A goal-directed functionally-based stylistic analyzer.* Master's thesis, Department of Computer Science, University of Waterloo, 1993.

Hoyt, Pat and DiMarco, Chrysanne (1994). "A goal-directed multi-level stylistic analyzer." *Proceedings, 10th Canadian Conference on Artificial Intelligence*, Banff, May 1994, 23–30.

Mann, William C. and Thompson, Sandra A. (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization." *Text*, **8**(3), 1988, 243–281.

Parsons, Kimberley J. (1997). Master's thesis, Department of Computer Science, University of Waterloo, expected Spring 1997.

Penman Natural Language Group (1989). "The Penman primer", "The Penman user guide", and "The Penman reference manual." Information Sciences Institute, University of Southern California.

Skinner, Celette Sugg; Strecher, Victor J.; and Hospers, Harm (1994). "Physicians' recommendations for mammography: Do tailored messages make a difference?" *American Journal of Public Health*, **84**(1), January 1994, 43–49.

Strecher, Victor J.; Kreuter, Matthew; Den Boer, Dirk-Jan; Kobrin, Sarah; Hospers, Harm J; and Skinner Celette S. (1994). "The effects of computer-tailored smoking cessation messages in family practice settings." *The Journal of Family Practice*, **39**(3), September 1994, 262–270.

Wanner, Leo and Hovy, Eduard (1996). "The HealthDoc sentence planner." *Proceedings of the Eighth International Workshop on Natural Language Generation*, Brighton, UK, June 1996.