18

Automating Usability Evaluation: Cognitive Walkthrough for the Web Puts LSA to Work on Real-World HCI Design Problems

Marilyn Hughes Blackmon
Dipti R. Mandalia
Peter G. Polson
University of Colorado, Boulder

Muneo Kitajima
National Institute of Advanced Industrial Science and Technology, Japan

When people navigate a relevant Web site for information needed to solve a problem, they encounter two subproblems. The first is navigating within the Web site to find a relevant web page(s) or a relevant document downloadable from the Web site. The second subproblem is comprehending the retrieved information.

For the past 5 years, we have addressed the first subproblem by developing a usability evaluation method (UEM) called the cognitive walkthrough for the Web (CWW). This chapter focuses primarily on how CWW employs LSA to identify and repair usability problems that impair navigation of large, complex Web sites. By now we have collected a large amount of evidence demonstrating that CWW reliably and validly predicts the usability problems that impede navigation of a Web site to retrieve information (Blackmon et al., 2002, 2003, 2005).

The special genius of LSA is its versatility to switch among a variety of different semantic spaces. Each LSA semantic space used by CWW is constructed from a scientifically sampled corpus of documents, emulating Zeno, Ivens, Millard, and Duvvuri (1995).[1] Scientifically sampled corpora ensure that the semantic space faithfully represents the background knowledge and general reading ability of a particular population, such as 3rd-, 6th-, 9th-, or 12th-grade general reading knowledge of American English. This versatility of LSA enables us to apply CWW to evaluate the usability of a given Web site for a diverse array of user populations. For example, CWW might simulate navigating a one-size-fits-all online encyclopedia to answer a range of information needs, predicting that users with college-level general reading knowledge would be successful, but that users with 3rd- or 6th-grade reading knowledge would experience frustration and probably fail to find the information they needed in the same online encyclopedia.

Recently, one of us (Mandalia, 2004) extended the research program in a new direction by addressing the second subproblem users face: whether users can comprehend the content that they find. Mandalia created a new LSA-based tool that integrates with the workflow of content writers. A section near the end of this chapter describes how this tool supports content writers in improving the comprehensibility of content material made available on the Web site, and targeting one or more particular user groups whose level of background knowledge differs markedly from the writers' own level of background knowledge.

THEORETICAL ROOTS UNDERLYING CWW

Theories of How People Navigate Complex Informational Web Sites

There is widespread agreement that a common problem-solving process underlies both navigating a Web site to find specific information (or to find

[1] The five tasa semantic spaces were built from the scientifically sampled corpora that Zeno et al. (1995) collected to create their definitive guide to word frequency for American educators. The five corpora are divided by applying the degrees of reading power (DRP) measure of readability found at http://www.tasaliteracy.com/

a product), and performing a novel task using a complex Macintosh or Windows application. This common problem-solving process is the process of learning or performing by exploration (Chi, Pirolli, Chen, & Pitkow, 2001; Kitajima, Blackmon, & Polson, 2000, 2005; Kitajima & Polson, 1997; Pirolli, 2005; Pirolli & Card, 1999; Pirolli & Fu, 2003; Pirolli, Fu, Chi, & Farahat, 2005; Rieman, Young, & Howes, 1996; Soto, 1999). Simulation models of performing by exploration have been derived from different theoretical foundations, but all share the underlying assumption that exploratory behavior is guided by perceptions of semantic similarity. The most common heuristic for directing exploratory behavior is to act on an object in a display (e.g., release on a menu item, click on a link) whose description is perceived by a user to be semantically similar to the user’s current goal.

When reporting how users perform Web site navigation tasks by exploration, there is consensus among researchers that users follow a trail of *scent*, or *information scent*, that involves users’ perceptions of semantic similarity between their current goal and each of the links they might click on any web page they visit in a particular Web site (Blackmon et al., 2002, 2003, 2005; Chi et al., 2001, 2003; Furnas, 1997; Pirolli, 2005; Pirolli & Card, 1999; Pirolli & Fu, 2003; Pirolli et al., 2005). Although CWW researchers follow the consensus and employ the term *information scent*, CWW researchers diverge from the consensus both by measuring information scent within a particular LSA semantic space and by distinguishing two dimensions of information scent: semantic similarity between the user’s goal and web page texts and familiarity of the terms in the web page texts.

CWW is designed to identify and repair design flaws in a Web site that would cause navigating by scent-following to fail, and the heuristic of guiding navigation by following a scent trail can fail in any of three different scenarios. In the first scenario, scent can be weak or nonexistent for the correct heading or link, failing to provide users with any guidance on what link to click. In the second scenario, the heading or link text may be unfamiliar and, thus, relatively meaningless to users (e.g., a correct link that uses a low-frequency scientific or medical term). In the third scenario, following a strong scent actively misdirects users, causing them to detour away from solution paths that would lead to achieving their goals.

Designers of Web sites, therefore, need to accurately predict users’ perceptions of similarity and familiarity, but designers and users frequently do not share common understandings and therefore have very different perceptions of similarity and familiarity. Even when compared to a typical college-educated user, the typical Web site designer has a great deal more knowledge than the typical user has, including exceptional knowledge of the content domain, specialized terms used in heading and link descriptions, and web page layout conventions. Thus, designers’ subjective judgments of similarity and familiarity, unaided by CWW, can be very different from perceptions of similarity and familiarity of target user populations. Thus, the validity of the CWW design evaluation and repair processes described in this chapter depends on designers being able to set CWW to a particular LSA semantic space to accurately simulate users' judgments of similarity and familiarity.

CoLiDeS Cognitive Model and the Construction-Integration Cognitive Architecture

Our CWW research began by adapting the Cognitive Walkthrough (CW), a widely used UEM originally developed to support the design and evaluation of application interfaces (Blackmon, 2004; Wharton, Rieman, Lewis, & Polson, 1994). CW has been applied to evaluating how well users could execute tasks on a walk-up-and-use interface (e.g., an ATM) or perform novel or infrequent tasks in a complex suite of applications, such as Microsoft Office. CWW retains a major advantage of CW: Designers can apply it early in the process of designing and building an application or Web site, avoiding the necessity of making expensive major changes to fix usability problems after the application or Web site has already been designed and built.

Kitajima and Polson (1997; Kitajima, Blackmon, & Polson, 2000, 2005) exploited the fact that their earlier theoretical models of learning and performing by exploration could be extended to analyze how to design Web sites to facilitate successful forward search (relying on hill-climbing, a general problem-solving method) and to prevent Web site design flaws that hinder successful navigation. The resulting model for Web site navigation is CoLiDeS, an acronym for *Co*mprehension-based *Li*nked model of *De*liberate *S*earch (Kitajima, Blackmon, & Polson, 2000, 2005). The action-planning processes of learning by exploration are comprehension based, and CoLiDeS is based on Kintsch's (1998) construction-integration (CI) theory of text comprehension and action planning. Information scent is generated by processes closely related to text comprehension, and scent, like text comprehension, depends on background knowledge.

A CI model of web navigation requires a robust, realistic model of attention processes, because clicking a link confronts the user with a new page containing many targets for action. Accordingly, CoLiDeS incorporates a model of attention management and a two-phase model of action planning. Each phase uses a pair of CI cycles. In the first phase of action planning, the user parses the web page into subregions and generates text descriptions of each region, perhaps describing one region as a "hierarchically organized side navigation menu" and another region as a "collection of links to physics, chemistry, and other physical sciences." Contingent on the user's own particular background knowledge, the user derives descriptions of subregions from (a) comprehension of heading texts, if any headings are used on the web page, and (b) knowledge of the functions of various subregions of a particular Web site (if the user has enough prior experience with that Web site) or default conventions for a typical Web site (if the user has prior experience with Web sites and can recognize, and know the conventional functions of, the side navigation bar, top navigation bar, content area, site logo, etc.). At the end of the first phase of action planning, a user focuses attention on a subregion of the web page whose description is familiar to the user and semantically similar to the user’s goal. In sum, during the first phase of action planning, a pair of CI cycles parses the web page into subregions and ends by focusing attention on a subregion of a web page most similar to the user’s goal (an attention action).

In the second phase of action planning, the first CI cycle identifies each possible target for action within the focused-on subregion (e.g., links, buttons, graphics), comprehends link label texts in that subregion, and generates a description for each target for action. Background knowledge is critical for understanding the consequences of clicking the various specific links, such as what to expect after clicking a link labeled “Music.” At the end of the second phase of action planning, the second CI cycle selects a specific action on an object from the attended-to subregion, for example, clicking the link labeled “Music.” CoLiDeS assumes the target for action is familiar to the user and semantically similar to the user’s goal.[2]

CoLiDeS uses LSA to model users’ perceptions of similarity and familiarity and combines them with prior experience with Web site widgets into a complex measure of information scent. In a full running simulation of CoLiDeS, the complex measure of information scent would be integrated into a single activation value, that is, a measure of the probability that the user will select a particular link or other screen object (Kitajima, Blackmon, & Polson, 2005).

When CoLiDeS simulates a human user successfully navigating a Web site by pure forward search, the heading and link that are semantically most similar to the user’s goal must also be *correct* and use *familiar* text labels. Throughout this chapter, the term *correct* consistently means that clicking the *correct* link nested under the *correct* heading in the *correct* subregion of the web page actually leads the user expeditiously to the target web page and, thus, to accomplishment of the user’s goal in the minimum number of clicks. The term *familiar* consistently means that the user can comprehend the term or terms used to label the heading and link and has sufficient background knowledge to select the correct heading and link. In contrast, if the correct heading and link are unfamiliar or not semantically similar to the user’s goal, CoLiDeS, simulating the human user, will flounder and be forced to backtrack or detour. If the problems are severe, then a human

[2] A demonstration of CoLiDeS can be found at http://www.staff.aist.go.jp/kitajima.muneo/CoLiDeS_Demo.html

user—or CoLiDeS simulating a human user—is likely to experience task failure (i.e., fail to accomplish the goal and give up navigating the Web site).

USABILITY PROBLEMS IDENTIFIED BY CWW

In order to identify usability problems with the navigation system of a web page, the CWW analyst must determine which link(s), heading(s), and subregion(s) are *correct* for accomplishing that goal on the web page. In contrast, CoLiDeS simulates human users and, like users, has no way of distinguishing whether a particular link, heading, or subregion is correct or incorrect. Thus, CoLiDeS, like human users, struggles when it encounters the following four types of CWW-identified usability problems while navigating Web sites to accomplish particular tasks. The first three of the following four types of usability problems have been verified by experiments reported in the earliest publications (Blackmon et al., 2002, 2003). The fourth has emerged more recently and, as CWW has evolved, we have made slight changes in the parameters for problem identification (Blackmon, Kitajima, & Polson, 2005). Appendix A lists the exact, automatable rules and LSA parameters that CWW now uses to identify these four usability problems.

An *unfamiliar correct link* problem can potentially occur whenever target users of the Web site lack sufficient background knowledge to comprehend a link text and accurately estimate its similarity to their current goals. A short LSA term vector length has empirically proved to be the most useful CWW index of insufficient background knowledge. Low word frequency in the selected LSA semantic space is, however, another important marker, because typical users cannot recognize or comprehend low-frequency technical terms and other low-frequency words. In some cases, such as *paleontology*, a link label is both a low-frequency word and a term that has a short-term vector length. In other cases, such as *anthropology* in the college-level semantic space, a link is not a low-frequency word but nevertheless has a short-term vector length. Thus, although college-level users recognize the word, they tend, even so, to have only sparse, vague knowledge of the full range of information that falls within the scope of anthropology. Both users and CoLiDeS tend to ignore links they fail to understand clearly, and ignoring an *unfamiliar correct link* causes serious problems because users must click that link to reach the target web page that accomplishes their goal. A preventive repair strategy is best to avoid *unfamiliar correct link* problems. Designers would ideally identify and repair all unfamiliar link texts on each web page before posting the web page on the Internet.

A *competing heading* problem arises when any heading and its associated subregion is semantically very similar to the user goal but does not contain a correct link that leads to accomplishing the user goal. Like users, CoLiDeS follows the information scent trail and has no way of knowing if a given subregion contains a correct link(s). Competing headings problems are liable to be serious problems, because they divert the user's attention away from the "correct" subregion. Users often click several links under a focused-on subregion before switching their attention from that subregion to another semantically similar subregion. Indeed, Blackmon et al. (2005) found that the best measure of competing heading problems is the number of attractive links within all competing subregions, called *competing links nested under competing headings*. Many high-scent links increase the user's perception that the correct link is somewhere within a high-scent competing subregion, so a user will probably click many links in that competing subregion, even exhaustively clicking relatively unlikely links, before leaving that subregion. Designers can prevent some competing heading problems by using high-quality link and heading labels (Blackmon et al., 2003; Miller & Remington, 2004), but some goals inevitably require cross-classification under two or more subregions (e.g., users search for information about music therapy under music, psychotherapy, and medicine links that necessarily belong in three different subregions).

A *competing link* problem occurs when a correct or competing subregion contains one or more links that are semantically similar to the user goal but not on the solution path. In recent work (Blackmon et al., 2005), we have begun using the more precise term *competing links nested under a correct heading*. Competing links located within a correct subregion may distract the user momentarily, but the user usually persists and eventually clicks the "correct" link before abandoning the "correct" subregion, but *competing links nested under competing headings* are more serious, as already indicated. A designer can prevent many competing links problems by changing link label text to reduce similarity among link labels within each subregion (Blackmon et al., 2003).

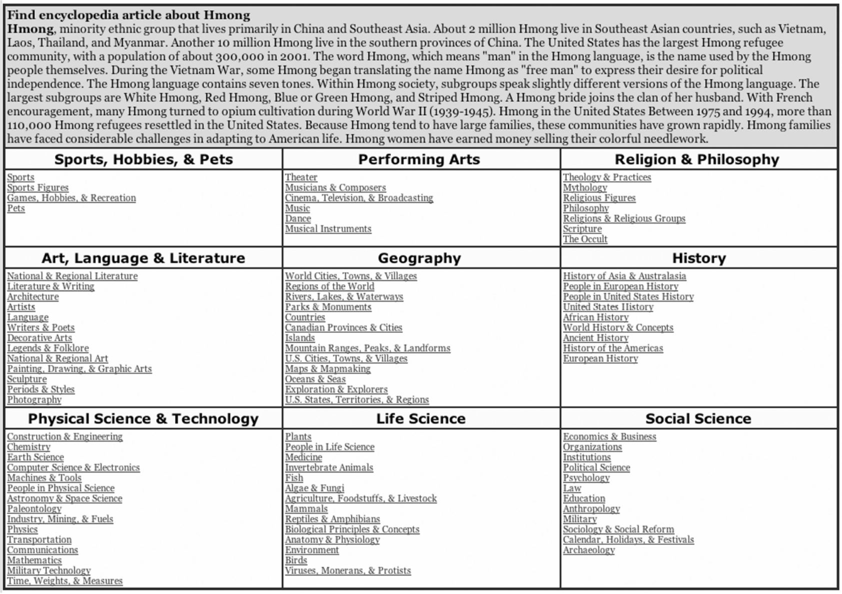
A *weak-scent correct link* problem refers to the situation when a correct link is semantically unrelated to the user goal (near-zero LSA cosine), and when there are no other correct links that have moderate or strong scent. CoLiDeS and human users generally ignore links that they perceive as semantically unrelated to the current user goal and have no way of knowing which link is correct. Therefore, a weak-scent correct link causes serious problems. Designers can prevent weak-scent correct links by devising high-quality link labels and testing the scent of each link label for a large set of typical user goals that will require users to click that link.

CONCRETE EXAMPLE OF AN EXPERIMENTAL TASK

A concrete example offers the simplest way to grasp how the CoLiDeS model guides experimental design, how CWW can be used to predict the difficulty of doing a specific task on a particular web page, and how we test the psychological validity of CWW predictions using controlled laboratory experiments. We have deliberately selected a task that CWW predicts will be very difficult for users (or for CoLiDeS simulations of human users), because users will encounter all four of the usability problems described previously while doing this task.

The particular task—Find Hmong—involves finding an article about the Hmong people by navigating an online encyclopedia. The main web page for the Find Hmong task in our controlled laboratory experiments is shown in Figure 18.1, a web page that closely simulates a popular online encyclopedia Web site and has 93 topic links nested under nine category headings. "Anthropology" is the only link that leads to the article on Hmong in the actual online encyclopedia Web site simulated in our experiment. Therefore, to complete the Find Hmong task on the experimental Web site, experimental participants had to click the correct link "Anthropology" nested under the correct heading "Social Science."

Experimental participants who did this task could see and read the Find Hmong user goal in the box at the top of the web page shown in Figure 18.1. The Find Hmong goal is a 205-word summary of the full encyclopedia article on Hmong, and accurately represents the actual article because it has an LSA cosine of .82 with the full article. (If the summary had contained ex-



Find encyclopedia article about Hmong

Hmong, minority ethnic group that lives primarily in China and Southeast Asia. About 2 million Hmong live in Southeast Asian countries, such as Vietnam, Laos, Thailand, and Myanmar. Another 10 million Hmong live in the southern provinces of China. The United States has the largest Hmong refugee community, with a population of about 300,000 in 2001. The word Hmong, which means "man" in the Hmong language, is the name used by the Hmong people themselves. During the Vietnam War, some Hmong began translating the name Hmong as "free man" to express their desire for political independence. The Hmong language contains seven tones. Within Hmong society, subgroups speak slightly different versions of the Hmong language. The largest subgroups are White Hmong, Red Hmong, Blue or Green Hmong, and Striped Hmong. A Hmong bride joins the clan of her husband. With French encouragement, many Hmong turned to opium cultivation during World War II (1939-1945). Hmong in the United States Between 1975 and 1994, more than 110,000 Hmong refugees resettled in the United States. Because Hmong tend to have large families, these communities have grown rapidly. Hmong families have faced considerable challenges in adapting to American life. Hmong women have earned money selling their colorful needlework.

Sports, Hobbies, & Pets	Performing Arts	Religion & Philosophy
Sports Sports Figures Games, Hobbies, & Recreation Pets	Theater Musicians & Composers Cinema, Television, & Broadcasting Music Dance Musical Instruments	Theology & Practices Mythology Religious Figures Philosophy Religions & Religious Groups Scripture The Occult
Art, Language & Literature	**Geography**	**History**
National & Regional Literature Literature & Writing Architecture Artists Language Writers & Poets Decorative Arts Legends & Folklore National & Regional Art Painting, Drawing, & Graphic Arts Sculpture Periods & Styles Photography	World Cities, Towns, & Villages Regions of the World Rivers, Lakes, & Waterways Parks & Monuments Countries Canadian Provinces & Cities Islands Mountain Ranges, Peaks, & Landforms U.S. Cities, Towns, & Villages Maps & Mapmaking Oceans & Seas Exploration & Explorers U.S. States, Territories, & Regions	History of Asia & Australasia People in European History People in United States History United States History African History World History & Concepts Ancient History History of the Americas European History
Physical Science & Technology	**Life Science**	**Social Science**
Construction & Engineering Chemistry Earth Science Computer Science & Electronics Machines & Tools People in Physical Science Astronomy & Space Science Paleontology Industry, Mining, & Fuels Physics Transportation Communications Mathematics Military Technology Time, Weights, & Measures	Plants People in Life Science Medicine Invertebrate Animals Fish Algae & Fungi Agriculture, Foodstuffs, & Livestock Mammals Reptiles & Amphibians Biological Principles & Concepts Anatomy & Physiology Environment Birds Viruses, Monerans, & Protists	Economics & Business Organizations Institutions Political Science Psychology Law Education Anthropology Military Sociology & Social Reform Calendar, Holidays, & Festivals Archaeology

Figure 18.1. Web page for Find Hmong Task.

actly the same text as the article the cosine would be 1.00, and if the summary had essentially no semantic similarity to the article, then the cosine would be approximately zero. Therefore, a cosine of .82 shows a high degree of semantic similarity.)

The 76 experimental participants who did the Find Hmong task were all undergraduates, so we selected the LSA semantic space for first-year college general reading knowledge. To determine which headings and links have the highest information scent for the Find Hmong goal, we performed a One-to-Many LSA analysis comparing the Find Hmong goal with each of the nine headings and 93 links shown in Figure 18.1. Then we sorted both the goal-heading cosines and the goal-link cosines by decreasing cosine value. To simulate the way that human beings elaborate text during comprehension, we had previously elaborated the link and heading texts with additional words that are both highly similar and highly familiar (see appendix B for details about elaboration of link and heading texts).

Following the CoLiDeS model, CWW assumes experimental participants will first parse the web page into subregions and focus on the subregion and heading most similar to the goal. History and Geography are the subregions with the highest scent for the Find Hmong goal (goal-heading cosines of .30 and .19, respectively), and both have stronger information scent than does the correct heading, Social Science (goal-heading cosine of only .08). Thus, CWW expects that users' attention will be actively misdirected from the correct heading to History and Geography and identifies *competing headings* problems for these two subregions. Nested under these two competing headings, History and Geography, are five high-scent links that CWW identifies as *competing links nested under competing headings* (goal-link cosines ranging from .37 down to .22), and CWW predicts that people will click many or all of these attractive links before switching attention to the correct subregion, "Social Science."

If and when the experimental participant finally gets to the correct Social Science subregion, there is still a fairly low probability of clicking the correct link. Examination of the goal-link cosines indicates that the correct link "Anthropology" has weak-scent (.08) for the Find Hmong goal—CWW calls this a *weak-scent correct link* problem—and that there is a higher scent competing link nested in the same subregion—called a *competing link nested under a correct heading*. In addition, the link "Anthropology" is an *unfamiliar correct link*, because the short-term vector for "anthropology" suggests that even college-educated users generally have inadequate background knowledge of anthropology to realize that anthropologists study the cultures and social organization of peoples like the Hmong.

In short, CWW predicts that users will encounter great difficulty finding the correct link to complete the Find Hmong task. Having identified the highest scent headings and links, however, it is now possible to design a repaired web page that would make it possible for people to do the Find Hmong task using pure forward search. The repaired web page built for the experiment makes it possible to find the encyclopedia article exactly where CWW predicts users are most apt to look for it, not just where an encyclopedia expert thinks the item is properly classified. More specifically, the repaired web page for Find Hmong makes it possible to find the Hmong article by clicking the highest scent links under History or Geography, as well as by clicking the "Anthropology" link designated correct by the designers. On the repaired web page, CWW predicts experimental participants will quickly click one of the links that actually leads to the Find Hmong article.

The performance of experimental participants closely matches CWW predictions, with 45% of the first-clicks falling on a link nested under History, 21% of the first-clicks falling on a link nested under Geography, and a mere 5% of the first-clicks falling on a link nested under Social Science. Table 18.1 shows the observed mean total clicks for the Find Hmong task for the 76 college students who did the task in our laboratory study, including 38 participants in the unrepaired web page condition and 38 participants in the repaired web page condition. For the unrepaired condition, the Find Hmong task took 9.026 mean total clicks compared with 2.135 mean total clicks in the repaired condition. The difference (measured in mean total clicks) between the two conditions of the Find Hmong task is significant, $F(1, 73) = 98.9$, $p < .0001$.

Only 26% of the students in the unrepaired web page condition ever found the Hmong article within the time limit of 130 seconds. In contrast, 100% of the students in the repaired web page condition were successful in finding the Hmong article and did it in a mean time of 41 seconds. Armed with an accurate way of predicting which links people are most apt to click, therefore, it is possible to build web pages where people accomplish their goals with pure forward search, the ideal situation according to the CoLiDeS model.

TABLE 18.1
Find Hmong Task: Repaired versus. Unrepaired Condition

Performance Measure	*Find Hmong Web Page condition*	
	Unrepaired	Repaired
First-click success rate	3%	43%
Actual mean total clicks	9.0	2.1
Success rate	26%	100%
Mean solution time	124 s	41 s
Experimental participants in each condition (76 total)	38	38

HOW THE RELIABILITY OF CWW DEPENDS ON LSA

Hertzum and Jacobsen (2003) have demonstrated that there is a disturbingly high “evaluator effect” for usability evaluation methods (UEMs) that rely on the human judgments of usability experts and developers. Similarity and familiarity judgments of human evaluators are subjective judgments anchored in the experience of individuals and can be far different from the perceptions of actual users. Hertzum and Jacobsen (2003) reviewed the available evidence for various UEM methods and demonstrated that agreement is unreliably low between the judgments of any pair of individual analysts. Increasing the number of analysts making a given set of judgments may improve the reliability of the judgments but drives up the cost of the usability evaluation.

CWW solves the UEM interrater reliability problem by substituting LSA measures of semantic similarity and familiarity in place of developers’ judgments of similarity and familiarity. CWW uses LSA measures of similarity (i.e., cosines) because LSA cosines are objective measures of semantic similarity that can be precisely replicated by any analyst using the same procedure and selecting the same semantic space. Similarly, LSA provides objective, replicable measures of familiarity (term vector length and term frequency).

A notable advantage of relying on LSA measures of similarity and familiarity is the capacity of LSA to make accurate, objective similarity and familiarity judgments for users very different from the analyst. LSA measures of similarity and familiarity are invaluable for designing heading and link labels that are usable by target audiences that have less advanced knowledge. It is particularly crucial to rely on objective LSA judgments of familiarity. People who design Web sites typically have fluent college-level general reading knowledge and high domain-specific background knowledge for the domain of the Web site. Unaided by LSA measures of familiarity, it is virtually impossible for a designer with advanced knowledge to accurately detect and flag all the terms that would be unfamiliar to users with third- or sixth-grade general reading knowledge, or unfamiliar to bicultural users whose native language and cultural background differ from the designer’s own native language and culture.

HOW THE PSYCHOLOGICAL VALIDITY OF CWW DEPENDS ON LSA

Blackmon et al. (2002, 2003) identified three types of usability problems with navigation systems—unfamiliar correct link problems, competing headings, and competing links problems—and reported a series of experiments that verified the psychological validity of CWW problem identifications and the success of CWW repairs. More recently, CWW added a new category of usability problems—called weak-scent correct link problems—and eliminated a confound in earlier data by distinguishing competing links nested under competing headings from competing links nested under a correct heading (Blackmon, Kitajima, & Polson, 2005). To identify and repair CWW problems, CWW harnesses the complete array of LSA analyses and measures and uses particular parameters in order to provide completely objective, fully automated measures of similarity and familiarity. For readers who wish to know all the details, the current procedures for CWW problem identification are precisely described in appendix A and appendix B.

After running many experiments (Blackmon et al., 2002, 2003, 2005), we had accumulated enough evidence to rise to the higher standard of proof for UEMs advocated by Gray and Salzman (1998a, 1998b). To meet the higher standard, we had to accomplish four subgoals, described in the following subsections. First, we started by building a multiple regression model and extracting a prediction formula from the model. Second, we cross-validated the regression model using an independent dataset, ensuring that our new CWW prediction formula was an accurate measure of problem severity. Third, to further test the prediction formula, we examined rates of hits versus false alarms and correct rejections versus misses. Finally, we tested the success rate for CWW-guided repairs of usability problems.

Multiple Regression Model for a Large Dataset

To create a multiple regression model and associated formula for predicting problem severity, the first step, as reported previously (Blackmon et al, 2005), was to compile results from completed CWW experiments that used pairs of tasks and met the following four specific criteria for inclusion. These four criteria for inclusion were met by 82 pairs of tasks, 164 tasks altogether, drawn from four different CWW experiments (reported in Blackmon et al., 2002, 2003, 2005), and no tasks done by these experimental groups were excluded from the dataset for reanalysis:

1. For each pair of tasks, the first criterion specified that the goal was identical for two well-matched experimental groups, but one experimental group tried to accomplish the goal on an unrepaired web page and a second group tried to accomplish the same goal on a repaired web page.
2. We set a minimum .76 cosine between the actual web page content article and the short, 100- to 200-word summary of the article that experimental participants saw on the web page (e.g., the description of

the Hmong goal in Figure 18.1). This criterion ensured that experimental participants had a fair, accurate representation of the complete target article they were trying to find in the Web site.

3. For the sample of tasks done in the unrepaired web page condition, the tasks manifested diverse combinations of unfamiliar links, and weak-scent links, competing headings, and competing links nested under both competing and correct headings.
4. Problem-solving data typically have high between-subject variance, so to ensure stable mean total click data for all 164 tasks, we required all tasks to be based on data from a minimum of 20 participants. In fact, the means for 144 of the 164 tasks were based on data from 38 or more experimental participants, and means for the remaining 20 tasks were based on data from 23 or more experimental participants.

The second step toward completing the multiple regression analysis was to develop a uniform, completely objective procedure for reanalyzing all 164 tasks in the dataset. This goal required iteratively rescoring the set of 164 tasks until we had created a set of automatable rules for identifying unfamiliar links, weak-scent links, competing headings, competing links nested under competing headings, and competing links nested under correct headings. Appendix A displays the automatable rules, and appendix B describes, step-by-step, the complex CWW procedure with the current edition of its parameters.[3]

The seven automatable rules in appendix A are all written as if–then production rules, so that a computer programmer can easily convert the rules to code. For example, two different automatable rules specify precisely defined conditions that can independently prompt classification of the heading as a competing heading. The first of the two rules specifies four conditions that must all be simultaneously met in order to trigger firing of this rule and consequent identification of a competing heading: (a) the particular heading is not a correct heading, (b) the goal-heading cosine of the heading must be greater than or equal to .8 times the goal-heading cosine of any correct heading, (c) the goal-heading cosine must be greater than or equal to .10 (i.e., not weak-scent), and (d) the highest goal-link cosine for the links nested under the heading must be greater than or equal to .20.

The seven automatable rules in appendix A paved the way for more complete automation of CWW by eliminating the subjective, time-consuming hand editing of LSA analyses[4] that we originally thought necessary

[3] For further details and concrete examples, download the file AutoCWWTutorialA.pdf at http://www.autocww.colorado.edu/~blackmon

[4] Hand editing uses human judgment to weed out likely false alarms, that is, headings and links that real people would be unlikely to focus on or select despite relatively high goal-heading or goal-link cosines.

(Blackmon et al., 2002). Brown (2005) built a new web-based interface for doing CWW, called ACWW (http://www.autocww.colorado.edu/~brownr/ACWW.php), which implements all the automatable rules in appendix A and follows the procedures defined in appendix B. ACWW automatically identifies usability problems and predicts task difficulty for any set of one or more goals performed on one or more web pages.[5]

The third step was to develop the multiple regression model of task difficulty. For our initial laboratory studies (Blackmon et al., 2002, 2003), we had deliberately selected tasks that each epitomized one class of usability problems, but in actual fact few tasks are pure examples of just one type of usability problem. Most tasks are afflicted by two or more types of usability problems, and some tasks (e.g., Find Hmong) are simultaneously afflicted by all of the CWW-identified usability problems.

By doing a multiple regression analysis of the 164-task dataset we tried to account for the variance in task difficulty, indexed by mean total clicks. For the full 164-item dataset, the mean total clicks ranges from 1.0 click to 10.3 clicks with a mean of 3.7 clicks. The multiple regression tested whether four hypothesized factors—number of competing links nested under competing headings, number of competing links nested under correct headings, unfamiliar correct links, and weak-scent correct links—were all statistically significant independent variables and how much each contributed to the overall difficulty level. For example, Find Hmong is a very difficult task (9.0 mean total clicks in the unrepaired condition), a task so difficult that it was completed by only 26% of the people who attempted to do it in the unrepaired condition. Could we have predicted that the Find Hmong task would be that difficult from knowing that it had five competing links nested under two competing headings, one competing link nested under the correct heading, an unfamiliar correct link, and a weak-scent correct link?

The multiple regression analysis explains 57% of the variance in observed mean total clicks as a function of the three hypothesized independent variables, F (4, 160) = 74.22, $p < .0001$, adjusted $R^2 = .574$. The fourth hypothesized independent variable, number of competing links nested under correct headings, was not significant. All three independent variables are statistically significant—number of competing headings, whether or not the only correct link was unfamiliar, and whether or not the only correct link as a weak-scent link—and the intercept is also significant. We were also able to show, with a secondary analysis, that the number of competing links

[5] The final step in the ACWW interface enables the analyst to choose to run multiple analyses on the selected goals and web pages, including running analyses to be run in two or more different semantic spaces, and specifying a unique set of parameters for elaborating the headings and links for different analysis run on a particular semantic space.

nested under competing headings explained a higher percentage of the variance than the alternate variable, number of competing headings.

The minimum solution path for all 164 tasks was a single click, but the statistically significant intercept of 2.199 reveals that even the non-problem tasks took an average of over two clicks to complete. The intercept and unweighted regression coefficients give us a working formula for predicting the mean total clicks:

Mean total clicks = 2.199
+ 1.656 if the correct link is unfamiliar
+ 1.464 if the correct link has weak-scent
+ 0.754 times the number of competing links nested under competing headings.

The next step after completing the multiple regression analysis was to apply the multiple regression formula to predict the mean total clicks for each of the 164 tasks in the dataset. Applying this formula to the Find Hmong task, for example, we predict 9.089 clicks for the unrepaired condition and 2.199 for the repaired condition, very close to the observed results of 9.0 for the unrepaired and 2.1 for the repaired condition. Figure 18.2 displays the accuracy of the predictions by comparing predicted and observed mean total clicks for all 164 tasks. The right half of Figure 18.2 displays nearly identical values of predicted and observed mean total clicks for the 82 unrepaired tasks in the 164-task dataset and for the 82 repaired tasks. To

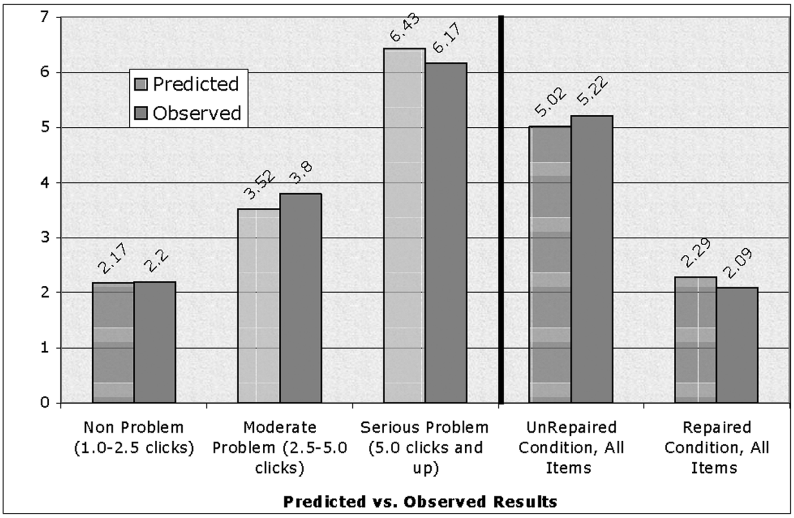


Figure 18.2. Comparison of observed and predicated scores for 164-item dataset.

test for consistency of the formula across increasing levels of task difficulty, the left half of Figure 18.2 first pooled the unrepaired and repaired tasks into a single set and then regrouped the 164 items into three groups: (a) one group of 83 *non-problem* tasks, mostly repaired tasks, that the CWW formula predicted people could do in less than 2.5 mean total clicks; (b) a second group of 35 *moderate problem* tasks, mostly unrepaired tasks, that the CWW formula predicted people would do in between 2.5 and 5.0 clicks; and (c) a third group of 46 *serious problem* tasks, all unrepaired tasks, that CWW predicted would take people 5.0 or more clicks. For all three groups, the predicted and observed mean total clicks are very close in value.

To test whether these thresholds—2.5 clicks and 5.0 clicks—were reasonable for distinguishing non-problems from moderate problems and moderate problems from serious problems, we found 100 tasks from the 164-item dataset for which we have recorded the percentages of experimental participants who could not complete the task within the allotted time (usually 130 seconds). A simple regression of the percentages of task failure per task on observed mean total clicks for the same task yields a correlation of .93, $F(1, 98) = 651.78$, $p < .0001$, adjusted $R^2 = .87$. Using the regression formula derived from the analysis—percent task failure = (.082 times observed mean total clicks – .154)* 100—we estimated a task failure rate of 5% at 2.5 mean total clicks (operationally defined as the threshold between non-problem and problem items), 26% at 5.0 mean total clicks (operationally defined as the threshold between moderate and severe problems), 51% at 8.0 mean total clicks, and 76% at 11.0 mean total clicks. It is a matter of judgment whether it is a “serious” problem when the task failure rate exceeds 25%, but in our judgment 25% task failure is an unacceptable failure rate, particularly considering that this is the failure rate for college-educated users, and we can assume that task failure rates will be higher for people with more modest levels of general reading knowledge.

Cross-Validation of Multiple Regression Model

It is crucial to replicate a multiple regression model on a completely new set of tasks, so we gathered new data that addressed the same four criteria (see earlier) but with two significant changes. Instead of comparing the same tasks on repaired and unrepaired web pages, as specified by the first criterion, we selected 28 tasks that CWW predicted to have usability problems and 36 tasks predicted to cause no problems. The 64 tasks were all done on a simulation of the online encyclopedia Web site with 93 links nested under nine categories, and for each task only one of the 93 links was correct. Thus, the comparison between problems and non-problems controlled for number of correct links, a variable that was not controlled in the 164-task dataset, where problem tasks generally had just one correct link but repaired tasks usually had two or more correct links.

Although the dataset is smaller (64 tasks instead of 164 tasks), the multiple regression analysis fully replicated the results of the original multiple regression analysis (as reported in Blackmon et al., 2005). The same three independent variables and intercept were all highly significant, the competing links under competing headings was still superior to the competing headings variable, and the competing links under correct heading variable was still nonsignificant. For the 64-task dataset, the multiple regression model explained 50% of the variance, *F* (3, 60) = 22.042, $p < .0001$, adjusted $R^2 = .50$. Table 18.2 compares the coefficients from the two multiple regression analyses, the original and cross-validation studies, putting the cross-validation study values in parentheses.

Hits Versus False Alarms

Table 18.2 shows a close alignment between the original and cross-validation study, and Figure 18.2 shows little discrepancy between predicted and observed mean total clicks. Nevertheless, to meet the standards advocated by Gray and Salzman (1998a, 1998b), we must also report the rates of hits versus false alarms for tasks that CWW predicts to have usability problems.

At the time the experiments were performed, all 82 unrepaired tasks from the 164-task dataset had been classified as usability problems by then-defined criteria, but by current refined CWW procedures, only 75 of the 82 tasks in the 164-task dataset are predicted to have usability problems (i.e., predicted to take 2.5 or more clicks to complete). The overall hit rate for these 75 tasks in the unrepaired condition is 69/75 (92%), and the false alarm rate was 6/75 (8%). In other words, 92% of the tasks predicted to require 2.5 mean total clicks or more actually did require 2.5 clicks or more,

TABLE 18.2
Multiple Regression of Actual Mean Total Clicks on Three Independent Variables for Two Datasets: 164-Task Original Dataset Versus 64-Task Cross-Validation Dataset Shown in Parentheses

Independent Variable	*Unweighted Coefficient*	*Standard Error*	*Standard Coefficient*	*t-Value*	*p-Value*
Intercept	2.199 (2.481)	.157 (.257)	2.199 (2.481)	14.010 (9.656)	< .0001 (< .0001)
Number of competing links under competing headings	.754 (0.551)	.070 (.123)	.578 (.423)	10.774 (4.492)	< .0001 (< .0001)
Unfamiliar correct link	1.656 (2.040)	.324 (.571)	.264 (.330)	5.104 (3.573)	< .0001 (.0007)
Weak scent correct link	1.464 (1.484)	.306 (.491)	.254 (.280)	4.785 (3.021)	< .0001 (.0037)

and the remaining 8% actually required fewer than 2.5 clicks. Because usability experts prioritize identifying and repairing the most serious usability problems, we narrowed our focus to the subset of 46 tasks predicted to pose serious problems (i.e., those problems that CWW predicted would take 5.0 or more mean total clicks), finding a hit rate of 46/46 (100%) for serious problems (meaning that experimental participants took at least 2.5 clicks to complete all of these tasks).

For the cross-validation study (64-task dataset), the hit rate was 26/29 (90%), and the false alarm rate was 3/29 (10%). For the subset of 17 serious problems (predicted mean clicks 5.0 or higher), the hit rate was 15/17 (88%).

Correct Rejections Versus Misses

The rates for correct rejections versus misses come only from cross-validation study data, because the cross-validation study was the only CWW experiment done to date that tested tasks predicted to be non-problems. For the 35 tasks predicted to be non-problem items, the correct rejection rate was 24/35 (69%), and the rate of misses (observed mean clicks equal to or higher than 2.5) was 11/35 (31%). Nevertheless, most of the misses posed only minor problems, and the remaining 4/35 (11%) of the predicted non-problems had observed mean clicks greater than 3.5 but less than 5.0. Thus, none of the misses were observed to pose serious problems.

Success Rates for Repairs

Another important question concerns the success rate for CWW-guided repairs of the usability problems that CWW identifies (Blackmon et al., 2003, 2005), a question that can only be answered by data from the original 164-task dataset that compared tasks done on repaired and unrepaired web page conditions.

A rigorous standard for defining a “successful repair” for a given task is to require statistically significant superiority in performance for the task done on a repaired web page compared to the same task done on an unrepaired web page. We can then tally the number of tasks that meet this rigorous standard and divide it by the total number of tasks. Out of the 82 unrepaired tasks in the original dataset, 75 are still predicted to be problems by current CWW criteria, and the overall success rate for repairs for these 75 problems is 83% (62/75). Because usability experts are often under time constraints that force them to prioritize identifying and repairing the most serious problems, it is important to narrow our attention to the 46 problems predicted to be serious problems. As already mentioned, there was a 100% hit rate for these 46 tasks, and the success rate for repairing the usability problems for these tasks was 93% (43/46).

CWW-Guided Methods of Repair

We completed one study comparing the performance of rigorous but time-consuming methods of repairing usability problems with the performance of a discount repair method, finding that the discount method delivers most of the performance gains with less investment of time (Blackmon et al., 2003). That study looked only at performance gains on the repaired web page, ignoring consequences for the Web site. For each user goal that encounters usability problems on a given web page, the discount method of repair Web site generally activates two or more links and the web developer must consequently continue these paths down through the hierarchy until they reach the target web page. Thus, the discount method quickly solves the usability problem on one web page, but there is a trade-off for the Web site as a whole: Each subordinate page must then be repaired to ensure that users can ultimately get to the target web page. The repairs to subordinate web pages are particularly costly if the Web site has a deep hierarchy. If multiple links are activated on these subordinate pages, then the effects branch out to require repairs on many web pages at many levels of the Web site.

In contrast, the more rigorous method of repair changes minimizes the need for activating multiple links by first improving the quality of the link and heading labels. Recently, Miller and Remington (2004) showed that improving "link quality" can make a deeply hierarchical site architecture function well. To improve "link quality," the rigorous CWW repair method uses LSA similarity and familiarity measures. One goal is to improve the coherence within each spatially distinct group of links, aiming for (a) high similarity between each pair of links within a group, (b) high similarity between the heading for the group and each link in the group, and (c) sufficient semantic distance between each pair of groups to minimize competing headings problems.

The rigorous CWW repair method also repairs unfamiliar problems by inserting or substituting familiar words. For example, the link label "Paleontology" has low word frequency and a short-term vector length, and it can be repaired by changing the link label to "Paleontology and Fossils" or "Fossils of Extinct Species," or "Fossils and Prehistoric Species." In the empirical study of repairs (Blackmon et al., 2003), changing the link label produced performance gains in some cases, but there is no quick and easy way to compensate for users' low background knowledge of a particular topic.

HOW LSA ENABLES CWW TO SCALE UP TO EVALUATING LARGE WEB SITES

Earlier UEMs had a serious problem of scale. Because it is very time consuming to perform these UEMs, they do not scale to the evaluation of large applications or Web sites. In contrast, the LSA component of CWW makes it feasible for CWW to scale up to the evaluation of large Web sites. Kitajima et al. (2005) tested this by writing computer programs to perform CWW for finding all the encyclopedia articles in a particular online encyclopedia, an encyclopedia with over 40,000 content pages. Blackmon (unpublished) subsequently built an experimental Web site to test 20 tasks that the automated usability analysis had predicted to be serious competing headings problems and 20 tasks predicted to be unfamiliar problems. All 40 unrepaired tasks proved to be serious or moderate usability problems when tested in the lab, producing a 100% hit rate for the sample of 40 tasks produced by the CWW of the large-scale Web site.

LSA is essential for automated CWW analyses of large Web sites. Our development of automatable rules for identifying usability problems, the multiple regression analysis of the 164-item dataset, and its cross-validation with an independent 64-task dataset, took a giant leap toward full automation of CWW. ACWW, the new, more automated web-based interface for doing CWW (http://www.autocww.colorado.edu/~brownr/ACWW.php) built by Brown (2005) currently requires the analyst to manually input a set of one or more user goals and a set of one or more web pages. ACWW then automatically identifies CWW usability problems (unfamiliar correct link, weak-scent correct link, and number of competing links nested under competing headings) for each goal on each web page. Then ACWW computes predicted mean total clicks for each goal on each web page, enabling the analyst to identify web pages that must be repaired for particular user goals/tasks. The modular design of ACWW makes it easy to more fully automate or refine each individual module independently. Thus, it would be possible to replace the current manual input of web pages with a more automated module that takes a web page(s) as input or even a set of URLs. It would also be possible to more fully automate the output, so that the analyst receives summary statistics for a web page or a Web site.

WRITERS' TOOL FOR IMPROVING WEB PAGE READABILITY

In order to design usable Web sites, designers must create content articles well matched to users' background knowledge and level of reading comprehension, and design Web sites that enable users to navigate by pure forward search and find the content easily. One of us (Mandalia, 2004) has developed a readability evaluation tool for developing content, and it can actually support both of the aforementioned design goals.

The readability evaluation tool affords a theory-based approach to content development by assisting the writer in monitoring and controlling three characteristics of expository text that have proven effects on comprehension and learning: percentage of low-frequency words, text coherence, and elaboration of key concepts. When using the tool to design content for optimal learning, published research on text comprehension and learning from text indicated that it would be best for writers to target about 5% low-frequency words, maintain high sentence-to-sentence coherence, and select familiar words and ideas to elaborate the key concept in a paragraph or section of text (the key concept is the concept, principle, process, or main idea that a writer is attempting to communicate in a passage).

When applying the tool to design or repair the navigation system—meaning the link and heading texts and groupings of links—the tool makes it easy for the designer to eliminate low-frequency words in link and heading texts, monitor coherence between pairs of links grouped together under a heading, and evaluate the heading text to make sure that it effectively expresses the key concept or relationship that unites a group of links.

Following user-centered design practices, Mandalia (2004) interviewed science and medical writers who have advanced degrees and who write science or medical articles intended for younger or less knowledgeable audiences (elementary, middle, and high school readers). Mandalia discovered that it was necessary to design the tool as an add-in to Microsoft Word to ensure seamless integration with the writers' workflow while revising texts for readability.

The multifunctional readability evaluation tool, like CWW, employs LSA measures of similarity and familiarity and represents the target user by a particular LSA semantic space. The first step in using the tool is to select the particular LSA semantic space that best represents the background knowledge of the target audience: 3rd-, 6th-, 9th-, 12th-grade, or first-year-college general reading knowledge. One function of the tool is to identify and highlight (in red font) all the low-frequency words in a text for a particular reading level and then to support the writer in attaining the target percentage of low-frequency words in the text. A second function supports monitoring and revision of the text to improve coherence, and a third function supports revision that optimally elaborates the key concepts the writer intends to communicate.

The readability evaluation tool was evaluated by doing user testing with the same science writers interviewed prior to building the tool. The user testing revealed that, apart from minor usability issues, the writers liked the overall organization and functionality of the tool. The psychological validity of the tool was empirically verified by a large experiment (n = 168) that found significantly higher learning gains for texts that the science writers had revised with the tool compared to texts revised without the tool. The empirical evaluation also provided evidence that the percentage of low-frequency words, text coherence, and key concepts elaboration do, indeed, influence text comprehension and learning gains, and the tool assists writers in better achieving these characteristics when producing content articles.

To verify these initial findings, it is necessary to run more experiments and replicate the results of the first experiment with more texts and with a broader sample of writers, and it is also necessary to test the tool with designers of navigation systems. The current version of the readability evaluation tool, the manual for using the tool, and papers about the user-centered design and the experiment are available for download on our research Web site.[6]

CONCLUSIONS

LSA has proved invaluable for CWW and for the readability evaluation tool, as it has for the practical applications and research reported in other chapters in this volume. LSA has enabled CWW to overcome all three of the crucial limitations of other UEMs. First, LSA provides reliable, objective ratings of similarity and familiarity that CWW substitutes for unreliable, subjective human judgments of similarity and familiarity. Second, CWW predictions, guided by the CoLiDeS cognitive model, have used LSA functionality to produce predictions of human navigation with high psychological validity. Finally, LSA enables us to solve the problem of scale, making it possible to build an automated version of CWW that can be applied to very large Web sites with 40,000 pages or more.

In addition to overcoming the limitations of other UEMs, each distinct LSA semantic space offers the means for simulating the influence of background knowledge on reading comprehension for a particular population of users. The versatility of LSA is its ability to simulate a multitude of user groups with high psychological validity by constructing a semantic space for each and every possible user group. Accordingly, an important goal for the immediate future is to extend the CWW research to other user groups and semantic spaces. Empirical evidence to date (Blackmon et al., 2002, 2003, 2005) has tested CWW predictions of heading and link selection only for college-educated users, using the college-level LSA semantic space. A driving motivation for the CWW research has been our hypothesis that we can successfully extend the CWW to evaluating Web sites for user groups who speak any language at any level of general reading knowledge. The first step toward verifying that hypothesis will be to make predictions from the sixth-grade semantic spaces with groups of experimental participants who have sixth-grade general reading knowledge.

[6] Download the readability evaluation tool, manual, and papers on the work at http://www.autocww.colorado.edu/~blackmon/Readability/ReadabilityTool.html

REFERENCES

Blackmon, M. H. (2004). Cognitive Walkthrough. In W. S. Bainbridge (Ed.), *Encyclopedia of human-computer interaction*(2 vols.). Great Barrington, MA: Berkshire Publishing

Blackmon, M. H., Kitajima, M., & Polson, P.G. (2003). Repairing usability problems identified by the Cognitive Walkthrough for the Web. *CHI Letters, 5: Proceedings of CHI 2003* (pp. 497–504). ACM Press.

Blackmon, M.H., Kitajima, M., & Polson, P.G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. *CHI Letters, 7: Proceedings of CHI 2005* (pp. 31–41). ACM Press.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive Walkthrough for the Web. *CHI Letters, 4: Proceedings of CHI 2002* (pp. 463–470). ACM Press.

Brown, R. (2005). *ACWW: Adding automation to the Cognitive Walkthrough for the Web (CWW).* Unpublished Master’s thesis, University of North Florida.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions and the Web. *CHI Letters, 3: Proceedings of CHI 2001* (pp. 490–497). ACM Press.

Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., & Cousins, S. (2003). The Bloodhound Project: Automating discovery of web usability issues using the InfoScent™ Simulator. *CHI Letters, 5: Proceedings of CHI 2003* (pp. 505–512). ACM Press.

Furnas, G. W. (1997). Effective view navigation. In *Proceedings of CHI’97* (pp. 367–374). ACM Press.

Gray, W. D., & Salzman, M. C. (1998a). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, *13*, 203–261.

Gray, W. D., & Salzman, M. C. (1998b). Repairing damaged merchandise: A rejoinder. *Human–Computer Interaction*, *13*, 325–335.

Hertzum, M., & Jacobsen, N. E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human–Computer Interaction*, *15* (1), 183–204.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, England: Cambridge University Press.

Kitajima, M., Blackmon, M. H., & Polson, P. G. (2000). A comprehension-based model of Web navigation and its application to Web usability analysis. In S. McDonald, Y. Waern, & G. Cockton (Eds.), *People and computers XIV—usability or else!* (pp. 357–373). New York: Springer.

Kitajima, M., Blackmon, M. H., & Polson, P. G. (2005). Cognitive architecture for website design and usability evaluation: Comprehension and information scent in performing by exploration. Invited session on Cognitive Architectures in HCI. *HCI-International 2005 Conference Proceedings* [CD-ROM].

Kitajima, M., Kariya, N., Takagi, H., & Zhang, Y. (2005). Evaluation of website usability using Markov chains and latent semantic analysis *IEICE Transactions Online, Vol. E88-B*, 1467–1475.

Kitajima, M. & Polson, P. G. (1997). A comprehension-based model of exploration. *Human–Computer Interaction, 12*, 345–389.

Mandalia, D. R. (2004). *User-centered design of a content analysis tool for domain experts*. Unpublished master's thesis, University of Colorado, Boulder.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: Implications for information architecture. *Human–Computer Interaction*, *19*, 225–271.

Pirolli, P. (2005). Rational analyses of information foraging on the Web. *Cognitive Science*, *29*, 343–373.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*(4), 643–675.

Pirolli, P. L., & Fu, W. (2003). SNIF-ACT: A model of information foraging on the World Wide Web (9th International Conference on User Modeling, June 22–26, 2003, Johnstown, PA). *Lecture Notes in Artificial Intelligence, 2702*, 45–54.

Pirolli, P. (2005). Rational analyses of information foraging on the Web. *Cognitive Science*, *29*, 343–373.

Pirolli, P., Fu, W-T, Chi, E., & Farahat, A. (2005). Information scent and web navigation: Theory, models and automated usability evaluation. Invited session on Cognitive Architectures in HCI. *HCI-International 2005 Conference Proceedings* [CD-ROM].

Rieman, J., Young, R. M., & Howes, A. (1996). A dual space model of iteratively deepening exploratory learning. *International Journal of Human–Computer Studies*, *44*, 743–775.

Soto, R. (1999). Learning and performing by exploration: Label quality measured by latent semantic analysis. *CHI Letters, 1: Proceedings of CHI'99* (pp. 418–425). ACM Press.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 105–140). New York: Wiley.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates.

APPENDIX A

Classify It as a Competing Heading

Rule 1

If the heading is not a correct heading
And if the heading has a goal-heading cosine ≥ .8 times the goal-heading cosine of the correct heading or the goal-heading cosine of the correct heading that has the highest goal-heading cosine if there are two or more correct headings
And if the goal-heading cosine of the heading ≥ .10 (i.e., NOT weak-scent)
And if the highest goal-link cosine of links nested under the heading ≥ .20

Then classify it as a competing heading.

Rationale. Users' attention is pulled to headings that are stronger than the correct heading, but it is meaningless to speak of "stronger than" when the higher goal-heading cosine is a weak-scent heading "in the noise." Requiring the competing heading to have a goal-heading cosine ≥ .8 times the goal-heading cosine of the correct heading is consistent with the way we compute competing links (i.e., as links with goal-link cosines ≥ .8 times the goal-link cosine of the correct heading).

Rule 2

I*f* the heading is not a correct heading
And if the highest goal-link cosine for any link nested under that heading ≥ .30 (i.e., strong-scent link)
Then classify it as a competing heading.

Rationale. Sometimes users are drawn to a particular heading by a strong information scent for a specific link(s) nested under the heading. Even if the strong-scent link does not work, the user will then search for other links similar to the strong-scent link under the same heading. For example, a person might first think "Chemistry" and then look for the heading where they would find the "Chemistry" link, that is, "Physical Science & Technology." Even if "Chemistry" turns out to not work, the user will think, "I must be close" and continue to search for other links with sufficient scent under the same goal.

Classify It as a Competing Heading Competing Link

Rule 1

If the link is nested under a competing heading
And if the goal-link cosine of the link ≥ .8 times the highest goal-link cosine of all the links nested under the competing heading
And if the goal-link cosine of the link ≥ .10
And if the goal-link cosine of the link is ranked no lower than fourth place when the goal-link cosines of links under the same heading are ranked in descending order, *or if* the goal-link cosine ≥ .30 (i.e., a strong-scent link)
Then classify it as a competing heading competing link.

Rationale. If the user's attention has been drawn to a competing heading, the user is apt to click links under that heading in order of decreasing information scent and then give up after clicking several high scent links under that heading, or even all links that are not weak-scent.

Rule 2

If the link is nested under a competing heading
And if the goal-link cosine of the link ≥ .20
And if there is no more than one link under the same heading with a higher goal-link cosine
Then classify it as a competing heading competing link.

Rationale. There are subregions where the highest-ranking link has such strong information scent that no other links under the same heading are ≥ .8 times the highest-ranking link. Nevertheless, users who focus on a heading and click the link with the highest goal-link cosine in that subregion, are likely to click at least one more link in that same subregion if they see one with fairly strong information scent (operationally defined as a goal-link cosine ≥ .20).

Classify It as a Correct Heading Competing Link

If the link is nested under a correct heading
And if the goal-link cosine of the link ≥ .8 times the goal-link cosine of the correct link
And if the goal-link cosine of the link ≥ .10
And if the goal-link cosine of the link is ranked no lower than fourth place when the goal-link cosines of links under the same heading are ranked in descending order, *or if* the goal-link cosine ≥ .30 (i.e., a strong-scent link)
Then classify it as a correct heading competing link.

Rationale. If the user's attention has been drawn to the correct heading, the user is apt to click links under that heading in order of decreasing information scent and then give up after clicking several high scent links under that heading, or even all links that are not weak-scent.

Classify It as a Weak-Scent Correct Link
If the link has a goal-link cosine < .10
And if the link is a correct link
And if there are no correct links with a goal-link cosine ≥ .10
Then classify it as a weak-scent correct link.

Rationale. A weak-scent link refers to the situation when a correct link is not semantically similar to the user goal and there are no other correct links that have moderate or strong scent. Weak-scent on the correct link makes the link an unlikely target of action, whether or not there is competition from other, higher scent links.

Classify It as an Unfamiliar Correct Link

If the text of a correct link is unfamiliar (i.e., *if* it has only one word and the word has a term vector length ≤ .55 *or if* the text of a correct link contains two or more words with a term vector length < .80)
And if there are no correct links that are not unfamiliar
Then classify it as an unfamiliar correct link

Rationale. Empirical evidence indicates that the term vector length is an approximate index of the amount of background knowledge that typical users have about a topic, and that unfamiliar problems happen when the term vector is low and typical users know little about the topic. The unfamiliarity partially or completely reduces the information scent, even in cases where the goal-link cosine is high. When some or all of the words in the link labels are low-frequency words, users may not even comprehend the meaning of the link, but these cases seem to be captured by short-term vector length without complicating the situation by examining word frequency.

APPENDIX B

Step 1: Select or Build and Select a Semantic Space

The first step in the CWW method is to select the most appropriate semantic space to represent a particular user group. Because the laboratory studies we have done to date all use college-educated experimental participants, we have consistently selected the semantic space for college-level general reading knowledge of American English. For a sixth-grade class in an American public school we would use the sixth-grade semantic space if about 50% of the students were proficient or above on the state achievement test for reading. Among both college-educated and sixth-grade groups, there are marked individual differences in reading ability and background knowledge, but in laboratory studies it is adequate to ignore these individual differences and use a single semantic space for the entire group.

If an appropriate semantic space does not exist already, it is possible to collect a scientifically sampled corpus of the documents (emulating Zeno et al., 1995) that are likely to have been read by a given user population. Using that corpus, the analyst can build a psychologically valid representation of the population. LSA semantic spaces can thus be built to provide a psychologically valid representation of virtually any user population with any level of background knowledge in any language or bilingual competence in two languages.

Step 2: Collect Set of User Goals

The next step is to collect a set of user goals to represent what that user group is likely to want to accomplish on the Web site under analysis. Ideally, user goals would be elicited by interviewing large samples of target users, and each user goal text would be a 100- to 200-word narrative description of what a particular user is looking for in a Web site.

As experimenters, we have not had the resources to collect user goals directly from users. In addition, it is useful to get data from 20 or more different persons completing the same assigned task. To perform a controlled laboratory test of the usability of the navigation system on an informational Web site, such as an online encyclopedia, we make the realistic assumption that the content articles presented on the Web site are valuable to users, and that users will invest considerable effort to surmount usability problems in a Web site if the Web site presents content that they find valuable. That assumption allows us to create user goal statements by using a summary of the target web page that we ask experimental participants to find in the Web site (e.g., the content article in an online encyclopedia).

The summary must be short (100 to 200 words), because experimental participants have been given only 130 seconds to complete the task. The summary also must be extremely similar to the actual article so that experimental participants have an accurate representation of the actual content they are trying to find in the Web site. We currently use the Summary tool of Microsoft Word to produce a summary of the complete online encyclopedia article we ask experimental participants to find. Then we use the LSA One-to-Many Comparison to compute the cosine between the text of the summary and the text of the complete article in the actual online encyclopedia. To ensure that the summary is an accurate representation of the content article, we aim for very high semantic similarity (operationally defined as a minimum cosine greater than .76). For the sample of 82 goal statements used in the first multiple regression analysis reported in this chapter, the mean summary-article cosine was .91, ranging from .76 to 1.00.

Step 3: Parse the Web Page and Identify Link and Heading Texts

Step three simulates how the user will parse the web page and identifies all the individual subregions of the web page. For example, CWW identifies nine subregions for the simple matrix web page layout in Figure 18.1, one subregion for each of the nine heading texts and cluster of links nested under the heading (e.g., the "Sports, Hobbies, & Pets" has a cluster of four links). The texts submitted to LSA include the link texts, meaning the texts that label each and every link on the web page. Heading texts will also be included in the LSA analysis if the web page designer decided to group related links together in the content area, or in one or more navigation bars. In some cases, the grouping of links has no heading text, but in that case the analyst can add together the texts of all the links in the group to create a pseudo-heading text, ensuring that LSA creates a single document vector for the link grouping. In other cases the designer uses an actual heading text to label the group. For example, Figure 18.1 shows a content area subdivided into nine groupings of links, and each grouping is labeled with a heading text that enables the user to scan the web page looking for the correct heading.

Step 4: Identify the Unfamiliar Heading and Link Texts

Familiarity measures include (a) term vector length for web page link texts and heading texts as an estimate of users' background knowledge of a given topic, and (b) word frequency within the selected semantic space. Current parameters identify a link or heading as an unfamiliar topic if the term vector length is .55 or less for a single-word text or less than .80 for a link/heading label with two or more words. Low-frequency words are defined as having a frequency of 15 or less in the corpus for the selected semantic space. CWW-guided repairs include substitution of familiar words for low-frequency words, but at present only the term vector length is used to identify unfamiliar heading and link texts.

Step 5: Elaborate the Heading and Link Texts

CoLiDeS and CWW also set simultaneous constraints on similarity and familiarity by using the near neighbors LSA analysis to simulate the process of elaboration that occurs during comprehension of short heading and link texts on a web page. The underlying CoLiDeS assumption is that the terms most likely to be activated by reading a web page text are those that are highly similar to the text and are also high frequency, familiar terms, so current parameters for elaborating texts with near neighbors specify a minimum document-to-document cosine of .50 and a minimum word frequency of 50. Elaborating link texts adds the near neighbors to the original, unelaborated link text that appears on the web page. The Elaborate Links function at http://www.autocww.colorado.edu can apply near neighbors analysis for up to about 50 link/heading texts input in a single batch with blank lines separating them.

Elaborating heading texts is more complicated than elaborating links. For example, under the heading "Life Science," the heading text expands to the string of words "Life Science science sciences biology scientific geology physics life biologist physicists," the link label "Birds" expands to the elaborated link label "Birds, birds bird feathers beak wings eagle nest nests fly wing geese hawk flew pigeons feather eagles owls fluttered flying," and the link label "Medicine" expands to "Medicine, medicine medicines doctor doctors prescription sick medical clinic." To create the full elaboration for the heading text for "Life Science" CWW combines the elaboration of the words "life science" with the elaborated link texts for all 14 links nested under Life Science, resulting in a 268-word text to represent the semantic meaning of "Life Science" for college-level users confronting the web page shown in Figure 18.1.

Step 6: Use LSA to Compute Goal-Link and Goal-Heading Cosines

The next step applies the LSA one-to-many analysis to compare the goal statement with the elaborated headings and links, set to produce document-to-document cosines. Thus, cosines are based on comparing a 100- to 200-word goal statement with the elaborated versions of each link and heading text, not just the link and heading texts printed on the web page.

We then separate the results into a group of heading-goal cosines and a group of link-goal cosines, sorting each group by decreasing cosine value. Next, we examine the sorted results and identify and mark the correct heading(s) and link(s), the ones that actually lead to accomplishing the goal in the actual online Web site being simulated.

Step 7: Apply Automatable Rules to Identify Usability Problems

The sixth step is to apply an automatable set of rules (see appendix A) for distinguishing unfamiliar correct links, weak-scent correct links, competing links nested under competing headings, and competing links nested under correct headings.

Step 8: Follow CWW Guidelines for Repairing Problems

Next we examine the results and see how to repair the problems. When fully repaired, the CoLiDeS model leads us to expect users to be able to accomplish the goal on the repaired page with pure forward search, following the high information scent on any of the competing heading and competing links. CWW does not yet have a set of automatable rules for repairing usability problems, but the general processes of repair are covered in Blackmon, Kitajima, and Polson (2003) and in the text of this chapter.

Step 9: Use CWW Formula to Predict Mean Total Clicks

At the final step, we apply a newly developed CWW formula for predicting the mean total clicks under both the repaired and unrepaired condition. The current formula is derived from a compilation of 228 tasks, pooling both the 164-task dataset and 64-task cross-validation study:

> Predicted mean total clicks required to find the item on a Web site = 2.292
> + 1.757 if there is an unfamiliar correct link
> + 1.516 if there is a weak-scent correct link
> + 0.655 times the number of competing links nested under competing headings