Repairing Usability Problems Identified by the Cognitive Walkthrough for the Web

Marilyn Hughes Blackmon[†], **Muneo Kitajima**[‡] **and Peter G. Polson**[†]

[†]Institute of Cognitive Science
University of Colorado at Boulder
Boulder, Colorado 80309-0344 USA
+1 303 492 5063
{blackmon, ppolson}@psych.colorado.edu

[‡]National Institute of
Advanced Industrial Science and Technology (AIST)
1-1-1, Higashi, Tsukuba, Ibaraki 305-8566 Japan
+81 298 61 6650
kitajima@ni.aist.go.jp

ABSTRACT

Methods for identifying usability problems in web page designs should ideally also provide practical methods for repairing the problems found. Blackmon et al. [2] proved the usefulness of the Cognitive Walkthrough for the Web (CWW) for identifying three types of problems that interfere with users' navigation and information search tasks. Extending that work, this paper reports a series of two experiments that develop and prove the effectiveness of both full-scale and quick-fix CWW repair methods. CWW repairs, like CWW problem identification, use Latent Semantic Analysis (LSA) to objectively estimate the degree of semantic similarity (information scent) between representative user goal statements (100-200 words) and heading/link texts on each web page. In addition to proving the effectiveness of CWW repairs, the experiments reported here replicate CWW predictions that users will face serious difficulties if web developers fail to repair the usability problems that CWW identifies in web page designs [2].

INTRODUCTION

Previously reported research used the Cognitive Walkthrough for the Web (CWW) usability inspection method [2] to identify usability problems in web pages and then confirmed the accuracy of CWW predictions by testing the web pages with actual users trying to accomplish specific goals on particular web pages with many links. In essence, CWW predictions sorted a representative set of user goals for a specific web page into two main categories. The first category was user goals that CWW predicted users could accomplish with no serious navigation problems and the second, more complex category was user goals that CWW predicted would encounter serious problems. The second category, goals with serious problems, was subdivided into three types of usability problems: *unfamiliar* heading/link texts, *confusable* heading/link texts, and *goal-specific competing* heading/link texts. Averaging over three experiments, users first click was on the correct link only 37% of the time for user goals with unfamiliar link texts, the problem type with the lowest success rate, but rose to 74% correct on first click for user goals for which CWW predicted no problems.

The two experiments reported in this paper move beyond earlier research by (1) defining CWW repair methods, (2) testing whether CWW repairs actually work on a laboratory simulation of an actual online encyclopedia, and (3) testing short-cut repairs to see if they capture most of the performance gains of full-scale repairs with less effort from web page designers. For both experiments we used CWW to identify usability problems in specific web pages, then applied CWW repairs to these web pages, and finally tested whether users can accomplish the same set of goals more successfully on the repaired page than on the original page.

The next section presents CWW repairs in the context of explaining how and why CWW can produce accurate predictions for a diverse range of user groups by relying on Latent Semantic Analysis (LSA) and on the CoLiDeS model of user website behavior. Then the paper presents the empirical work demonstrating the benefits of using CWW to repair usability problems, one section for each of the two experiments. The final section enlarges the discussion to related research and draws conclusions.

REVIEW OF CWW [2]

The Cognitive Walkthrough for the Web (CWW) is a theoretically-based usability inspection method [12] for detecting and correcting design errors that interfere with finding information on a website [2]. CWW, like the original Cognitive Walkthrough [15], simulates step-by-step user behavior for a given task and assumes that users perform goal-driven exploration. But CWW is specially tailored to simulate users navigating a website and better fits a realistic website design process [2], considering three features specific to website design. First, CWW uses realistic narrative descriptions of user goals that incorporate rich information about users' understanding of their tasks and underlying motivation.

Second, CWW assumes that generating an action on a web page (e.g., clicking a link, button, or other widget) is a two step process. Step one is an attention process that parses a web page into subregions and attends to the subregion of the page that is semantically most similar to the user goal.

CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA.

Step two is an action selection process that selects and acts on a widget from the attended-to subregion, the widget semantically most similar to the user goal.

This two-step CWW web navigation mechanism is derived from a theory of the cognitive processes that control goal driven exploration, CoLiDeS [9]. CoLiDeS, an acronym for Comprehension-based Linked model of Deliberate Search, extends a series of earlier models [8] of performing by exploration based on Kintsch's [7] construction-integration theory of text comprehension and problem solving processes. CoLiDeS is part of a broad consensus among theorists and website usability experts [3,4,5,6,11,14] that problem solving processes, guided by users' goals and *information scent*, drive users' information-seeking or search behaviors when exploring a new website or carrying out a novel task on a familiar website.

Third, the CWW evaluation process can balance competing constraints by working on one web page at a time in relation to a whole set of representative user goals. The CWW evaluation process can start with a detailed description of the home page and a rough outline of its immediate successor pages. CWW can then be applied repeatedly to incrementally design and evaluate each successor page down through the hierarchy.

Design Questions

CWW [2] identifies usability problems by simulating step-by-step user behavior for a given task using a prototype interface, and by having the design team answer the following four questions at each simulated step: *Q1) Will the users try to achieve the right effect? Q2) Will the correct action be made sufficiently evident to the user? Q3a) Will the user connect the correct subregion of the page with the goal using heading information and her understanding of the site's page layout conventions? Q3b) Will the user connect the goal with the correct widget in the attended-to subregion of the page using link labels and other kinds of descriptive information? Q4) Will the user interpret the system's response to the chosen action correctly?* (quoted from [2], p. 463).

Questions Q1, Q2, and Q4 are retained from the original Cognitive Walkthrough, whereas Q3a and Q3b are specifically adapted to the users' web navigation process specified by CoLiDeS. Q3a corresponds to the attention process, and Q3b corresponds to the action selection process in the CoLiDeS model.

How CWW Employs Latent Semantic Analysis (LSA)

CWW answers the design questions Q3a and Q3b by applying Latent Semantic Analysis (LSA) [10], and it answers question Q1 by using goal statements long enough for accurate LSA predictions (100-200 words). LSA is a machine learning technique that builds a semantic space representing a given user population's understanding of words, short texts (e.g., sentences, links), and whole texts. The meaning of a word, link, sentence or any text is represented as a vector in a high dimensional space, typically with about 300 dimensions. LSA generates the space by applying singular value decomposition, a mathematical procedure similar to factor analysis, to a huge terms-by-documents co-occurrence matrix.

The degree of semantic relatedness or similarity between any pair of texts, such as the description of a user's goal and a link label on a web page, is measured by the cosine value between the corresponding two vectors. Cosines are analogous to correlations. Each cosine value lies between +1 (identical) and -1 (opposite). Near-zero values represent two unrelated texts. CWW uses LSA to compute the semantic similarities between user goals and subregion heading and link labels or descriptions of other widgets. CWW predicts that users attend to the subregion with the highest goal-heading (or goal-subregion) cosine value and the link or widget in the attended-to subregion with the highest goal-link (or goal-widget) cosine value.

Another important measure provided by LSA is term vector length, a measure that is correlated with word frequency, and that estimates how much knowledge about a word or phrase is embedded in the designated LSA semantic space (e.g., the space for grade 9 general reading knowledge). A semantic space representing a given user population is generated from a large corpus of written materials (including books, magazines, and newspaper articles) read by typical members of that population. Words not included in the corpus are not represented in the semantic space. Words with low frequency in the corpus (e.g., specialized technical or scientific terms) have short term vector lengths. When a heading/link has a short term vector length, CWW predicts that users modeled by the semantic space will perceive it to be relatively meaningless, reducing the probability that users will attend to or click on them.

CWW Repairs for Usability Problems CWW Identifies

The current iteration of CWW identifies and repairs three classes of usability problems: confusable heading/link, unfamiliar heading/link, and goal-specific competing heading/link. CWW detects usability problems by consulting measures LSA provides to the analyst: term vector lengths, semantic similarity between pairs of headings/links, and semantic similarity between representations of users' goals and headings/links. Specific detection criteria follow, using examples from the experiments reported in this paper.

Unfamiliar Headings/Links and Confusable Headings/Links

Unfamiliar and confusable problems can cause problems for any user goal, so it is strategic to repair these two types of problems first – before using CWW to identify and repair goal-specific competing headings/links.

Any pair of headings or any pair of links yielding a cosine of 0.6 or more in the LSA analysis is tagged as confusable. For example, a confusable pair of links on the Unrepaired Humanities web page is the pair United States History and People in United States History, a pair with a cosine of 0.97. Our repair changed the second link label to Leaders in American History, reducing its similarity with United States History to a cosine of 0.23. In addition to reducing the similarity between the two links, the repair must improve users' accuracy in predicting what they will find when they click that link. The repaired link label Leaders in American History is a better topic label than People in United States History, because the list of articles users see when they click the link contains, without exception, articles about famous individual American leaders, e.g., Abraham Lincoln. None are articles about a people, e.g., Cheyenne, Pueblo, or Amish. Finally, the repair needs to result in a high heading-link similarity with the heading under which the link is nested and low heading-link similarities with all other headings on the same web page – a criterion that can necessitate regrouping links.

A heading/link can cause serious trouble if it is unfamiliar (the user does not know what the heading/link means and/or has little background knowledge about the topic). Low word frequency is a good index of a single unfamiliar word, but a term vector length in LSA is a versatile measure that covers either single or multiple words. We thus define a heading/link as unfamiliar if it has a term vector length of less than 0.8 for the two most meaningful words. For example, The Occult has a term vector length of 0.08. CWW predicts that first-year college students will know little about The Occult. Even if user goal has its highest cosine with The Occult, the user can perceive the similarity only if the link is meaningful to the user. Basic repair strategies include replacing or elaborating unfamiliar heading/link texts with higher frequency words. In this experiment, for example, the repair used for the unfamiliar link Paleontology (0.06 term vector length) was to rename it Paleontology & Fossils (1.55 term vector length). Similarly, we repaired The Occult by renaming it Magic, Supernatural, & Spirits (best-available repair raises term vector to 0.76 for magic + spirits, 0.88 for all three words).

Goal-specific Competing Heading/Link

Unfamiliar or confusable headings or labels are bad whatever the user's goal might be, but some problems emerge only for some goals. For example, two headings may not be very similar to one another, but may both be equally similar to a possible goal. If the similarity of a heading to the goal is equal to or greater than the similarity of the correct heading to the goal, the analyst marks the intruder as a goal-specific competing heading unless the analyst judges the similarity a false alarm – a heading not likely to attract users' attention for accomplishing that goal.

In this experiment, the web designer designated the link Industry Mining & Fuels, nested under the heading **Physical Science & Technology**, as the correct link to find an article about Fisheries. CWW predicts that people looking for Fisheries would be more likely to focus on the heading **Life Science** than on **Physical Science & Technology**, so **Life Science** is a goal-specific competing heading for Fisheries. We repaired this problem by making it possible to find Fisheries by focusing on the competing heading **Life Science** and then clicking at least one highly similar link (we chose the link Fish) nested under **Life Science**. The repair also preserved the option of finding Fisheries by clicking the "correct" link Industry Mining & Fuels under the heading **Physical Science & Technology**.

The CWW standard for a goal-specific competing link has three criteria: (1) the competing link label must be under the same heading as the correct link, (2) the competing link label must have a goal-link cosine value that equals at least 80% of the goal-link cosine for the correct link label, and (3) not be judged by the analyst as a *false alarm*, i.e., a link that real users would probably not select. For example, CWW predicted that users' attention would be strongly drawn to the heading **Social Science** to find an article about Child Labor. Although there are no competing headings for Child Labor CWW predicted competing links nested under **Social Science**. CWW predicted that users would be more likely to select Sociology & Social Reform or Law than the "correct" link, Economics and Business. We repaired the competing links problem by making all competing and correct links (Sociology & Social Reform, Law, and Economics and Business) lead to the Child Labor article.

In sum, *competing* heading/link problems are repaired by making it possible for users to find things in more than one way, putting links to the item all the places where CWW predicts users are most likely to look.

A more extensive description for how to identify and repair each of the usability problems identified by CWW is provided in a CWW tutorial on the website where CWW analyses can be done [http://AutoCWW.colorado.edu].

EXPERIMENT 1: TEST FULL-SCALE REPAIRS

The goal of Experiment 1 was to use CWW to identify and repair usability problems on web pages from an actual website, and then to verify that fully repaired versions of these web pages were significantly more usable.

Method

Materials

We constructed an experimental website to parallel the hierarchical structure of a widely used online encyclopedia as closely as possible, ensuring that human performance on our laboratory simulation would approximate performance on the actual website. Our simulated website used two top-level web pages: (1) a Humanities main page with 40 topic links nested under five headings, and (2) a Sciences main page with 34 topic links nested under three headings. One important difference is that the simulated website combined the top two levels of the actual website into a single level, a design choice that traded breadth for depth and probably improved performance on the simulated website compared to the actual website [11]. On the actual website users must select a category on the top-level categories page and drill down to a list of topics for that particular category on level two. Our simulation presented a rectangular matrix that used category labels as headings for cells in the matrix and nested the topics under the headings in exactly the same order as they are listed on the actual website. The simulated and actual websites used exactly the same category/heading texts, topic texts, and article titles. At level three the actual website has a web page for each separate topic with an alphabetized list of articles classified under that topic, and users click an article-title link in the list to reach the encyclopedia articles (at level four). Users follow exactly the same procedures on our simulated online encyclopedia, except that our simulation contains a small fraction of the encyclopedia articles available in the actual website, and alphabetized lists of articles were correspondingly shorter.

The experiment asked participants to find a series of 32 target articles, such as an article on Cotton or on Courtly Love. Regardless of experimental condition, participants could always find each target article under the same topic under which it is classified in the actual website. For the Unrepaired condition of Experiment 1, however, *only one* topic link leads to the target article, since people using the actual website must usually click a single "correct" topic link to navigate to that particular article in the actual website. Encyclopedia classification experts determined which topic is the "correct" topic(s) for each article.

The Humanities and Sciences web pages used for the Unrepaired condition in Experiment 1 were the web pages that simulate the actual website. In addition, two fully repaired pages were created for the RepairedExamples condition of Experiment 1. The RepairedExamples Humanities and Sciences web pages retained the same topics and number of topic links (34 links for Sciences, 40 for Humanities), but unfamiliar and confusable link texts for these topics were repaired as described in the previous section. In addition, links were regrouped and/or given new heading labels until each link on the page was more similar to the heading under which it was nested than to any other heading on the page. To produce more cohesive link groupings the number of headings was increased from five to six for the Humanities page and from three to seven for the Sciences page. Finally, we used LSA to evaluate all examples from the alphabetically arranged list of encyclopedia articles and selected the 5-10 best examples to elaborate and clarify each topic link. The examples selected met two constraints: (1) article titles were highly similar to the topic link text, and (2) article titles were highly familiar as measured by term vector length in the semantic space. You can see the web pages used in the experiment and/or try doing the experiments yourself at psych.colorado.edu/~blackmon/Expt011015Home.html.

The 32 target articles were divided into two sets: 16 Humanities items and 16 Sciences items. Each of these content subsets was subdivided into four problem types: Unfamiliar, Competing Headings, Competing Links, and Two or More Problems. The Sciences and Humanities Unrepaired pages each had four items for each problem type. LSA was used to select items predicted to have these problems on either the Humanities or Sciences web page.

For the Unrepaired Humanities and Sciences web pages all 32 items were problem items. For the RepairedExamples Humanities and Sciences web pages all 32 items were repaired problems. As a result people were expected to find each of the 32 target articles faster and more easily on the web page for the RepairedExamples condition than on the corresponding web page for the Unrepaired condition.

Both Unrepaired and RepairedExamples web pages were simplified compared to actual web pages, eliminating the usual features of main pages in websites (site-wide logo, navigation bars, news/features links and advertisements). The content area of the experimental web pages was pared down to a rectangular table, one or two rows high by three or more columns wide. Each table cell forms one subregion, and heading texts in large, bold, contrasting-color font top each of these subregions. For example, the Unrepaired Sciences web page (34 links nested under three headings) had a single row divided into three columns. Each column contained a vertically arranged list of topic links topped by one of the three heading texts: **Physical Science & Technology** (13 links), **Life Science** (12 links), and **Social Science** (9 links).

The top part of the web page (the part of the web page normally occupied by the logo) had an attention-grabbing background color of bright yellow and contained a description of the target article. The text describing the goal was between 100 and 200 words and was a faithful summary of the online encyclopedia article. When creating a summary, LSA was used to assess the degree of semantic similarity between the summary and the article, and the minimum LSA similarity measure (cosine) between the article and its summary was set at 0.8, a very high degree of similarity for any pair of texts. The purpose was to ensure that experimental participants had an accurate conception of the article they were looking for, maximizing the information scent for the various target articles.

Experimental Participants

The participants in Experiment 1 were 119 undergraduates doing the experiment to complete a course requirement for the Introduction to Psychology course in which they were enrolled for the semester. Experimental participants were randomly assigned to one of four experimental groups, a 2 X 2 design of two groups (A1 and B1) times two presentation orders for the goals (Original and Reverse): A1Original, A1Reverse, B1Original, and B1Reverse.

Procedure

All participants searched for a total of 40 target articles, presented by the website in a fixed order. Participants were given 150 seconds to find an article, and anyone who failed to find the encyclopedia article before the time expired saw a "Time expired" screen and clicked a link to move on to the next goal in the sequence. Participants completed the experiment in 35-55 minutes.

All participants first searched for eight practice target articles on an online encyclopedia web page with 51 links nested under 8 headings, covering topics in both sciences and humanities. After completing the practice items, the website presented two sets of 16 target articles. The A1Original and B1Original groups did the 32 target articles in sequential order, 9-24 followed by 25-40. The A1Reverse and B1Reverse groups did the two sets in reverse order, articles 25-40 and then articles 9-24. For each target article, therefore, half the data come from participants who did the goal in the first set of 16 items and the other half from participants who did the goal in the second set of 16 items. This design made it possible to determine if performance on the 32 target articles differed if done in the second half of the experiment than in the first half (due to extra experience that improved performance or to fatigue that lowered performance).

All four experimental groups alternated between a Humanities web page on odd-numbered target articles and a Sciences web page on even-numbered target articles. For the A1Original and A1Reverse groups the Humanities web page was in the RepairedExamples condition and the Sciences page was in the Unrepaired condition. Just the opposite was true for the B1Original and B1Reverse group: the Humanities web page was in the Unrepaired condition and the Sciences page in the RepairedExamples condition. This design allowed each participant to alternate between harder items (Unrepaired condition) and easier items (RepairedExamples condition), as well as making it possible to compare performance for the two conditions on both Humanities and Sciences pages.

Results

The left half of Figure 1 summarizes the results of Experiment 1 averaged across Humanities and Sciences pages. The performance measure is the number of mean clicks experimental participants made on the main web page in order to find the target article. Clicking a link on the main page produced an intermediate topic page with an alphabetical list of articles belonging under that topic. Subjects rapidly scanned the list and clicked the target article, if present in the list, or else clicked the back button. Clicks on intermediate pages were virtually error-free, so we only tallied clicks on main pages. The data were analyzed using a Repeated Measures ANOVA for a mixed design of between- and within-group variables. We averaged the data from the four individual target items for each problem type nested under Humanities or Sciences, giving eight means per participant. The statistics presented below are for these means collapsed over individual items.

The between-group variable for Condition (Unrepaired vs. RepairedExamples) was significant for both the Humanities web page, *F* (1, 110) = 157.499, *p*<.0001, and the Sciences web page, *F* (1, 111) = 107.574, *p*<.0001. The mean on the Humanities page was 4.51 clicks per item for Unrepaired compared to 1.97 for RepairedExamples, and the mean on the Sciences page were 3.24 clicks per item for Unrepaired compared to 1.68 for RepairedExamples. Averaging across the two web pages, the ratio was 2:1.

There was a significant difference for the within-group variable of Problem Type (Competing Links, Competing Headings, Unfamiliar, TwoOrMoreProblems) for both the Humanities web page, *F* (3, 330) = 10.92, *p*<.0001, and the Sciences web page, *F* (3, 333) = 19.846, *p*<.0001. As shown in the left half of Figure 1, TwoOrMoreProblems was the most difficult and Competing Links the least difficult of the problem types (averaged across conditions).

The presentation order for the goals (Original vs. Reverse), a between-group variable, did not significantly affect performance for either the Humanities or Sciences web pages, and none of its interactions were significant.

There was only one significant interaction, the interaction between Condition and Problem Type, and it was significant for both the Humanities web page, *F* (3, 330) = 7.96, *p*<.0001, and the Sciences web page, *F* (3, 330) = 4.65, *p*<.005. This interaction results from the fact that the performance gain due to repairs was greater for Competing Links and Competing Headings problem types than for the Unfamiliar type, as shown in the left half of Figure 1.

Discussion of Experiment 1

The strategy used in designing the web pages for the Repaired conditions was to apply every technique possible to improve performance. Unfamiliar headings/links were replaced with paraphrases made up of high-frequency words familiar to users. We repaired and juxtaposed confusable heading/link texts to facilitate discrimination and then clarified each link with 5-10 examples. We regrouped links and altered heading texts to ensure that each link was highly similar only to the heading under which it was nested. Then we repaired goal-specific competing problems to ensure users could access each target article via all of the most similar headings and links.

The repaired pages produced dramatic improvements in performance, an almost 2:1 reduction in the mean clicks to solution. The repair process, however, entailed two undesirable trade-offs. First, it was very time-consuming to repair all confusable and unfamiliar problems and to select 5-10 relevant examples to elaborate each link, raising the development cost of real-world websites. Second, elaborating the links added text to repaired pages, making

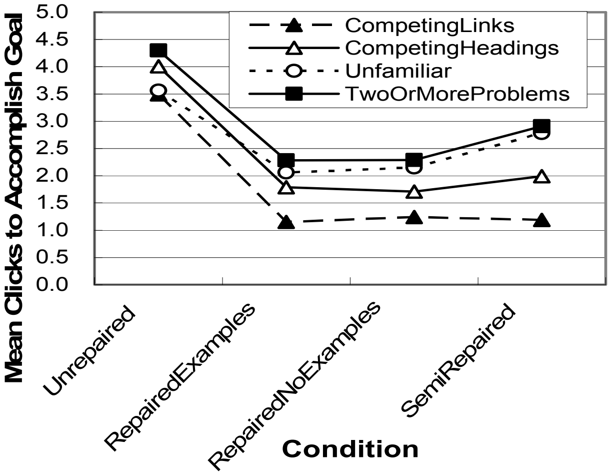

Figure 1. Summary of Experiments 1 and 2: Mean clicks on main page to accomplish goal

them possibly harder to scan than unrepaired counterparts.

EXPERIMENT 2: TEST SIMPLIFIED CWW REPAIRS

Experiment 2 tested two quick-fix repair processes that eliminated the undesirable trade-offs of full-scale repairs. Procedures and materials were identical to Experiment 1 except for the new versions of Humanities and Sciences home pages generated by simplified repair methods. The RepairedNoExamples condition just eliminated the topic-elaborating examples from the Sciences and Humanities pages used in the RepairedExamples condition of Experiment 1. For the SemiRepaired condition we repaired all goal-specific competing heading/link problems but made no modifications to the heading/link texts on the pages used in the Unrepaired conditions. As a consequence, simplified repairs in the SemiRepaired condition ignored making repairs for unfamiliar and confusable problems.

Method

Experimental Participants

The 85 participants in Experiment 2 were undergraduates enrolled in Introduction to Psychology, as in Experiment 1, and were randomly assigned to one of four groups in a 2 groups (A2, B2) X 2 orders (Original, Reverse) design: A2Original, A2Reverse, B2Original, and B2Reverse.

Materials

The crucial change between Experiment 1 and Experiment 2 was in the Humanities and Sciences web pages presented to participants. SemiRepaired Humanities and Sciences web pages were substituted for their Unrepaired equivalents in Experiment 1. SemiRepaired pages had exactly the same surface appearance as Unrepaired pages, but hidden beneath the surface of SemiRepaired pages were repairs for all the goal-specific competing heading/link problems identified by CWW, exactly matching the repairs for goal-specific competing heading/link problems on the RepairedExamples and RepairedNoExamples pages.

By simply deleting the lists of examples for each link label, RepairedNoExamples Humanities and Sciences pages were created from and substituted for RepairedExamples pages used in Experiment 1. RepairedNoExamples pages are a shortcut repair compared to RepairedExamples pages, because designing the lists of examples is a very time-consuming process. The RepairedNoExamples pages are also more scannable than RepairedExamples pages – a potential advantage unless deleting the lists of examples significantly reduces users' accuracy when selecting links.

Procedure

The procedure for Experiment 2 was identical to the procedure for Experiment 1.

Results

Since Experiments 1 and 2 were identical except for the change in web page materials, the data from the two experiments was combined, pooling the data from all 204 people who participated in the two experiments. The full set of data were then analyzed with a three variable mixed between- and within-group Repeated Measures ANOVA that had four conditions instead of the original two: Unrepaired, RepairedExamples, RepairedNoExamples, and SemiRepaired.

The between-group variable for Condition was significant for both the Humanities web page, *F* (3, 188) = 87.195, *p*<.0001, and the Sciences web page, *F* (3, 190) = 69.936, *p*<.0001. As Figure 1 shows, the RepairedNoExamples shortcut appears to show performance equivalent to RepairedExamples, and the SemiRepaired shortcut appears to have captured most of the performance gains but not be quite as good as RepairedExamples. Bonferroni/Dunn Post Hoc tests confirm these visual observations. For both the Humanities and Sciences web pages, all three repaired pages were significantly better than the Unrepaired pages (consistently *p*<.0001).

In addition, the ANOVA Post Hoc tests showed no significant differences between RepairedNoExamples and RepairedExamples for either Sciences or Humanities pages. For the Sciences web page the RepairedExamples and RepairedNoExamples conditions were significantly better than SemiRepaired condition (*p*=.0009, *p*=.0017, respectively). The difference between RepairedExamples and SemiRepaired approached significance for the Humanities web page (*p*=.0288, not quite meeting the significance criterion of .0083 for the Bonferroni/Dunn).

There was a significant difference for the within-group variable of Problem Type (Competing Links, Competing Headings, Unfamiliar, TwoOrMoreProblems) for both the Humanities web page, *F* (3, 564) = 26.051, *p*<.0001, and the Sciences web page, *F* (3, 570) = 65.674, *p*<.0001. As Figure 1 shows, TwoOrMoreProblems was the most difficult problem type, and Competing Links, least difficult.

The presentation order for the target articles (Original vs. Reverse), a between-group variable, did not significantly affect performance for either the Humanities or Sciences pages, and none of its interactions were significant.

Only one interaction was significant, the interaction between Condition and Problem Type that was significant for both the Humanities web page, *F* (3, 564) = 6.142, *p*<.0001, and the Sciences web page, *F* (3, 570) = 5.649, *p*<.0001. This interaction results from better performance gains for Competing Links and Competing Headings problem types than for Unfamiliar and TwoOrMoreProblems types, as shown in Figure 1.

Discussion of Experiment 2

The results of Experiment 2 make it very clear that the effort required to generate examples used in the RepairedExamples conditions of Experiment 1 was not worth it. It is possible that participants did not bother to carefully process the examples. Nielsen [13] and others have repeatedly made the point that user skim web pages and act without bothering to process all of the material provided by developers.

GENERAL DISCUSSION

The large gains for the Competing Link and Competing Heading items can be explained in large part by the nature of the task: search through hierarchically structured categories for a target item (e.g., looking for an article in an online encyclopedia or browsing for a book on a given topic). People find a particular item semantically similar to more than one category and are faced with two or more competing choices. Such confusions will be especially prevalent in users who are not domain experts, and who do not have the knowledge necessary to make the fine grain distinctions made by the domain experts who constructed the category hierarchy. The proper repair is easy: use LSA to identify all of the headings and links that typical users will perceive as being semantically similar to the target item and offer access via all likely paths. All three repaired conditions did this and yielded strong performance gains.

The data on unfamiliar problems from Experiments 1 and 2 make clear some of the inherent constraints on the process of developing successful websites for users with different backgrounds. Many unfamiliar headings and links problems are difficult to repair because users, in this case college freshman, have limited domain knowledge. For example, the corpus used to generate the LSA semantic space for first-year college students did not include much material on such topics as anthropology or paleontology, and paraphrasing link labels with more familiar words cannot easily compensate for low background knowledge of such topics. On the other hand, repairing link labels worked well for a few items. For example, subjects took 7.2 clicks to find Medicine Man when the correct link was labeled The Occult, but only 2.6 clicks for the repaired link Magic, Supernatural & Spirits. A different repair strategy for unfamiliar problems would let users find the target by clicking the familiar topic that has the highest goal-link cosine, but we have not yet tested this repair strategy.

Blackmon et al. [2] used a different measure of performance – percent correct on first click – to compare goals that CWW identified as having problems with goals for which CWW identified no problems. For comparability to previously reported work, Figure 2 shows the results for the percentage correct on first click. All three repaired conditions had similar performance, and all three elicited far better performance than the Unrepaired condition. Humanities and Sciences pages got similar results.

Usability Inspection Methods and Repairs

Usability inspection methods are a class of techniques for evaluating a user interface by examining and critiquing it. The critique would normally be based on experience, psychological principles, or a set of previously defined guidelines. These techniques include Guideline Review [16] in which evaluators check guideline conformance, Cognitive Walkthrough [15] in which evaluators simulate users' problem solving, Heuristic Evaluation [12] in which evaluators identify violations of design heuristics, etc. Repairs for the identified problems are devised by the design team that is conducting the evaluation. The design team can successfully repair usability problems for users *highly similar* to themselves in background knowledge.

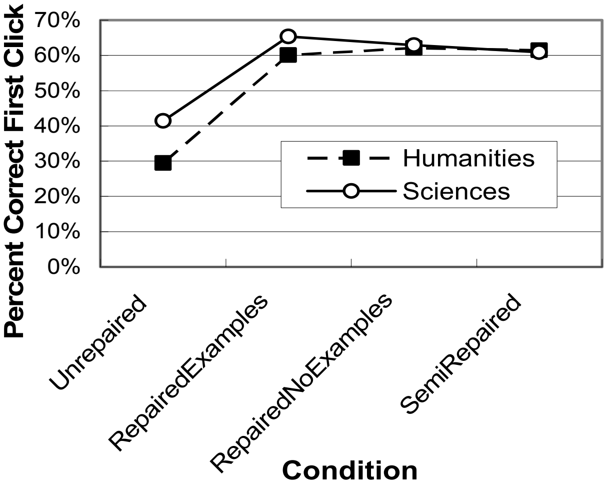

Figure 2. Experiments 1 and 2: Percent correct on first click for 4 conditions on Humanities and Sciences web pages

CWW offers a significantly different method for devising repairs. CWW can identify and successfully repair usability problems caused by a mismatch between the background knowledge of developers designing the website and the background knowledge of users visiting the website in search of information. The greater the disparity in background knowledge, the more difficult it is for designers to devise successful repairs. To repair an unfamiliar link label, for example, the design team, consulting the word-knowledge base in their own heads, would have difficulty listing words familiar to the intended users. CWW, in contrast, can identify and repair usability for any particular user group, provided CWW has a semantic space to accurately represent the user group. CWW currently offers semantic spaces for college-level general reading knowledge in English and French, and for 3rd-, 6th-, 9th-, and 12th-grade reading knowledge in English. Efforts are underway to expand the number of semantic spaces to span multiple reading levels in various languages.

Usability experts and developers can and do make judgments of familiarity and similarity of headings and links, but these intuitions cannot replace LSA. Members of a team developing a website are, or rapidly become, domain experts on the content of the site. A designer with expert domain knowledge and the individuals modeled by the semantic space for first-year-college general reading knowledge will make very different judgments of familiarity and similarity. LSA would model a domain expert by processing a corpus made up of a large amount of reference material from that domain, producing a semantic space very different from the first-year college space, with resulting differences in familiarity and similarity ratings.

CONCLUSIONS

An important claim made by Wharton, Rieman, Lewis, & Polson [15] was that the theory underlying the original Cognitive Walkthrough could be used to derive repairs for the original design, in addition to being used to detect interface flaws that would prevent learning by exploration. The results reported here support the same claim for CWW.

The experimental encyclopedia website used in our experiment exemplifies many similar tasks performed by users of websites. These tasks all involve browsing through a set of hierarchically organised headings and links to find a target page that will enable the user to accomplish his goal (read a target article, purchase a product with desired characteristics, etc.). In the following, we will assume that the task involves selecting between one of several links.

Our results show that even with well designed link descriptions, there can be multiple reasonable link selections for a given target or goal. Our most important results show that an effective solution to this problem is to provide multiple access paths to the same target page and that LSA can be used to identify these alternative paths.

Our results also make clear the limitation that developers of a website face where trying to serve user populations who have limited knowledge in the content domain of the site (e.g., medicine, computers, cameras, etc.). For example, categorizing products by the technical specifications will generate a set of links that are incomprehensible to users without the relevant domain knowledge. There are very similar problems with medical/health websites that use medical terms to label links. Most medical/health websites are accessible only to individuals with college and advanced degrees, not to users with low background knowledge of medicine and low reading knowledge [1]. CWW can accurately identify unfamiliar link labels, but more research is needed to expand CWW repair options for unfamiliar heading/link problems. Making the link text more familiar only sometimes solves the problem.

Finally, using CWW and LSA enables developers to overcome an important limitation of the interaction design processes. It is possible for an experienced developer to generate a set of links that seem comprehensible and not confusable to her, based on her expert knowledge of the content domain. Using CWW enables a developer to very quickly check if these newly designed links are also comprehensible and not confusable to her intended users, and if not, repair them.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 01-37759 to the first author. We are grateful to Dr. David J. Wade-Stein, Research Associate at The Center for Spoken Language Research at the University of Colorado. He generously shared his LSA code adapted for Linux to install on our new web server at http://AutoCWW.colorado.edu, and he has been an invaluable resource to us for his LSA expertise.

REFERENCES

1. Berland, G.K. at al. Health information on the Internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association*, 285, 2612-2621, 2001
2. Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. Cognitive Walkthrough for the Web. *Proceedings of CHI 2002*, ACM Press, 463-470, 2002.
3. Chi, E., Pirolli, P., & Pitkow, J. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a website. In *Proceedings of CHI 2000,* 161-168, ACM Press, 2000.
4. Chi, E., Pirolli, P., Chen, K., & Pitkow, J. Using information scent to model user information needs and actions and the Web. In *Proceedings of CHI 2001*, ACM Press, 490-497, 2001.
5. Furnas, G. W. Effective view navigation. In *Proceedings of CHI'97*, ACM Press, 367-374, 1997.
6. Katz, M. A., & Byrne, M. D. Searching and browsing on e-commerce sites: Frequency, efficiency, and rationale. Manuscript submitted for publication.
7. Kintsch, W. *Comprehension: A Paradigm for Cognition*, Cambridge University Press, 1998.
8. Kitajima, M. & Polson, P. G. A comprehension-based model of exploration, *Human-Computer Interaction*, **12**, 345-389, 1997.
9. Kitajima, M., Blackmon, M. H., & Polson, P. G. A Comprehension-based model of Web navigation and its application to Web usability analysis. In *People and Computers XIV*, Springer, 357-373, 2000.
10. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240, 1997.
11. Larson, K., & Czerwinski, M. Web page design: Implications of memory, structure and scent for information retrieval. In *Proceedings of CHI'98*, ACM Press, 25-32, 1998.
12. Nielsen, J. & Mack, R. L. *Usability Inspection Methods*, New York: John Wiley & Sons, Inc., 1994.
13. Nielsen, J. *Designing Web Usability*, New Riders Publishing, 2000.
14. Pirolli, P. & Card, S. Information foraging. *Psychological Review*, **106**, 643-675, 1999.
15. Wharton, C., Rieman, J., Lewis, C., & Polson, P. The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods*, New York: John Wiley, 105-140, 1994.
16. Vanderdonckt, J. Development milestones towards a tool for working with guidelines. *Interacting with Computers,* **12,** 81-118.1999.